Microsoft

2026 Responsible AI Transparency Report

Content

Foreword

AI that earns and sustains trust has the potential to benefit people, organizations, and communities in every part of the world. But that trust, and those benefits, are not a given. At Microsoft, we recognize that how we build and deploy AI, support customers, and act as a technology provider helps determine whether AI is trusted and whether its benefits are realized broadly. That recognition drives a rigorous commitment to doing and showing our work to govern AI: understanding the risk landscape, investing in safeguards, and collaborating to strengthen governance as technology evolves—so that our customers and partners can use AI with justified confidence, and in service of their goals.

For us, responsible AI is not a one-time exercise but a continuous discipline. We strengthen safeguards by probing evolving capabilities and limitations, observing how systems perform in the lab and in the real world, and iteratively improving how AI is designed, evaluated, and governed. As AI capabilities advance at extraordinary speed and new questions emerge about how these systems are developed, adapted, deployed, and monitored over time, our continuous discipline guides how we embed responsible AI across our business and into our broader strategy.

Through a combination of people, policies, processes, and tools, we empower teams across Microsoft—and the customers who build on our platforms—to develop and deploy AI responsibly. We invest in this program with the same seriousness we bring to advancing AI systems and enabling customers, as rapid technological change and rising societal expectations demand governance that is both robust and adaptable.

Accelerating agility: A re-engineered Responsible AI Standard

This past year marks a significant step forward in our program's evolution. Our AI principles have proven durable across successive eras of AI development—from traditional machine learning to generative AI and increasingly agentic systems—providing continuity as technology has advanced. Building on that foundation, we have re-engineered our Responsible AI Standard and strengthened our organizational muscle to operationalize it at enterprise scale.

This re-engineering is more than a technical update. The Standard's new architecture, described in Sections 2 and 3, reflects a fundamental refresh that sharpens how we manage risk across the AI tech stack and the lifecycle by which AI systems are developed and deployed. It enables us to respond more nimbly as AI capabilities accelerate, deployment patterns evolve, and technical and governance questions emerge. As the trends highlighted in Section 1 underscore, the need for agility has never been clearer.

Leading with impact in an ever-changing discipline

As AI systems become more capable and integrated into everyday work and life, responsible AI will remain an ever-changing discipline—requiring continuous learning from our own experience and through collaboration with others. When AI systems fall short or are misused, Microsoft is prepared with an incident response program designed to accelerate continuous improvement. Reported issues, as catalogued in Section 2, also inform our evolving governance.

Partnerships help us stay sharp and extend our reach. As detailed in Sections 2 and 4, to strengthen evaluation capabilities, we have invested in collaborations with the US Center for AI Standards and Innovation and AI Safety and Security Institutes; contributed to MLCommons benchmarks; and shared new tools with customers and the ecosystem. To expand research and understanding of priority risks, we launched an External Red Team Alliance with 18 universities across six continents, and we're listening to teens seeking healthy AI interaction.

We have also taken concrete steps to protect and support people. These include implementing policies to manage risks in sensitive human-AI interactions, providing training for older adults impacted by AI-enabled fraud and scams, and investing in multilingual AI capabilities to support more reliable system performance in languages that are not well represented in typical training datasets.

The 2026 Responsible AI Transparency Report reflects our ongoing commitment to sharing information about how we govern AI at Microsoft: what we have learned, how our program is changing, and where important challenges remain, motivating expanded collaboration across the AI ecosystem.

Responsible AI is not just about anticipating known risks; it is about building governance that remains effective as the technology, its uses, and societal expectations continue to evolve. Trust is essential to unlocking AI's benefits, and we remain focused on earning and sustaining it through effective governance—disciplined action, continuous improvement, and openness about our progress and priorities.

Natasha Crampton
Chief Responsible AI Officer

Key takeaways

We are designing and implementing more adaptive governance as AI rapidly advances.

As AI capabilities advance, adoption grows, and societal expectations evolve, governance cannot remain static. This year, we re-engineered our Responsible AI Standard so that our requirements can adapt more readily to new capabilities, uses, and risks while applying consistently across the models, platforms, and applications we develop and deploy. We also strengthened pre-release reviews, supporting more than 450 reviews of Sensitive Uses and nearly 2,500 generative AI product and feature reviews from July 2025 through June 2026.

We are strengthening safeguards across the AI lifecycle, from anticipating emerging risks to responding to real-world use.

Increasingly agentic, interconnected, and consequential AI systems require stronger approaches to identifying, preventing, monitoring, and responding to risks. Over the past year, we updated our approach to frontier and sensitive interaction risks; expanded evaluation, red teaming, controls, and monitoring of agentic systems; and strengthened defenses and incident response readiness for fraud, cyber threats, and other forms of misuse. These efforts reflect a broader shift from release and point-in-time assessments toward continuous governance.

We are making responsible AI more practical, transparent, and accessible across the AI ecosystem.

Microsoft is expanding the tools, training, and technical guidance that customers and developers need to govern AI in practice. We are also investing in more reliable AI performance across languages and cultural contexts and working with governments, researchers, civil society, and industry on shared approaches to evaluation, transparency, and assurance. These efforts aim to broaden access to responsible AI capabilities, strengthen accountability, and build trust through evidence and collaboration beyond any one company.

Taken together, these efforts reflect a simple commitment: as AI becomes more powerful and more present in people's lives, Microsoft will continue evolving governance, strengthening safeguards, and expanding the practical tools needed to earn and sustain trust.

Section 1

2026 trends: What is changing and how we are adapting

Five trends impacted Microsoft's responsible AI program over the last year—reflecting where risks and opportunities are emerging, why they matter, and how we are adapting. Trust is being shaped not just by model performance but also by how systems behave in real-world settings and how organizations govern them over time. In response, our program is focused not only on technical progress but also on coordinated governance, operational discipline, and the ability to learn systematically from deployment and user experience.

AI innovation and diffusion are accelerating—even as gaps widen across geographies

AI technologies are rapidly advancing while becoming more deeply embedded in everyday tools, workflows, and decisions. At the same time, our Global AI Diffusion Report highlights a widening divide in who is accessing and benefiting from AI, with disparities across geographies and languages.[1]

Why it matters—and how we are adapting

The growing pace of AI innovation and diffusion requires more adaptable and scalable governance. Over the last year, we have responded by:

- **Re-engineering our Responsible AI Standard** to address the complex and evolving AI tech stack and lifecycle through core and scenario-specific requirements that apply to our development and deployment of AI models, platform services, and applications.
- **Increasing rigor and efficiency in pre-release reviews** of how high-impact AI capabilities, features, products, and deployment scenarios implement our Standard. We supported 450+ Sensitive Use reviews from July 2025 to June 2026, where over 30% of cases were related to agentic AI, integrating responsible AI expertise upstream in the development lifecycle. Over the same period, we conducted pre-release reviews for nearly 2,500 generative AI features or products, increasing the coverage of capabilities, risks, and quality metrics measured through our central testing platform.
- **Investing in our internal responsible AI community** through training and community-based learning about prompt injection defenses and threat modeling for agentic systems.

Increasing AI adoption heightens the need for customer and ecosystem support in a wide range of deployment scenarios. Over the last year, we have met that demand by:

- **Strengthening our readiness to work with customers** on responsible AI and customized deployments by supporting members of our internal responsible AI community who focus on building custom solutions for Microsoft customers. These team members submitted about 30% of all Sensitive Use cases reviewed from July 2025 to June 2026.
- **Expanding responsible AI tools available open source and via our Foundry platform,** including coverage of risk and performance domains, technology capabilities, and lifecycle stages.
- **Scaling customer and ecosystem training,** including hands-on training with security experts, publicly available red teaming resources, and broader AI literacy programs through Microsoft Elevate, such as "teach the teacher," which connects school leaders with resources to confidently integrate AI in the classroom.

Increasing but uneven diffusion has underscored the importance of investing in AI performance across languages and cultural contexts. Over the last year, we have focused on:

- **Researching AI evaluation methods across languages and cultures**, with 6 locales represented in an MLCommons AILuminate working group.[2]
- **Building data and model capability for digitally underrepresented languages** through LINGUA Europe and LINGUA Africa.[3]
- **Co-designing AI technologies with local communities** in East Africa, South Asia, and beyond, informing new playbooks for multilingual AI capabilities and helping inform the launch of PazaBench, the first automatic speech recognition leaderboard, with initial coverage of 39 African languages.[4]

+

As AI adoption accelerates globally, we are scaling our AI governance program alongside it—supporting diverse deployment contexts with scaled pre-release checks, tools, and training while addressing gaps in performance across languages.

Generative and agentic AI capabilities are expanding—raising imperatives for observability and control

In traditional machine learning, risks were often assessed primarily in relation to a single model, its training data, and a defined use case. Early generative AI applications similarly functioned largely as standalone tools with relatively simple architectures. Today, however, generative AI applications often integrate multiple models with external tools, data sources, and services. Agentic systems extend this further, introducing greater interconnection, persistence, and autonomy in interactions. In multi-agent scenarios, that complexity is multiplied.

As agentic capability improves, risk is increasingly shaped by how multiple models, tools, and systems are dynamically assembled, configured, and integrated; how access and permissions are scoped; how behavior evolves through use and memory; and how actions are chained together and executed over time.[5]

Why it matters—and how we are adapting

Where risk profiles are more dynamic, governance approaches based on individual model evaluations and static assessments are less effective at addressing system-level interactions and emergent behaviors. Agentic capabilities can also increase the risk of loss of effective control, raising the importance of guardrails, monitoring, and response across the AI lifecycle, including during pre-deployment testing. Over the last year, we have increased investments in policies, practices, tools, and partnerships to strengthen the foundations of how systems are designed for these new paradigms and to monitor and manage them over time, including by:

- **Deepening readiness to assess and manage emerging risks.** We updated our Frontier Governance Framework to address loss of control risks, and we refined policies and practices to address systems retaining "memory" about their interactions with users, building upon findings from a multistakeholder workshop.[6] We initiated research collaboration with the US Center for AI Standards and Innovation on methods to detect techniques that compromise agent workflows. We also scaled our monitoring capacity and strengthened rapid investigation of reported issues.
- **Enhancing risk identification techniques and evaluation tools.** We developed and integrated into our Security Development Lifecycle and Responsible AI Standard new threat modeling techniques for agentic AI, scaling implementation through internal deep-dive workshops that drew 3,000+ engineers. We also released an open-source framework, ASSERT, and tools on Foundry that enable customers to turn written AI agent policies into repeatable tests as well as examine the end-to-end outcomes and the step-by-step execution of agentic systems. These evaluators focus on capabilities like task adherence, tool selection, and tool call accuracy as well as implementation of guardrails.[7]
- **Advancing capabilities for identity, runtime controls, and observability** in applications with agentic capabilities or services that support development or deployment of agents, including:
  - Access control and greater accountability through verifiable agent identities (Microsoft Entra Agent ID) and portable runtime policies (Agent Control Specification) that define which agents may access which tools, resources, and actions under what conditions—enabling approval gates for sensitive or irreversible actions and administrative controls to activate, block, or revoke agents;
  - Telemetry on agent actions and runtime policy enforcement events, supporting human or administrator intervention when anomalies, test failures, or policy violations arise; and
  - Layered defenses against adversarial manipulation, including repeatable evaluations that can regularly monitor whether systems meet defined safety and security claims (RAMPART).

+

As AI systems become more integrated, autonomous, and persistent, we are enabling customers to manage and monitor these systems' actions in real time and working with partners on industry standards and novel research.

AI misuse is scaling—increasing people's exposure to fraud, scams, and cyber exploits

Like other technologies, AI can be powerful in the hands of people using it for broad benefit or harm. The increased capabilities of generative systems have lowered barriers to producing convincing messages, voices, and images, and malicious actors have leveraged AI to scale fraud, impersonation, deepfakes, and socially engineered attacks. More capable models and agentic systems have also heightened risks that malicious actors can more rapidly discover critical code vulnerabilities and develop powerful exploits.

Why it matters—and how we are adapting

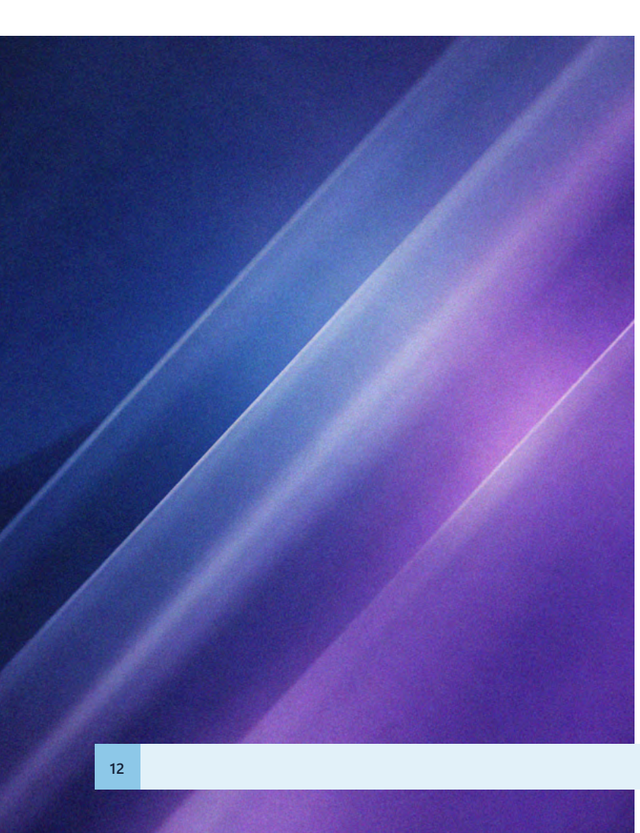
Fraud and scam risks disproportionately affect vulnerable populations while broadly undermining societal trust in AI technologies. While AI is key to rapidly transforming cyber defense, misuse of advanced cyber capabilities could also heighten exposure to risks of significant cyberattacks, including in the context of critical infrastructure. Defending against AI misuse requires a multi-layered strategy, especially because those using AI for harm are likely to be persistent, constantly employing new techniques and tactics to circumvent or break through guardrails. Over the last year, we have invested in:

- **Probing for adversarial robustness and applying layered mitigations.** In 2025, our AI Red Team executed more than 100 operations. We strengthened safeguards against violative image generation through iterative safety testing and guardrails across the AI tech stack, from MAI-Image models to Bing Image Creator. To defend against prompt injection attacks—where malicious inputs attempt to override intended safeguards—we enhanced classifiers, including Prompt Shields,[8] and implemented "spotlighting" to help models distinguish trusted from untrusted content.
- **Scaling readiness to mark AI-generated content and attach provenance metadata** across relevant product surfaces, supporting downstream detection and mitigation of deepfakes and fraud. We are pairing watermarking with C2PA-based secure provenance metadata[9]—a verifiable record showing where content originates—with central solutions to monitor at scale.
- **Intensifying focus on incident response readiness** to strengthen rapid and coordinated investigation, containment, and learning. In 2025, we formalized cross-functional incident response playbooks across 13 topics and processed over 11,000 submissions to the Microsoft Security Response Center, scaling processes to incorporate learnings into policy and safeguards.

Collaboration is essential to address ecosystem-wide misuse. Over the last year, we have also been:

- **Collaborating to harden code bases and detect adversarial behavior.** We joined Project Glasswing to help safeguard software systems worldwide as more capable AI models and agentic systems risk dramatically accelerating the discovery and exploitation of software vulnerabilities. We also collaborated to develop TaskTracker, a new technique for detecting indirect prompt injection attacks on large language model (LLM)-powered systems, as well as a public dataset of over 200,000 attack prompts.[10]
- **Disrupting criminal misuse of AI technologies.** In 2025, our Digital Crimes Unit worked with international law enforcement to dismantle a transnational tech support fraud network that used AI-enabled tactics to target older adults.[11]
- **Supporting populations disproportionately impacted by fraud and scams.** We partnered with the American Association of Retired Persons (AARP) and its affiliated nonprofit, Older Adults Technology Services (OATS), to deliver fraud-prevention messaging, and committed to scaled industry efforts to help protect people from scams through the Industry Accord Against Online Scams & Fraud.[12]

> As AI lowers the cost and increases the scale of misuse, we are investing in iterative risk management, strengthening incident response, and coordinating across technology researchers, civil society, and the public sector to help prepare defenders and users.



People are turning to conversational AI for support—amid growing awareness of risks

Conversational AI services have become a primary interface for interacting with AI models, with expanding use among non-experts and in more sensitive contexts like personal advice, health-related questions, and emotional support. Analysis of aggregated and de-identified Microsoft Copilot usage data indicates that health-related queries dominate mobile usage and that nearly one in five health-related conversations involve personal symptom assessment.[13]

Why it matters—and how we are adapting

While conversational AI offers clear benefits—such as accessibility and ease of use—it also raises important risks, including emotional dependency or misplaced trust in AI outputs. These risks are heightened for vulnerable users, including children, underscoring the urgency of strengthening safeguards as chatbot usage grows. Over the last year, we have focused on:

- **Evaluating risks to individuals and society and implementing layered mitigations** for consumer-facing experiences, in line with our new Sensitive Interactions Policy. Mitigations include system prompts, content classifiers, and system responses to harmful content and unsafe behavior, including self-harm or suicidal ideation. Additional protections for vulnerable users include age-based access controls, enhanced protections for adolescents, and parental controls through Microsoft Family Safety.[14]
- **Pairing product-level interventions with ongoing research and learning directly from youth.** Microsoft Research is examining what design choices promote human agency and appropriate reliance. We are also hearing directly from teens across regions through an AI Futures Youth Council and in workshops, facilitated in partnership with Cyberlite, to co-design "by teens, for teens" guidance on AI companions.[15]
- **Advancing shared approaches to assessing psychosocial risks and mitigations.** With academic, civil society, and industry partners in the MLCommons Alluminate Psychosocial Task Force, we are discussing how risks are conceptualized and how the impact of mitigations can be measured, aiming to develop shared benchmarks covering multiple risk areas, including crisis prevention and unqualified AI advice.

> As conversational AI evolves as a primary interface, we are focused on layered mitigations, research on user experience, and shared measurement tools.

AI regulation is evolving rapidly—while trust remains a key adoption enabler

Global regulatory activity has intensified over the last year, with the EU AI Act moving into enforcement and new laws taking effect in US states and elsewhere. At the same time, uncertainty remains around interpretation, enforcement timelines, and alignment across jurisdictions. Meanwhile, AI deployers in regulated sectors continue to require confidence from their AI supply chain, and end user trust remains pivotal.

Why it matters—and how we are adapting

Global regulatory dynamics have heightened concern that overly broad or prescriptive regulation could slow innovation and adoption—or that persistent gaps could undermine trust. In this environment, Microsoft has focused on:

- **Providing clarity and confidence for customers.** We were among the first providers to release a general-purpose AI (GPAI) model in the EU under the GPAI Code of Practice, meeting transparency obligations through the publication of a data summary and model card for MAI-Image-1.[16] In this third iteration of our Transparency Report, we are also offering insights into how we are using AI responsibly.
- **Enhancing compliance readiness.** In re-engineering our Responsible AI Standard, we helped teams across Microsoft align more efficiently with evolving regulatory requirements. We analyzed more than 100 enacted and proposed laws and refreshed our process for continuous analysis and adaptation going forward, strengthening the agility of product engineering and customer solutions teams by re-factoring how the Standard integrates requirements dynamically.
- **Partnering to develop best practices, standards, and tools** that can advance ecosystem maturity and support more streamlined and integrated assurance. We are contributing to Frontier Model Forum technical guidelines on areas like managing advanced cyber risks and to MLCommons as it develops reliability benchmarks and testing capacity.[17] To support AI assurance efforts across AI supply chains and deployment sectors, we also joined the Appia Foundation as a founding member.

We have leaned into information sharing as a stabilizing signal for regulators, customers, and users alike, aiming to understand how improved transparency can be leveraged for more targeted and adaptable governance, including by:

- **Refreshing our transparency templates** that support teams developing data summaries and model, platform, and application cards and implementing them across the company throughout 2026.
- **Contributing to external reporting frameworks that support interoperability,** including the OECD-led informal task force to develop the Hiroshima AI Process (HAIP) Reporting Framework version 2.0, which streamlines reporting, expands and clarifies applicability across the AI value chain, and integrates new topics like agentic AI.[18]
- **Investing in better understanding the impact of transparency in practice**—analyzing what information is needed, by whom, and at what point in the AI lifecycle. This includes support for Berkeley AI Research (BAIR) Lab academic research that informed the development of a playbook on transparency in generative AI systems.

In a rapidly evolving regulatory landscape, we are providing clarity and confidence for customers and product teams while helping inform practical and interoperable regulatory approaches.

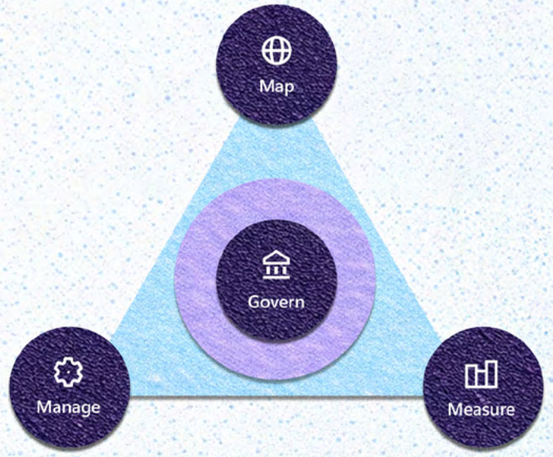
As AI technologies themselves become more capable and integrated, so too must responsible AI across the development and deployment lifecycle. Over the last year, our focus has been on scaling governance to match adoption, designing safeguards for real-world use, learning from deployment, and supporting customers and partners navigating complex regulatory and operational environments. Integrating technical, operational, and governance measures helps us translate AI innovation into outcomes that people and organizations can trust.

Section 2

How we develop, release, and operate AI responsibly

Summary of our risk management steps and progress over the last year

Our commitment to responsible AI begins with the AI technologies we develop, release, and operate. At Microsoft, we align this work with the NIST AI Risk Management Framework (RMF) functions of Govern, Map, Measure, and Manage, complemented by centrally managed pre-release oversight processes to verify that risks have been appropriately addressed before release. In practice, as the RMF recognizes, this is an ongoing, iterative process spanning both pre- and post-release stages. In this year's report, a dedicated sub section describes post-release stages.



Govern

Define policies and drive implementation.

We re-engineered our Responsible AI Standard; developed new tooling; and invested in our Responsible AI Governance Community.

Manage

Implement mitigations to reduce risks to an acceptable level.

We implemented new or expanded guardrails, including prompt injection defenses and safety classifier coverage; established new templates for transparency documentation; and strengthened governance capabilities for agents.

Map

Identify risks that apply to AI technologies.

We enhanced AI risk taxonomies; introduced new agentic AI threat modeling practices; and launched an emerging tech program to address new risks from frontier AI capabilities.

Release

Review decisions to make AI technologies available.

We assessed nearly 2,500 generative AI releases, 30% of which were related to agentic AI, and reviewed 450+ Sensitive Uses cases, where nearly 30% were custom solutions developed for customers, from July 2025-June 2026.

Measure

Assess risks and the impact of mitigations.

We expanded internal evaluation capabilities in risk coverage and tools; established partnerships with government institutes to advance evaluation; and leveraged a new external testing framework for Copilot Health.

Post-release Map, Measure, and Manage

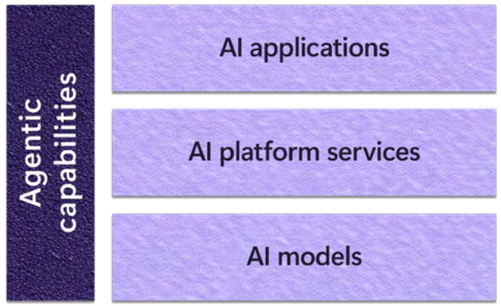
Address evolving risks and mitigations and implement ongoing improvement.

We scaled AI incident response capabilities, formalizing 13 incident response playbooks and strengthening capabilities to accelerate response timelines and address new attack patterns.

Case studies across the AI tech stack

Microsoft's AI technologies span from models to platforms that support AI development and deployment to applications that bring AI to life for users. We apply responsible AI governance at every layer. This Report's case studies across the tech stack help to demonstrate what that means in practice.

- At the base of the stack are our AI models, including general-purpose, domain-specific, proprietary, and open-weight models. Microsoft's MAI-Thinking-1 model is a general-purpose AI model with capabilities such as content generation, reasoning, and task assistance. MAI-Image-2 is a text-to-image model integrated into Bing Image Creator and available for customers to deploy in applications via Foundry.[19] Fara-7B is Microsoft's first agentic small language model designed specifically for computer use.
- Our platform services, such as Foundry Agent Service, enable developers, enterprises, and power users to build, customize, and manage AI applications, integrating our proprietary and open-weight models or those developed by partners or other third parties. Depending on their specific features, platform services may provide infrastructure for training, fine-tuning, orchestration, evaluation, memory, identity, and governance—supporting the development and deployment of AI responsibly and at scale.
- At the top of the stack are our AI applications and features, such as M365 Copilot (including features such as Copilot Cowork), Browse with Copilot, and Copilot Health. Applications may integrate multiple Microsoft and/or third-party models and leverage platform-level capabilities (e.g., infrastructure and APIs that enable applications to integrate models), delivering AI-powered experiences to consumers and enterprise users.



Agentic capabilities and the AI tech stack

Agentic capabilities apply across layers of the AI tech stack. Agents are enabled by models that are optimized for agentic use and may operate as part of platform services or applications, performing multi-step tasks by managing how models interact with tools, data, or environments. For example, agentic capabilities and multi-agent systems can appear as research or coding agents embedded in applications (e.g., GitHub Copilot Agent) or integrated with services (e.g., the multi-model agentic scanning harness MDASH in Microsoft Defender). Platform services may also embed agents with orchestration, planning, memory, or tool execution capabilities (e.g., for model routing or coordinating other workflows).

Describing agentic capabilities according to the stack without assigning them to a fixed layer helps clarify the architecture while remaining future-proof. As AI systems evolve toward more autonomous, multi-step, and multi-agent workflows, the core role of agents—as the mechanism for coordinating models, tools, and actions—remains consistent even as their placement within products varies.

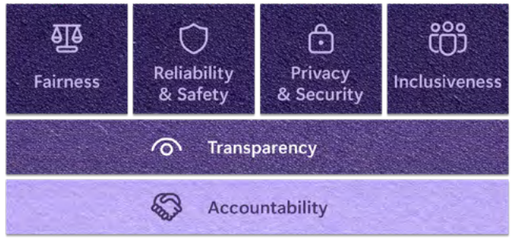
Products at multiple layers of the tech stack also enable customers to build and manage agents. For instance, Copilot Studio provides a low-code platform for building agents that can run independently or appear inside M365 Copilot (where Copilot acts as a meta-interface that routes tasks to the appropriate specialized agent). Foundry provides developers with a programmatic environment where they can design and operate agents at scale for custom applications or services. Foundry[20] (for developers) and Agent 365[21] (for IT administrators) enable management of agents, providing capabilities for agent identity, permissions, policy enforcement, and observability.

Govern: Defining policies and driving implementation

Microsoft has built an adaptive AI governance program anchored in enduring AI principles, a re-engineered Responsible AI Standard, and an engaged community and accountable oversight structure.

Principles, policies, practices, and tools

Microsoft's AI principles continue to serve as the north star of our approach to AI governance. Over multiple waves of innovation—from traditional machine learning to generative AI and now increasingly agentic AI—these principles have provided continuity, informing the creation of our first version of the Responsible AI Standard in 2019 and guiding its evolution.



Our responsible AI policies have evolved amidst rapid and compounding technology and governance change. As AI capabilities have continued to advance, the AI value chain has grown more complex and interconnected—with organizations increasingly operating across roles, from adapting models to developing and deploying applications. At the same time, regulatory and policy expectations have accelerated globally, and best practices have continued to mature.

Key developments in the history of our principles and Responsible AI Standard

Microsoft's commitment to respect human rights has shaped our responsible AI efforts from the outset and continues to guide our program's evolution.

- Our Global Human Rights Statement, which outlines our commitment to respect and advance human rights across our operations and technologies,[22] is foundational to our AI principles. It is informed by instruments such as the UN Guiding Principles on Business and Human Rights, the Universal Declaration of Human Rights, and the Convention on the Rights of the Child.
- In 2018, we commissioned a human rights impact assessment of AI that helped establish a rights-respecting approach to AI governance.
- In 2024, we commissioned a second human rights impact assessment focused on generative AI, examining emerging capabilities (such as autonomy and multimodality) and deployment patterns affecting our responsibility to respect human rights.[23]

Our Responsible AI Standard has also evolved with technology and as we have learned.

- The Standard was updated in 2022 to reflect ongoing research and lessons from implementing our governance program.
- In 2023, we formalized a set of generative AI requirements based on the Standard.
- In 2024 and 2025, we complemented our Standard with targeted guidance addressing new risks and design considerations, such as multimodal capabilities, persistent memory, computer use agents, and increasingly autonomous agentic AI. During this period, we also developed and publicly released our Frontier Governance Framework. It guides how we track, assess, and mitigate advanced AI model capabilities that could be misused to threaten national security or pose at-scale public safety risks.

Our re-engineered Responsible AI Standard

Developer and deployer

New chapters in the Standard clearly distinguish developer and deployer responsibilities, reflecting how AI moves from development into real-world use.

AI tech stack

Requirements are now tailored across AI models, platform services, and applications, and safeguards match how agentic capabilities weave through the AI tech stack.

Risk-based rules

The Standard now separates core requirements from scenario-specific ones, applying additional safeguards where AI use poses higher potential risk.

Security aligned

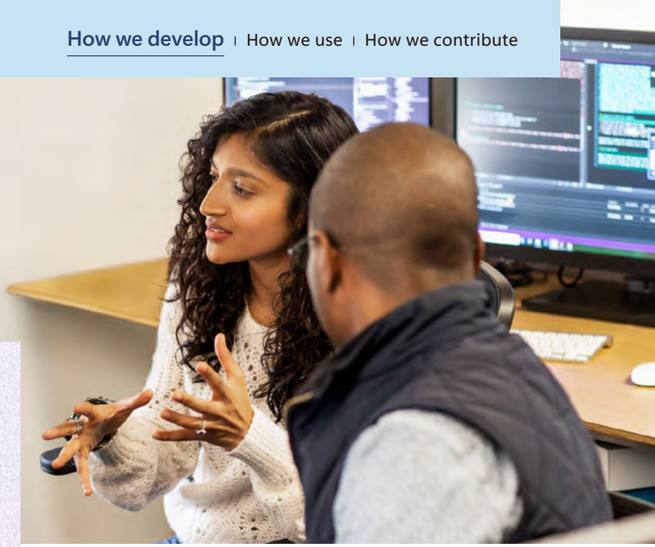
The refreshed Standard integrates previously standalone policies and closely aligns with evolving security requirements defined in the Security Development Lifecycle.

A new Standard for adaptability in a dynamic AI era

Over the past year, we re-engineered our Responsible AI Standard to reflect how Microsoft's AI investments and responsibilities have evolved and to clarify how governance requirements are integrated and updated in a dynamic environment, including by:

- Distinguishing Microsoft's responsibilities across the AI value chain—covering how we develop and deploy AI technologies in a Developer Chapter as well as how we deploy third-party AI applications for internal and external use in a Deployer Chapter. This marks the first time Microsoft has integrated a Deployer Chapter in the Standard. We also defined responsibilities for custom AI solutions within the developer framework.
- Recognizing differences across the technology stack—including for **AI models**, **AI platform services**, and **AI applications**—and addressing agentic capabilities that may apply across layers.
- Organizing map, measure, and manage requirements into two types: **core** requirements, which apply to each tech stack layer, and **scenario** requirements, which apply in specific cases, such as frontier models, agentic AI, or **sensitive interactions**. Scenario requirements enable the Standard to adapt with greater agility. Additionally, for policies maintained distinctly, such as our **Frontier Governance Framework**, these scenario requirements simplify alignment and integration with the Standard.

The Standard also continues to align closely with requirements in adjacent domains, including updated Security Development Lifecycle (SDL) practices for threat modeling, agent identity management, access control, memory safeguards, logging, and reliable shutdown mechanisms.



Defining a Sensitive Interactions Policy as a set of scenario requirements

We defined a new Sensitive Interactions Policy to address AI interactions that involve open-ended conversation, memory, and personalization across sessions. The Sensitive Interactions Policy is designed to address the risks of social and emotional interactions with AI systems. It requires techniques to prevent anthropomorphic behavior and sets boundaries to help users avoid emotional entanglement, with enhanced safeguards for youth.

Building the Standard in alignment with an evolving Frontier Governance Framework

The re-engineered Standard is designed to simplify alignment with standalone policies, such as Microsoft's Frontier Governance Framework, first published in February 2025. Through the core and scenario model requirements, teams arrive at the same expectations defined in the Frontier Governance Framework. In early 2026, Microsoft updated our Frontier Governance Framework to align with emerging regulatory requirements and reflect feedback from internal and external stakeholders. The update clarifies how Microsoft tracks high-risk capabilities such as harmful manipulation and loss of control, and it introduces new disclosures about our ongoing red teaming and risk modeling practices. While the changes do not reflect shifts in operational policy, they provide greater transparency into how Microsoft evaluates and monitors frontier model risks, engages third-party evaluators, and supports internal reporting and accountability.

Developing and implementing a more adaptable Responsible AI Standard

The re-engineered Standard was developed through an iterative, consultative process that brought together 80+ employees representing 30+ teams across Microsoft—spanning responsible AI research, engineering, and policy as well as security and product development.

To support implementation at scale, we are pairing policy with practical enablement. Teams receive iterative guidance to translate policy-level requirements into engineering-ready implementation. The guidance is maintained collaboratively by policy, engineering, and research teams. Internal tooling and direct engagement with responsible AI experts help teams implement requirements as part of their regular workflow, with resources at the ready when questions arise.

We are also continuing to progress internal tooling to support data governance and responsible AI documentation and review processes. Tooling helps increase end-to-end visibility into risk assessment, adversarial testing, systematic measurement results, and incident response data so product, governance, and leadership teams can track risk trends and compliance progress.

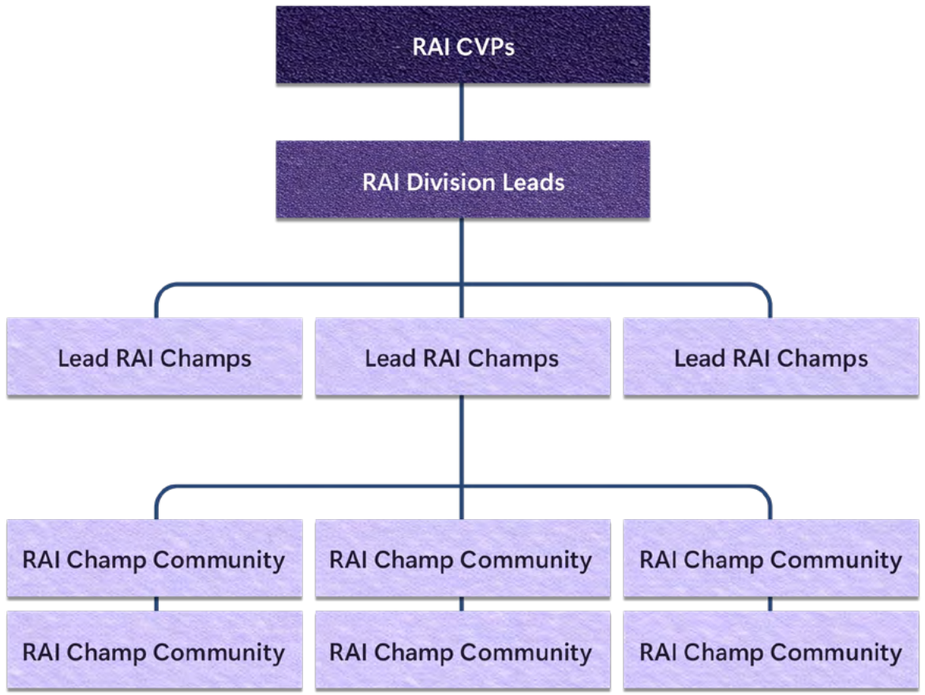
Looking ahead, leveraging the new architecture, the Standard's requirements will be updated regularly, informed by evolving technology trends, societal and regulatory expectations, and feedback from teams implementing the Standard in practice. Strong governance processes and an engaged community are key to ensuring the success of our Standard and its continued evolution.

People and processes

Microsoft's Responsible AI Governance Community—the people who energize it and the processes that facilitate their collaboration and accountability—are the engine behind our cross-company effort.

- At the executive level, Microsoft's responsible AI program is supported by senior leadership, including through quarterly oversight meetings with our Responsible AI Council co-led by Vice Chair and President Brad Smith and Chief Technology Officer and Executive Vice President of AI Kevin Scott.
- The Council is comprised of Corporate Vice Presidents (CVPs) representing all areas of the business, including engineering, sales, and business functions. Those CVPs are supported by selected division-level leaders that are responsible for operationalizing policy updates.
- This structure scales through the responsible AI champions program embedded across product groups. Working with division-level leaders, this community provides expertise and supports day-to-day implementation of our Standard.
- Our responsible AI leaders also regularly engage with Microsoft's Board of Directors through the Board's Environmental, Social, and Public Policy Committee to enable visibility into progress and integrate responsible AI considerations into broader strategic and risk discussions.

Responsible AI Governance Community



We are committed to strengthening our Responsible AI Governance Community where needed to ensure our readiness to support product groups, business functions, and customers. Over the past year, we have deepened our investment in team members who develop custom solutions, preparing us to help customers customize and deploy AI technologies.

Training and upskilling are a core part of this governance ecosystem, with required and optional courses covering responsible AI with different lenses relevant to engineering, sales, policy, legal, and other internal audiences:

- "Trust Code," our standards of business conduct training, applies to the broadest cross-company audience. Responsible AI is one of the most prominent topics in our latest edition, which 99% of employees have completed.
- In 2025, Microsoft offered over 500 AI-related educational resources for employees. These offerings included a range of learning modalities, from live learning sessions to recorded training to written reference materials.
- In 2026, Microsoft updated our internal responsible AI learning program by introducing role-based learning paths. These learning paths provide consistent, centralized responsible AI education across a range of topics tailored to different roles and responsibilities across the company.
- Since January 2026, Microsoft's Machine Learning, AI, & Data Science (MLADS+) program has delivered over 45 learning sessions related to responsible AI, reaching more than 4,000 attendees across these sessions.

We are committed to listening to our community. Over the last year, we implemented updates to our whistleblower policy and process, including to address AI concerns. Our goal is to support employee confidence in reporting AI-related issues, reinforcing transparency and trust in how Microsoft addresses AI risks.

Together, our principles, policies, practices, tools, people, and processes form the foundation of Microsoft's AI governance program across the NIST RMF map, measure, and manage functions. They not only set expectations but enable teams to meet them with consistency, at scale, and in ways that can evolve as AI technologies and their operating environments change.

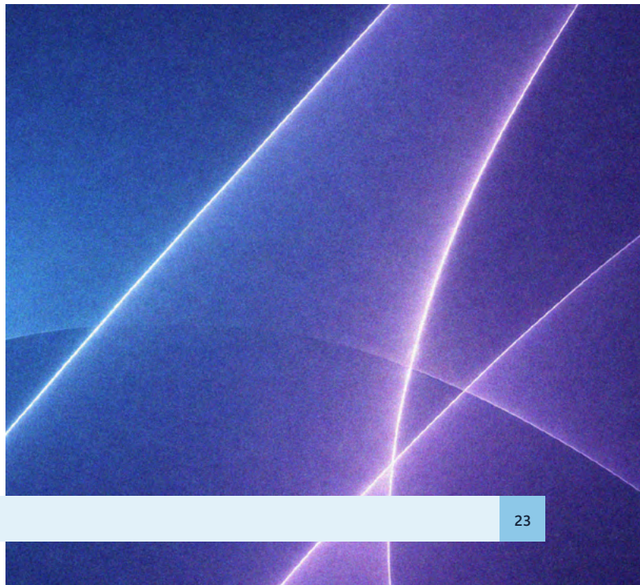
Map: Identifying risks

The steps we take to map and prioritize risks are about understanding the role of AI technologies we are building and cataloging risks that need measurement and mitigation. Over the last year, we advanced these practices in three key ways:

- Aligning on AI technology terminology and inventory management through our re-engineered Responsible AI Standard.
- Advancing our risk taxonomy and integrating it with other trust domains.
- Investing in new practices and tools to support rigorous and at-scale implementation of risk mapping.

AI technology terminology and inventory management

First, our re-engineered Standard aligns how we define and scope AI technologies, including AI models, AI platform services, and AI applications—three categories that structure our Developer Chapter. For each category, the first "Core Requirement" is to "know" the AI technology, which means determining capability integration or use scenarios that trigger additional risk assessment and mitigation. Along with broader ongoing work on our internal tooling, these steps build from existing AI model and service inventories and facilitate component reviews across the AI tech stack. Ultimately, they support clearer understanding of where AI is used and which risks matter most—so that workflows for risk assessment and mitigation can be built in from the outset.



Risk taxonomy

Second, we integrated practical taxonomies for known and emerging risks to improve coordination. This centralized, dynamic catalog of legal and policy risks—tied to how Microsoft builds, deploys, and uses AI—supports required risk assessments that strengthen Microsoft's AI risk management and compliance. Our risk taxonomy tracks dozens of risks across a number of categories, including security, privacy, and psychosocial risks.

This central risk tracking also correlates with Microsoft's Vulnerability Severity and Content Classifications for AI Systems (i.e., AI Security & Safety Bug Bar), which defines categories and severity levels for common AI security vulnerabilities and content-related issues.[24] Over the last year, as new classification categories were launched within the Bug Bar, our aligned taxonomies and tracking processes facilitated more seamless coordination as we assess, mitigate, detect, and respond to risks, issues, and incidents pre- and post-release.

Implementation practices

Third, as agentic AI systems become more capable, Microsoft has developed and scaled implementation of a new generation of threat modeling practices tailored specifically to agentic AI. These practices account for dynamic agent behavior, tool orchestration, memory, and evolving interaction patterns over time.

The updated threat modeling framework introduces structured methods to identify and assess risks such as tool misuse, permission escalation, memory poisoning, and cross-agent interference. Over the past year, training sessions helped prepare product teams to proactively surface and mitigate risks. This work has also been instrumental in shaping Microsoft's broader approach to agent governance, which is strengthening readiness for scalable deployment of agentic systems.

In addition, we have continued to invest in practices to ensure a strong feedback loop in how we identify, catalog, and prioritize AI risks. A mix of practices, from threat modeling to AI red teaming, customer feedback, incident response, and external research, helps strengthen our understanding of which risks are most relevant to which AI technologies and deployment scenarios. As our understanding of the risk and threat landscape continues to evolve, our re-engineered Standard and agile update process also facilitate operationalizing those learnings at scale.

Measure: Assessing risks and mitigations

Measuring AI risks—an iterative process of pre- and post-release assessments of risks and mitigation effectiveness—sits at the heart of a strong AI risk management process. Over the past year, we:

- Increased our focus on strong internal evaluation, including distributed product team assessments and centralized adversarial testing and systematic measurement.
- Prioritized collaboration with external experts, including where risks are novel or benefit from at-scale investment.

Product team assessments

Our product group-specific efforts are grounded in our re-engineered Standard, which includes risk assessment requirements for different AI technologies and product scenarios. As part of this update, we provided teams with simplified guidance for baseline evaluations of risks and the impact of mitigations as well as clarity regarding steps to address elevated risks at the outset or as capabilities could emerge. Our centralized tooling supports tracking and documentation of assessments.

Centralized internal evaluation

Our centralized teams and processes play a critical role in risk evaluations for high-impact AI technologies and use scenarios, including more capable general-purpose models, generative AI, and where our Sensitive Use criteria are met. Over the last year, we advanced our adversarial testing capabilities; boosted the scope and rigor of our systematic measurement, with an emphasis on psychosocial, offensive cyber, and insufficient robustness, reliability, or accuracy (i.e., product malfunction); and increased coverage for evaluations of agentic systems.

Our AI Red Team conducts human-led and tool-based adversarial testing for novel, high-impact technologies and risk scenarios. In comparison to last year, our AI Red Team testing increased 1.5x, executing more than 100 operations. The AI Red Team also deepened their expertise on novel attack vectors—such as adversarial memory poisoning, where a system is manipulated into remembering false information, and techniques to detect backdoors in open-weight LLMs.[25]

We also centrally manage systematic measurement of risks and mitigations of generative AI technologies. Over the last year, we made advancements by investing in research and tools to help scale evaluations, including:

- **Open-source benchmarks:** We have onboarded additional open-source benchmarks to an internal benchmarking platform. Through this effort, we have expanded benchmarks covering a broad range of capabilities and risks, including:
  - Capabilities for biology research
  - Capabilities for agentic AI
  - AI companionship
  - Fair representation of people
  - Hazardous content
  - Non-compliant behaviors
- **Tools for developing automated evaluations:** In addition to increasing evaluation coverage for capabilities, quality, and metrics, we have invested in science and tooling to improve our process of developing new evaluations. We created an evaluation framework that translates risk definition into measurement techniques. With policy and domain experts in the loop, it automates risk systematization, targeted query generation, and annotation and metrics calculation. These investments enable us to scale evaluation coverage while advancing the state of the art in measurement.

Advancing evaluation science and practice—with real-world benefits

Our approach to developing automated evaluations is grounded in work by the Microsoft Research Sociotechnical Alignment Center (STAC) to advance the scientific foundations of generative AI evaluations through a social science measurement framework. The framework defines how an issue is measured and whether measurements are valid, enabling evaluations of AI capabilities, behaviors, and risks to produce results that are more reliable, interpretable, and actionable. STAC has also identified significant validity gaps in commonly used red teaming metrics and in LLM-as-judge evaluations (where one AI system assesses the outputs of another AI system), showing that common approaches can lead to unreliable conclusions about system safety.

In 2026, we introduced the Adaptive Spec-driven Scoring for Evaluation and Regression Testing (ASSERT) open-source framework, building on the STAC framework to help turn natural-language behavior specifications into valid, executable evaluations. Altogether, this research and practical implementation advance evidence-based tools to support more reliable measurement of AI technologies built on Microsoft's platforms and across the broader ecosystem.[26]

External expert evaluation

To complement internal adversarial testing and systematic measurement of risks, we also collaborate with external firms and government research institutes to assess our AI technologies. Over the last year, we have worked with ten firms that have general AI risk and red teaming expertise, strengthening our capacity for in-depth assessments. Additionally, we partnered with eight firms and government research institutes with deep subject matter expertise in high-severity risk areas, including biological, chemical, and cyber weapons.[27] Going forward, through an External Red Team Alliance (EXTRA), we are also building a distributed network of researchers, practitioners, and regional specialists who can contribute directly to red teaming in areas where specialized expertise is needed, helping advance AI safety evaluation and testing practices.

To facilitate external evaluations in collaboration with government research institutes, Microsoft has put in place a Memorandum of Understanding with the US Center for AI Standards and Innovation (CAISI) and a Collaboration Agreement with the UK AI Security Institute (AISI).[28] Through these collaborations, qualified external evaluators with national security expertise can help examine how our AI models and systems behave under challenging and unexpected conditions, including attempts to misuse or bypass safeguards, and generate actionable feedback to strengthen risk mitigation. These collaborations complement evaluation research partnerships with US CAISI, UK AISI, and institutes in Australia and Singapore as discussed in Section 4.

Manage: Mitigating risks

At Microsoft, we implement a layered approach to mitigating AI risks—combining technical safeguards to reduce risks and transparency practices to inform how technologies are deployed and used. These mitigations are tailored to the specific characteristics and risk profiles of different AI technologies, including models, platform services, and applications. Over the past year, we enhanced prompt injection defenses, expanded our product guardrails for agents, and strengthened our transparency documentation.

Product guardrails and features

We have continued to invest in secure-by-design defenses and layered mitigations to reduce risks across the AI lifecycle and address emerging agent risks. Key examples include:

Secure AI development workflows. We improved the secure and reliable handling of access credentials and sensitive data in AI development workflows, including through identity and access management controls, automated scanning, and other zero-trust enhancements to Microsoft's broader security infrastructure.

Expanded classifier coverage and prompt injection defenses. We broadened the scope and precision of classifiers used to detect and block disallowed content and defend against prompt injection, including indirect or cross-prompt injection attacks (XPIA). Broader prompt injection defenses include system prompt hardening; "spotlighting" to distinguish trusted from untrusted content (a technique developed by security researchers as discussed in Section 4 and available as a feature in our Foundry platform);[29] and fine-grained permissions and access controls that mitigate the impact of any successful XPIA.[30]

AI transparency features. We advanced transparency in user experience design and readiness to apply marking and attach provenance metadata to AI-generated content across relevant product surfaces. Our layered implementation strategy pairs C2PA-based secure provenance metadata[31] with watermarking as complementary mechanisms, with central solutions to help support latency stabilization, production hardening, and monitoring at scale.

Governance infrastructure for agents. As AI systems become more agentic—planning across multiple steps, invoking tools, and taking actions across connected systems—effective risk mitigation increasingly depends on the governance infrastructure that supports those systems in operation. Over the past year, Microsoft has expanded our capabilities to authenticate, constrain, and monitor agent behaviors. Examples of our investments include enhanced capabilities to:

- Identify agents and assign them permissions;
- Define and enforce runtime policies to govern which agents can invoke which tools, access which resources, and perform which actions under specified conditions;
- Evaluate agent behavior pre- and post-deployment;
- Maintain observability and audit logs; and
- Preserve meaningful human oversight, including to:
  - Configure authorization and policy settings, such as approval gates;
  - Activate, block, or delete agent instances; and
  - Review evaluation results and logs through user interfaces.

Human approval gates add a layer of control for sensitive or irreversible actions. In user-facing experiences, these controls are surfaced directly: Browse with Copilot lets users see actions in real time and interrupt or take control; Copilot Cowork makes actions visible in the conversation and requires explicit approval for sensitive actions before execution. These approaches help ensure that as agents perform more consequential actions, people retain the ability to supervise and intervene as well as remain accountable.

Implementing Model Context Protocol

Microsoft implements Model Context Protocol (MCP) for agent tool use and is an active contributor to its core maintainer leadership community. We integrate MCP with a defense-in-depth security strategy: MCP standardizes agent access to tools and data while layered security controls—such as strong identity and authorization, scoped permissions, and secure execution environments—enhance security in interactions. Critically, MCP interactions are also instrumented so that agent actions and tool calls are observable. This combination supports agentic systems that are both powerful and enterprise-ready, with safeguards that facilitate reliability and accountability.

Transparency documentation

Transparency remains one of Microsoft's core AI principles and an area in which we have demonstrated compliance leadership. By clearly explaining the capabilities, limitations, and intended uses of AI models, platform services, and applications, we empower downstream developers, deployers, and end users to make informed decisions and have context for compliance as well as appropriate reliance.

Over the last year, Microsoft refreshed our data summary and model card templates to incorporate learnings and address regulation, including California's AB 2013 requiring training data transparency and the EU AI Act General-Purpose AI (GPAI) Code of Practice Transparency Chapter. Ahead of AB 2013's effective date, Microsoft published data summaries for in-scope models, including MAI-Image-1 and Phi-4-Reasoning. For MAI-Image-1, which qualified as a GPAI model under the EU AI Act, we were also among the first providers to publish both a Data summary and Model card under the GPAI Code of Practice Transparency Chapter. We have continued to publish documentation as new models, such as MAI-Image-2, have been released, and periodic internal review by a cross-functional policy, legal, and engineering team facilitates ongoing compliance.[32]

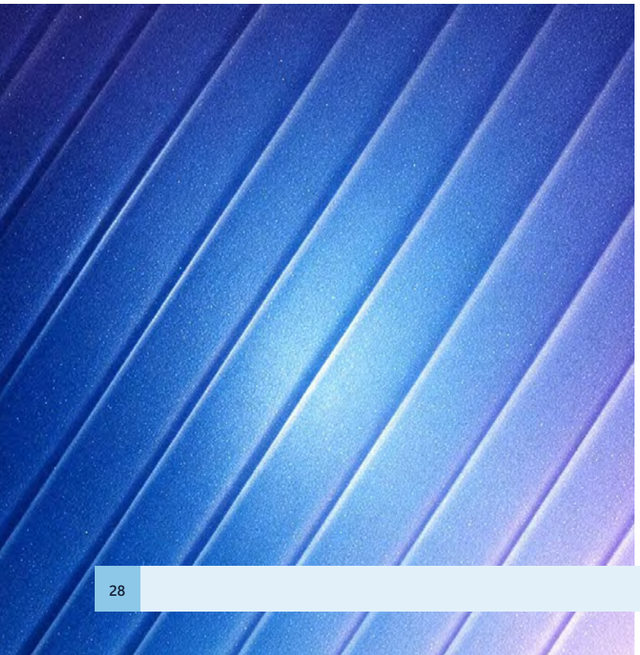
Beyond new data and model card templates, Microsoft has scaled our documentation program across the AI tech stack, launching new templates for AI applications and platform services and beginning to implement them in 2026. The templates were shaped not only by evolving regulatory requirements but also by partner research (see Section 4) and customer interest in concrete guidance on responsible AI use. As a result, new platform and application cards provide clearer context on technology behavior, capabilities, limitations, safety components, and governance practices. Since January 2026, we have published 20 application and platform cards that reinforce shared responsibility between Microsoft and customers by clarifying our role in the responsible development and deployment of AI technologies along with actions that customers can take to deploy and use them with trust.

Release: Reviewing decisions to make AI technologies available

Throughout 2025 and 2026, we continued to invest in pre-release oversight for generative AI technologies as well as products and features that trigger review through our Sensitive Uses program. We made progress in ensuring our pre-release review keeps pace with emerging capabilities. We also applied these processes across models, platform services, and applications—as described by case studies that illustrate our approach in practice.

Generative AI technology reviews

From July 2025 to June 2026, nearly 2,500 generative AI cases were submitted to receive guidance from experts across teams focused on governance and our responsible AI community. Through tools and training, we simplified processes to tailor risk management routes while implementing quality control to strengthen rigor. Specifically, we updated the rules that inform how we categorize risk levels, enabling employees across Microsoft to provide answers to a small set of fact-based questions and then be automatically routed to refined sets of relevant requirements—while implementing new quality control processes to strengthen confidence that release assessments are appropriately tailored. These refinements both support non-experts in understanding the initial risk profile of their technologies and help right-size requirements to the characteristics of the launch.



Sensitive Use reviews

Our Sensitive Uses and Emerging Technologies program provides oversight for and strengthens our readiness to evaluate risks of high-impact uses of AI. In implementing this program, the Office of Responsible AI works closely with product engineering as well as customer and partner solutions teams to provide guidance on the trusted development and deployment of AI technologies where foreseeable use or misuse triggers one of the following criteria:

- A consequential impact on an individual's legal status or life opportunities, e.g., systems that prioritize access to healthcare services.
- Significant physical or psychological injury, e.g., workplace safety alert systems in industrial environments.
- The potential to restrict, infringe upon, or undermine the ability to realize an individual's human rights, e.g., computer vision-based systems deployed in public spaces that could restrict rights to assembly.

As regulatory frameworks reflect the concept of "high-risk" AI applications, Microsoft has focused on translating our criteria and legal requirements into concrete, operational policies and practices that teams can apply consistently in real-world development and deployment decisions. We are working to rationalize regulatory concepts, covering areas such as essential infrastructure and public services, healthcare, law enforcement, mental health, and biometric identification, into a shared internal understanding of elevated risk and corresponding safeguards. The Sensitive Uses and Emerging Technologies program plays a central role in this approach by serving as an early indicator for identifying AI technologies that may be scoped into high-impact categories.

Since 2019, the Sensitive Uses team has processed over 1,900 submissions, including 450+ from July 2025 to June 2026. Of the 450+ cases submitted during that period:

- More than 30% of cases were related to Computer Use Agents (CUA), agentic AI experiences, or autonomy in productivity contexts.
- Nearly 30% of cases were custom solutions developed for customers.
- Of the cases that were built as custom solutions for specific Microsoft customers, nearly 21% were related to healthcare and life sciences, and 10% were related to financial services.

Emerging Technologies program

In 2025, the Office of Responsible AI launched the Emerging Technologies (EmTech) program to proactively examine frontier AI capabilities—such as memory and Computer Use Agents—and identify emerging risks that may not yet be addressed by formal policy. The program's objective is to provide early steer responsible AI guidance, grounded in Microsoft's AI principles, to support teams developing cutting-edge AI technologies during periods of rapid innovation and regulatory uncertainty. Our re-engineered Standard simplifies how relevant guidance may inform evolving core or scenario-specific requirements.

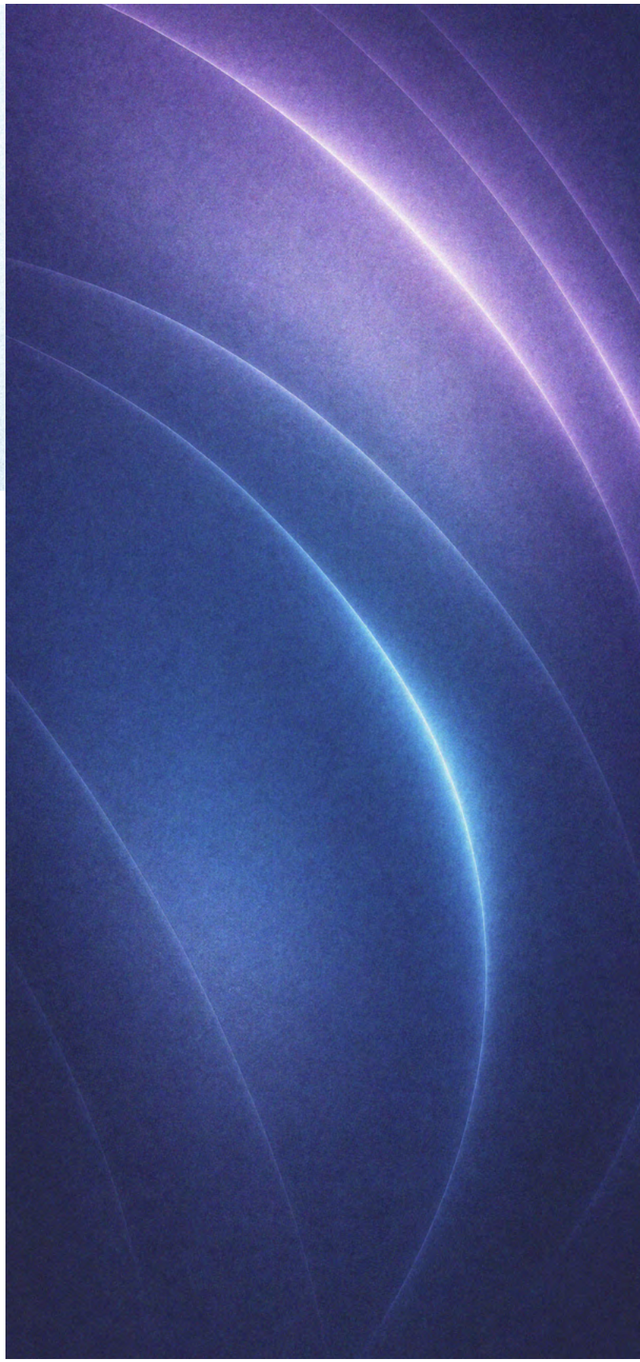
Applying pre-release reviews in practice: AI technology case studies

To illustrate how our risk management and pre-release review processes work in practice, the following case studies from across the AI tech stack—available in this Report or online—highlight how teams assess products, features, and deployment scenarios according to our Responsible AI Standard. Product and assessment teams validate that risks have been appropriately mapped, measured, and managed through an iterative process pre- and post-release.

1. Models: Case studies at the tech stack's model layer

At Microsoft, we build and modify general- and specific-purpose models, as well as proprietary and open-weight models. Over the last year, examples include:

- MAI-Thinking-1, a proprietary model designed to support a range of capabilities such as content generation, reasoning, and task assistance.
- MAI-Image-1 and MAI-Image-2, proprietary text-to-image models; Image-2 is integrated into Bing Image Creator and made available for customers to integrate into applications via Foundry.[33]
- Fara-7B, an open-weight, ultra-compact model designed specifically for computer use.



Case study

MAI-Image-1 and Image-2: Building safety in layers

MAI-Image-1 and MAI-Image-2 are text-to-image AI models that generate images from written descriptions. MAI-Image-1, launched in fall 2025, produces high-quality, realistic images that closely align with user prompts, including detailed scenes, lighting, and composition. MAI-Image-2, launched in spring 2026, builds on this foundation by further improving image quality, visual detail, and accuracy. Microsoft applied a defense-in-depth approach to safety across the model development, release, platform deployment, and application integration lifecycle, tracing the layers of safety applied across this lifecycle.

Map

The model development teams identified key risks in focus for MAI-Image-1 and Image-2 as content depicting explicit violence and nudity or invoking child safety concerns.

MAI-Thinking-1 general-purpose model

Microsoft MAI-Thinking-1 is a medium-sized reasoning model designed to support a range of scenarios. It can be integrated into products and services to enable capabilities such as content generation, reasoning, and task assistance. Given the broader range of scenarios, as part of mapping and relative to MAI-Image models, development teams identified a broader range of key risks in focus for measurement and management processes. These included harmful content, agentic risks, and cyber misuse risks. As part of a responsible release process, MAI-Thinking-1 was initially made available through a private preview with a select group of customers, followed by broader testing and public preview release in August 2026.

Measure and manage

Layer 1: Training data curation

From the outset, the teams embedded responsible AI in data curation decisions. During training, teams applied safety-focused filtering to the underlying datasets, reducing the model's exposure to harmful content such as explicit violence or child sexual exploitation and abuse imagery (CSEAI). By removing these categories from training data, the models were less likely to learn detailed or realistic representations of such content. This early intervention helped establish safer baseline behavior.

Layer 2: Model evaluation

The models then underwent multi-stage safety evaluations designed to assess potential risks, identify gaps in mitigation, and iteratively strengthen safeguards prior to deployment. Lessons learned from Image-1 were applied and built upon for Image-2. These processes combine adversarial testing, large-scale image generation for both manual and automated review, and targeted tuning of automated safety systems.

Image-1

Step 1: Testing the raw model. The evaluation of Image-1 began with testing the raw model without any external safety guardrails enabled. In this configuration, like nearly all image generators without guardrails in place, the model was able to produce a limited set of harmful or misaligned outputs when prompted with toxic or adversarial inputs—especially when those inputs are crafted as "jailbreaks," which are adversarial prompts designed to manipulate the model into bypassing its internal safety guardrails (coded messages, unusual phrasings, etc.). This baseline testing confirmed that, without additional safeguards, the model could generate limited examples of content inconsistent with Microsoft's requirements.

Step 2: Enabling text and image safety filters. The team enabled text and image safety safeguards, including automated classifiers that screened both inputs (user prompts) and outputs (generated images) for policy violating content. Overall, the filters demonstrated strong baseline performance, while the remaining edge cases highlighted areas where additional refinement was needed. Our prompt enhancer, which interprets and enriches user prompts and frames them in a way that the generator is best able to ingest, adds a layer of risk mitigation.

Step 3: Testing using internal adversarial datasets. To assess real-world behavior, the team evaluated model outputs with safety filters enabled and identified a small subset that warranted closer review. These cases were generally mild or ambiguous rather than clearly harmful, but they highlighted opportunities to strengthen existing safeguards. In response, additional safety detectors were added to the safety stack to improve coverage for edge cases and reinforce existing text and image filtering.

Step 4: Testing using external benchmarks. The enhanced safety stack was then tested using a public benchmarking platform. The updated safeguards blocked additional images, and the remaining cases were highly ambiguous—for example, prompts with unclear instructions or where the resulting image was not unsafe.

Step 5: Focusing on child safety. Finally, the team conducted a focused review of child safety risks and identified one minor risk (that children could be generated in contexts where they are implied but not explicitly asked for—e.g., "people jumping on a trampoline"). To reduce the frequency of depicting children and the likelihood of depicting children in potentially concerning situations, the sensitivity threshold was adjusted. This conservative adjustment increased blocking behavior, including some false positives, but was deemed appropriate given the heightened risk profile associated with child safety.

Evolving the evaluation approach

Modern AI systems are inherently probabilistic, making comprehensive and consistent safety testing difficult. While relying exclusively on human experts to manually test for every potential harmful outcome is neither scalable nor sustainable, structured templates that automatically generate test scenarios must also reflect the evolving environment. Following the launch of Image-1, the Microsoft AI (MAI) team built on learnings to develop a scalable, product-specific safety evaluation platform designed to systematically identify and address safety vulnerabilities and complement centralized testing. This platform supported the launch of Image-2 in spring 2026 and contributes to our expanding library of centralized and product-specific safety tests.

Image-2

Image-2 built directly on the safeguards and mitigations implemented for Image-1 and used the same baseline safety stack. Image-2 also underwent a manual review of more than 2,000 generated images produced during private benchmarking. Across this dataset, only a very small number of images were identified as potentially harmful, and these occurred infrequently. Again, the observed cases were ambiguous rather than clearly unsafe—for example, a fully clothed child crying while holding a toy. Further testing showed that such outputs appeared sporadically, indicating random variation rather than predictable failure modes.

Importantly, all sexualized or exploitative content involving children was blocked by the safety system, and no systematic patterns of policy violation were identified. Based on these findings, the empirical risk profile for Image-2 was assessed as low and consistent with Image-1, supporting continued use of the existing safety mitigations.

Layer 3: Platform deployment and application integration safeguards

Beyond model-level evaluation, additional safety measures were applied as Image-1 and Image-2 moved into real-world deployment—through platform availability on Foundry or integration into applications such as Bing Image Creator. Each context introduced its own safeguards, reflecting the different ways developers and users interact with models.

Platform deployment safeguards

Before deploying MAI-Image-2 on Foundry in public preview, the team conducted centralized evaluations to address common risks associated with image generation, including the potential creation of harmful or other disallowed content. Based on the results of these evaluations, the team implemented additional safeguards, such as classifiers to filter problematic user requests and model outputs, protections designed to prevent misuse, and stricter content filtering for imagery involving children. Foundry also provides transparency into model catalog capabilities, including access to model cards, filtering by modalities such as text-to-image, and insight into different deployment options.[34]

Application integration safeguards

When integrating the models into consumer-facing applications, teams applied tailored safeguards to the product context and user experience. These measures operate across four areas: evaluating inputs, steering models, assessing outputs, and providing transparency.

Evaluating inputs. A foundational component of application-level safety is the use of evaluators—assessment tools that review model inputs for potential risks, such as policy-violating content, with platform-level guardrails in place, including classifiers. Evaluators help teams systematically identify where model behavior is clearly safe, clearly problematic, or falls into more ambiguous "gray areas," providing a consistent way to measure risk and inform downstream decisions during and after application integration.

Steering models. Insights from evaluators inform the use of system-level steering mechanisms—safeguards designed to guide how the model responds when high-risk or uncertain scenarios are detected. Rather than relying solely on hard blocks or refusals, steering mechanisms, such as system-level messages, influence model behavior by encouraging more cautious responses, reinforcing existing safety protections, or redirecting interactions within predefined boundaries. These mechanisms act as backstops that complement the model's baseline safety filters, helping improve coverage for edge cases and reducing the likelihood of harmful or unintended outcomes while preserving product functionality.

Assessing outputs. In addition to input- and model steering-focused safeguards applied at the platform and system levels, teams also applied AI image judges to evaluate model outputs. Image judges are automated, platform-level assessment mechanisms that review generated images after they are produced, scanning for disallowed content.

Providing transparency. The team provided information about models integrated into Bing Image Creator, including each model's capabilities and what users can expect when selecting them. They provided guidance on how to use the application, craft effective prompts, and report harmful content. The team also explained how responsible AI informed development, along with the consequences of violating policies.[35]

Release and post-release monitoring and refinement

At every release checkpoint in this case study, assessment teams reviewed evaluation and risk mitigation results. Across both platform deployments and application integrations, safeguards also operate as part of a continuous improvement loop. Teams monitor real-world performance, analyze patterns in flagged inputs and outputs, and iteratively refine protections over time. Findings from platform-level evaluations, such as emerging risk patterns identified during Foundry deployment, inform updates to application-level controls—and vice versa. This cross-context feedback cycle helps ensure that safety improvements compound in deployment, supporting responsible use of the models wherever they are accessed.

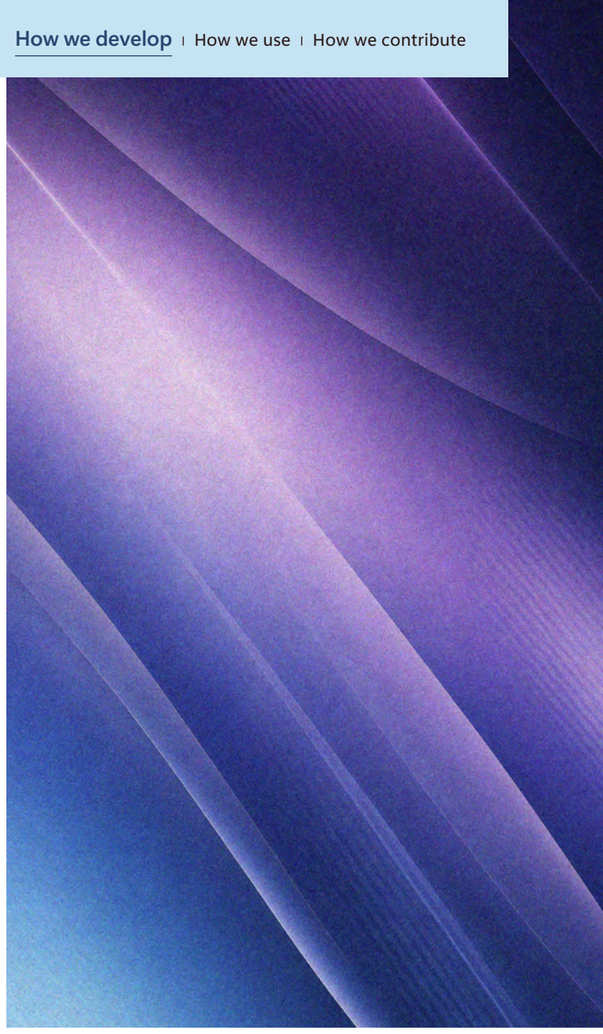
2. Platform services: Case studies at the tech stack's platform services layer

At Microsoft, we provide platforms that enable customers to build, modify, and manage AI models, applications, and agents. One example from the past year is Foundry Agent Service, which has integrated capabilities allowing developers to create and manage long-term memory for AI agents.

3. Applications: Case studies at the tech stack's application layer

Microsoft provides a wide range of applications. This includes general- and specific-purpose applications that may integrate a combination of first- as well as third-party models and that are made available for consumers or enterprise customers. Over the last year, examples of applications—or agents or features within applications—include:

- Copilot Cowork, an AI agent within M365 Copilot that carries out tasks across a user's M365 enterprise environment.
- Copilot Health, an experience where users can bring health questions, wearable data, and health records together for proactive insights and guidance that are personalized and secure.
- Browse with Copilot, an AI feature within Edge enabling agentic AI to perform web-based tasks.
- M365 Copilot with memory capabilities, which enables the AI assistant to retain information about users across conversations to provide contextually relevant responses.



Case study

Copilot Cowork: Addressing agentic capabilities in the M365 environment

Copilot Cowork is an AI agent that carries out tasks across a user's M365 enterprise environment. Users describe what is needed using natural language, and Cowork executes actions. These actions could include sending emails through Outlook with attachments or scheduling meetings and managing calendars. Cowork will check in before executing certain actions, and the user must explicitly approve them.

Because of Cowork's agentic capabilities, it qualified for a Sensitive Use review under Microsoft's Responsible AI Standard.

Map and measure

Working in conjunction with Microsoft's AI Red Team, the Sensitive Uses team identified jailbreaks and cross-prompt injection attacks (XPIA), harmful content generation, and accuracy and robustness as key risks associated with agentic systems like Cowork.

To evaluate these risks, the team developing Copilot Cowork conducted custom evaluations to better understand Cowork's ability to detect and avoid generating harmful outputs; its resilience against jailbreaks and XPIA attacks; and its ability to adhere to tasks and remain grounded in relevant content. The results showed a higher degree of resistance across these risk areas while maintaining performance.

Manage

To address mapped and measured risks and iteratively improve evaluation results, the team implemented multiple layers of safety mitigations in addition to those already present in the M365 safety stack. For example, every action that sends or modifies content requires explicit user approval before execution if the user did not already give permission in the initial instructions. Further, Cowork operates entirely within the user's existing M365 environment's permissions and does not access additional data or capabilities beyond what the user already has.

Cowork includes guardrails that prevent it from generating harmful content, and it will refuse requests to evaluate individual employee performance or profile people based on protected characteristics. Like most AI agents engaging with web-based content, Cowork can be exposed to XPIA attacks. To mitigate this risk, Cowork treats external data with caution and flags suspected injection attempts to the user. Outputs include attribution indicating AI assistance, ensuring recipients know when they are interacting with AI-produced content.

Release

Copilot Cowork became available through Microsoft's Frontier program in late March 2026 and moved to general availability in June 2026.[36] This phased approach limited exposure as the team collected feedback to continue improving the experience, allowing them to identify real-world issues and validate that safety mitigations perform as intended before broader availability. In addition, the team published an application card to inform users on Cowork's capabilities, intended uses, limitations, and best practices for leveraging the tool responsibly.[37]

Case study

Copilot Health: Adapting to a more sensitive deployment context

Over 50 million health questions are asked every day across Microsoft's consumer products, including Bing and Copilot. Questions reflect interests in everything from better understanding symptoms to finding the closest care provider. Whether they're tracking wellness, searching for health tips, or managing daily routines, our users are increasingly turning to conversational AI to help them live healthier lives.[38]

In response, the Microsoft AI Health Team developed Copilot Health, a dedicated space within Copilot where users can bring their health questions, wearable data, and health records together for insights and guidance that are personal and secure.

Prior to deployment, the Microsoft AI Health Team mapped, measured, and managed risks associated with using generative AI in the context of health.

Map

Given the sensitive nature of health information, a Sensitive Use review of Copilot Health was conducted. Working closely with the Office of Responsible AI, the product team identified potential risks associated with using generative AI in health contexts, including (1) responses not grounded in credible health sources and (2) conversational chats that lacked the context necessary to answer a user's question accurately and safely.

The risk of providing responses not grounded upon credible health information sources was a particular concern. Grounding responses in, and pointing users towards, credible health sources is what allows people to trust the information they receive and to explore it further. For example, when a user asks about a health symptom, the response is far more valuable when it draws on credible, well-established health sources, such as Harvard Health, compared to an unknown personal blog post.

Additionally, the risk of not seeking further user information when necessary was a concern because users often omit pertinent details in their initial query, such as the specific nature of their symptoms or any relevant medical history. Without the full picture, even a factually correct answer may not be useful for a user's specific context. For example, a user who describes seemingly innocuous calf pain may not mention that they had recently traveled out of the country on a long-haul flight—critical details that would let Copilot Health provide a more relevant and safer response.

Measure and manage

From the outset, the team embedded our AI principles in the design and development of Copilot Health to proactively mitigate risks. This included using large-scale automated testing, co-designed by clinical and safety experts, to ensure that responses are safe, accurate, and helpful in health-related scenarios. In parallel, over 250 licensed clinicians from over two dozen countries and members of the public undertook human evaluations to assess the product's clarity, tone, helpfulness, and overall safety. The team also undertook targeted red-teaming exercises to probe edge cases and potential failure modes. Further, the team engaged Microsoft's responsible AI, disability, and accessibility teams early in development to ensure the experience is inclusive and accessible to users of all abilities.

Moreover, to address the specific risks of not grounding answers in credible health sources, the team applied a health source credibility framework, independently published by the National Academy of Medicine, to proactively surface highly credible information sources. In addition, for more than 500 of the most common health conditions and treatments, Copilot Health surfaces answer cards authored by Harvard Health during the user's conversation.[39]

Additionally, to manage the risk of not seeking further user information when required, the team designed Copilot Health to ask, with user consent, clarifying follow-up questions, as informed by clinical best practices. This approach enables Copilot Health to gather missing or ambiguous details like symptom duration or severity, recognizing that health conversations are inherently context-specific and that complete information is necessary to deliver accurate, helpful, and safe answers.

Release

Before releasing Copilot Health, the team conducted user research to obtain product feedback from various groups of different socio-economic backgrounds in order to better understand user expectations and safety considerations and promote user trust.

Subsequently, in May 2026, Copilot Health was made available to select US users aged 18 and older, reflecting the demographics of the group with which the team conducted pre-release user research. This phased rollout was designed to enable the team to systematically gather further user feedback and iteratively improve the overall user utility of the product.

Post-release mapping, measuring, and managing

Once AI technologies are made available to customers and users, Microsoft maintains post-release monitoring, incident response, and learning mechanisms to enhance visibility and understanding of whether systems are operating as intended and to adapt to evolving risks.

Post-release monitoring

We expanded our monitoring capabilities to strengthen real-time detection of emerging risks and to inform continuous improvement through logging and observability, detection innovation, and behavioral insights.

Logging and observability. We implemented new AI-focused requirements within Microsoft's Security Development Lifecycle (SDL) and integrated AI systems into observability platforms. These platforms support privacy-reviewed, post-release monitoring, incident response, and auditability across the AI stack.

Detection innovation. Our security team improved privacy-reviewed automated detection capabilities, including a new system that accelerates investigation and detection of advanced threats in M365 Copilot logs. The system supports real-time and retrospective analysis of multi-turn jailbreak attempts, prompt injection, memory poisoning or misuse, and other emerging risks. Faster cross-product analysis of threats helps us detect issues before they escalate.

Behavioral insights. We also invested in research to better understand how users interact with AI systems in practice. For example, the Copilot Usage Report 2025 analyzed 37.5 million aggregated and deidentified conversations to uncover patterns in how people use Copilot across devices and times of day.[40] This research revealed how Copilot is integrated as part of users' lives—supporting both professional and personal tasks and surfacing expectations for responsible design.

Incident response

As AI systems become more capable and widely deployed, rapid and coordinated incident response is increasingly essential. In 2025, we scaled AI incident response capabilities to meet emerging and anticipated demands.

Standardized playbooks. We formalized incident response playbooks across 13 topics, supporting teams from Microsoft Security Response Center (MSRC) and product and legal groups in following a unified process when AI incidents arise. These playbooks are reinforced through tabletop exercises and readiness drills.

Novel issue investigation. Our security response capabilities were tested in response to a novel AI-related issue that revealed a potential attack vector. External researchers found that attackers could exploit Copilot to access files without those access points being recorded in Unified Audit Logs, a centralized record that captures user and administrator activities across M365 services. However, the activity was captured in our enhanced logging and data platform, enabling Microsoft to conduct a large-scale forensic investigation—analyzing over 90 million security logs to assess scope and identify potentially impacted customers. Based on this investigation, we did not identify evidence of customer harm or confirmed exploitation associated with the reported behavior. The underlying logging issue was remediated, restoring expected audit log coverage for Copilot interactions.

AI-driven response infrastructure. Since our novel issue investigation, our enhanced logging and data platform has been used to streamline threat-hunting queries—targeted searches used to proactively identify threats—improving the speed and efficiency of investigations involving AI systems. These enhancements reflect a broader transformation in Microsoft's incident response posture. AI-driven capabilities are enabling stronger support for forensic and threat-hunting teams across cloud environments and global regions, improving both proactive and reactive investigation.

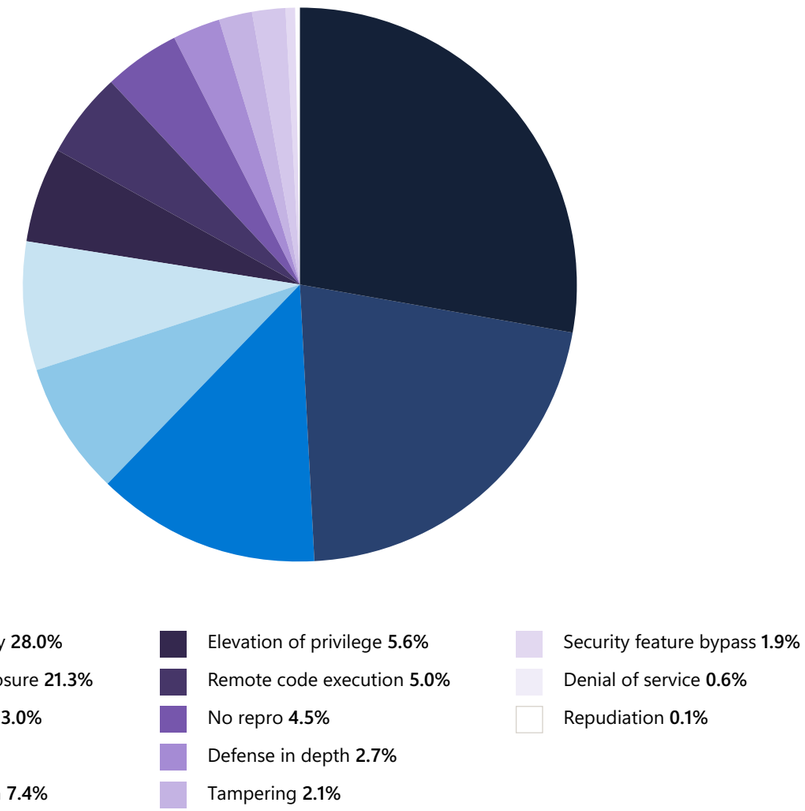
Learning from issues

We treat every incident, vulnerability, or reported or discovered issue as an opportunity to learn and improve. In 2025, the MSRC, which is tasked with investigating reported or discovered issues, received over 11,000 submissions. Of these submissions, more than 600 escalated to full AI safety or security cases, as depicted in the chart below. The most frequent subjects of these submissions were malicious use; chemical, biological, radiological, and nuclear (CBRN) risks; and command injection.[41] Prompt injection remained frequently reported, but with notable growth in new surfaces, especially coding-assistant agents and agentic code generation workflows. There was also a rise in multimodal prompt injection techniques using audio, images, QR-encoded inputs, and other novel formats, reflecting broader model capabilities and researcher experimentation. Insights from these cases informed updates to guardrail tuning, especially in Copilots involving user-generated content or cross-surface prompts.

2025 trends reveal that 28% of reported cases were classified as "not a vulnerability," reflecting ongoing learning about what falls within our "Bug Bar" scope. This exchange both helps clarify boundaries and signals the importance of discussions and partnership with our researcher community. Issues in other cases reinforce priorities for layered defenses and responsible AI development to strengthen resilience.

Together, these post-release capabilities give Microsoft the ability to detect, investigate, and respond to AI issues at cloud scale—reinforcing our governance rigor and our commitment to continuous learning.

Total cases by impact in 2025[42]



Section 3

How we use AI responsibly

At Microsoft, our commitment to responsible AI extends beyond how we develop, release, and operate AI technologies to how we use them. Strong governance should be grounded in a deployer's use case as well as its organizational practices to evaluate and manage risks and to support end users.

As AI increasingly influences how work is performed, responsible use requires attention to human judgment, role evolution, and skills. Workers and managers need technical proficiency and contextual understanding to determine appropriate and responsible use—keeping people at the center of how we deploy and govern AI systems and use them to expand human capability.

Adapting our Responsible AI Standard

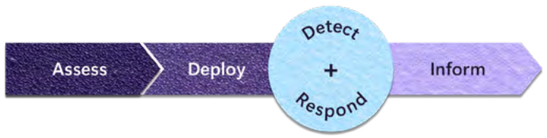
Our approach to responsible use is now embedded in Microsoft's re-engineered Responsible AI Standard. While the Developer Chapter, as introduced in Section 2, applies to first-party AI technologies, for the first time, our Standard also includes a Deployer Chapter. The Deployer Chapter establishes requirements for when Microsoft deploys third-party AI applications, either for use by Microsoft employees or by external users through Microsoft web properties and products. This structure reflects differences between deploying first- and third-party AI applications, including available context on responsible development practices.

The Deployer Chapter of Microsoft's Responsible AI Standard is organized around four objectives: assess the underlying AI application, deploy trustworthy AI applications, detect and respond to issues, and provide information to support responsible deployment. In implementing these objectives, Microsoft works independently and, where relevant, with the original developer. For example, developers and deployers may work together to detect and respond to issues, sharing context on how technology is designed and where risk emerges.

Responsible AI Standard

Developer Chapter	Deployer Chapter
Four core objectives for developing and deploying first-party AI models, applications, and platform services	**Four core objectives for deploying third-party applications**
1 | Know the AI model, platform service, or application	1 | Assess the underlying application
2 | Build for trustworthiness	2 | Deploy trustworthy AI applications
3 | Detect and respond to issues	3 | Detect and respond to issues
4 | Inform customers and users	4 | Provide information to support responsible deployment

Key steps in AI deployment governance



Assess the underlying AI application

Responsible deployment begins with a clear understanding of the use case and the application's intended use, which is informed by the original developer's documentation in support of the AI application. Specifically, under our Responsible AI Standard Deployer Chapter, teams assess whether the application qualifies as a Sensitive or Restricted Use and developer documentation of intended uses, capabilities, limitations, supported contexts, and potential misuse scenarios. While deployments in restricted categories cannot proceed, Sensitive Uses undergo review to determine whether additional risk management steps are required (Section 2: Release).

In reviewing the underlying AI application, teams may also assess:

- How the application marks or applies content provenance to AI-generated or modified content;
- Mitigations for disallowed self-harm content;
- The application's post-deployment monitoring and feedback mechanisms—such as logging, performance tracking, robustness metrics, and user feedback channels; and
- Whether the application supports responsible user interaction, such as enabling human oversight.

In addition to assessing the application itself, teams may consider how the application could shift workflows or roles. By identifying shifts in responsibility or reliance early, teams can reinforce human judgment and set clear expectations.

Once a deployment plan is in place, teams document key information—such as intended uses, operating context, and known limitations—in an internal AI asset registry. Clear documentation about AI capabilities and limitations informs appropriate reliance.

Deploy trustworthy AI applications

Building from the process of assessing the application and documenting key information, the next step in the Deployer Chapter is to evaluate risks that may arise from how the AI application will be used in practice. These assessments consider both intended uses and reasonably foreseeable misuse, including risks related to insufficient accuracy, reliability, or robustness. As appropriate for identified risks, teams evaluate their likelihood and potential impact, implement appropriate mitigation measures, and document residual risk along with the rationale for proceeding.

Deployments involve steps supporting responsible human–AI interaction, such as providing users with information about the application's appropriate uses, designing experiences so users interpret outputs correctly, and implementing safeguards discouraging inappropriate reliance. To support effective human oversight, guidance is provided for operators responsible for monitoring system behavior, detecting anomalies, and intervening when necessary. Oversight mechanisms also include the ability to override or halt system operations and appropriate training for personnel responsible for supervising deployed AI applications.

Detect and respond to issues

As with development, responsible deployment does not end at launch. To reduce risk, AI applications may be introduced through pilots and phased releases internally to validate performance and allow for continuous iteration based on early feedback.

Under the Deployer Chapter, teams implement appropriate post-deployment monitoring methods such as product analytics, logging, and user feedback channels. These mechanisms help teams identify potential anomalies and detect issues requiring investigation or mitigation.

Teams also document an incident response plan and, where feasible, coordinate closely with the developers of underlying AI applications to address vulnerabilities or emerging issues. Signals from deployment—such as incidents, user feedback, or performance changes—are assessed and, where appropriate, communicated back to developers.

Inform users to support responsible deployment

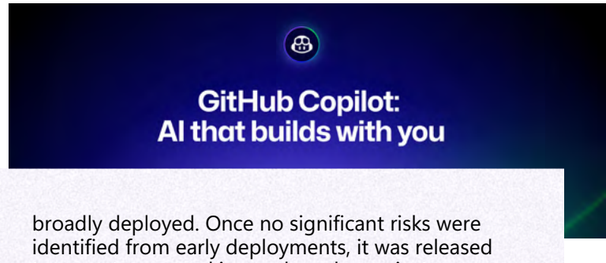
Finally, Microsoft emphasizes transparency about how AI applications are deployed. Teams inform users about an AI application, including its intended uses, its limitations, and the safeguards implemented to support responsible operation.

How does the Responsible AI Standard's Deployer Chapter enhance transparency and accountability, two of Microsoft's foundational AI principles?

- Makes deployment decisions more visible and reviewable.
- Provides more clarity to support human accountability at deployment.
- Standardizes information expected across deployments.
- Enables greater traceability from assessment to real-world use.
- Strengthens oversight and regulatory readiness.

AI deployment governance in real-world use cases

We apply our Responsible AI Standard when deploying AI applications across engineering and business operations. Although the Deployer Chapter currently applies to third-party AI applications, the following examples highlight how Microsoft implements consistent practices when deploying first-party AI applications (subject to requirements in the Developer Chapter) for employee use or on Microsoft web properties that support customers.



Case study

GitHub Copilot

Accelerating software development

Many Microsoft engineers use GitHub Copilot as part of their development workflow.[43] This internal deployment provides a helpful learning ground for responsible AI practices in software engineering environments.

Context and value

GitHub Copilot supports developers across the entire software development lifecycle, from writing and reviewing code to finding and fixing issues. Recent advances in agentic capabilities are also expanding AI-assisted workflows. Early in the generative AI era, internal studies revealed that developers using GitHub Copilot completed certain coding tasks 55% faster while also reporting higher satisfaction and improved focus on complex engineering work; with Copilot's agent mode, they can now perform complex multi-step tasks such as analyzing entire codebases, generating multi-file updates, executing tests, and proposing bug fixes.[44]

Responsible AI deployment practices

- **Pre-deployment assessment:** Prior to broad deployment, GitHub Copilot's capabilities, limitations, and appropriate development contexts were evaluated.
- **Gradual and phased deployment:** GitHub Copilot was piloted with a few teams before being more broadly deployed. Once no significant risks were identified from early deployments, it was released to more teams working on broader projects.
- **Risk-based mitigations:** For more sensitive code, additional mitigations were implemented, such as limiting the specific AI models used.
- **Human oversight:** Human oversight is integrated into the development process through mechanisms such as review and approval of AI-generated code suggestions, targeted review of high-risk changes, or automated checks with human escalation. Standard engineering safeguards—including pull request reviews, branch protections, and automated testing—help ensure that AI-generated code is subject to appropriate review before deployment.
- **Security integration:** Microsoft's secure development practices, including static analysis and vulnerability scanning through tools such as CodeQL and GitHub Advanced Security, apply to GitHub Copilot workflows. These safeguards help ensure that AI-assisted development maintains the same security expectations as traditional development workflows.

Governance insights

Operating GitHub Copilot at enterprise scale provides valuable insights into how AI tools influence developer workflows. Embedding periodic reviews, monitoring, and continuous learning into our responsible deployment practices informs updates to engineering guidance, secure development practices, and responsible AI as applied to developer tools.

Case study

Copilot and Dynamics 365 Customer Service

Modernizing employee support with AI

Microsoft's Human Resources (HR) organization deployed Copilot capabilities within Dynamics 365 Customer Service to improve internal employee support.

Context and value

HR teams manage a high volume of employee inquiries covering benefits, policies, and operational processes across many countries. Prior to this deployment, resolving complex cases often required manual research across multiple knowledge sources.

Copilot capabilities introduced AI-assisted case summaries, automated drafting of responses, and a unified HR knowledge base spanning 120 countries. The deployment has helped improve operational efficiency and service quality. Microsoft HR teams reported a 20% increase in case throughput, while adoption among advisors reached 72% monthly active usage.

Responsible AI deployment practices

- **Responsible handling of sensitive data:** HR scenarios involve highly sensitive employee information. The deployment incorporated strong data governance controls and restricted access policies to ensure that AI outputs reflect appropriate data protection practices.
- **Human decision authority:** HR advisors remain responsible for reviewing and finalizing responses generated with Copilot assistance. This helps ensure that sensitive communications are handled appropriately.
- **Training and change management:** Microsoft invested in training and internal communication to support HR teams in understanding both the capabilities and limitations of the application. Context on Microsoft's AI principles was integrated into this rollout to help employees use the application appropriately.
- **Operational monitoring:** The HR team tracks adoption, satisfaction, and performance metrics to continuously refine workflows and improve system performance.

Governance insights

Deploying AI in internal enterprise support functions highlights how responsible AI practices must account for sensitive data environments, employee trust, and organizational change management.

Case study

Copilot Studio – Ask Microsoft web agent

Deploying scalable AI agents for customer experiences

Microsoft's marketing organization uses Copilot Studio to build and deploy AI agents that guide customers through product support, discovery, evaluation, and purchase journeys across Microsoft web properties.[45]

Context and value

The Ask Microsoft web agent, built on Copilot Studio, helps customers across Microsoft's commercial product ecosystem by answering questions related to product support, selection, pricing, and licensing. The solution uses a multi-agent architecture in which specialized agents provide domain-specific guidance across offerings such as Azure and Microsoft 365.

Deployment of the Ask Microsoft web agent has improved responsiveness and reduced latency while helping customers find relevant information more quickly. It has also reduced the volume of customer chats that require human support while increasing product signups.

Responsible AI deployment practices

- **Gradual deployment and experimentation:** Teams used staged rollouts and A/B testing to evaluate application behavior and performance under real-world traffic conditions, enabling gradual expansion while maintaining the ability to pause or roll back changes.
- **Operational monitoring:** Engineers use dashboards to track latency, errors, usage patterns, and user feedback. These signals help identify performance issues and guide ongoing improvements.
- **Grounded responses and transparency:** The application provides links to source material so users can verify responses and explore additional information. Teams developing and deploying the Ask Microsoft web agent also run automated groundedness and safety evaluations, leveraging Foundry evaluation tools,[46] and halt deployment if quality bars are not met.
- **Customer-centric design:** The application integrates with human support channels, enabling seamless escalation to live agents when needed.

Governance insights

This deployment demonstrates how AI applications can be introduced safely through phased releases, monitoring, and human fallback mechanisms to support robust and scalable customer experiences.

Case study

Azure AI Search

Improving enterprise knowledge access with AI

Microsoft's internal Licensing Navigator illustrates how AI applications can support complex knowledge retrieval tasks while maintaining high standards for accuracy and transparency.[47]

Context and value

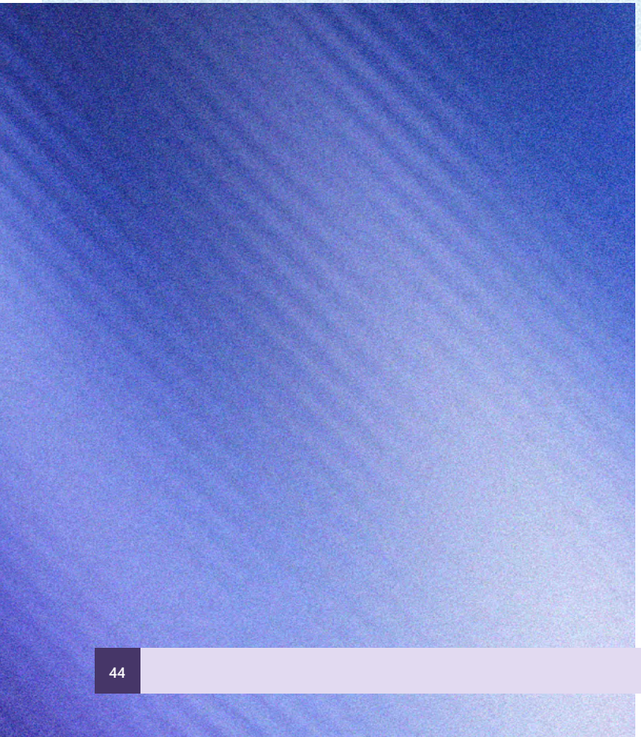
Licensing questions are often complex and require consulting multiple contractual and informational resources. Previously, employees could spend significant time researching answers. The Licensing Navigator agent uses Azure AI Search and Copilot Studio to retrieve and synthesize information from the Microsoft Product Terms, Services Provider Use Rights, and over 300 additional curated Microsoft web pages, enabling employees to generate accurate responses significantly faster. The application now answers more than 1,000 questions per week for hundreds of employees across roles, including sales, support, and operations. It also supports partners through the Partner Center AI assistant.

Responsible AI deployment practices

- **Curated and trusted knowledge sources:** The team carefully selected authoritative documentation to ground the system's responses, reducing the risk of inaccurate or unsupported answers.
- **Retrieval-augmented generation (RAG):** The architecture retrieves relevant content segments and uses them to generate responses, ensuring outputs are grounded in verifiable documentation.
- **Continuous accuracy improvement:** The team regularly refines content and model configuration based on user feedback and performance analysis.
- **Source transparency:** Responses include "deep links" to the specific relevant content, such as clauses in Microsoft Product Terms, making it easier for users to validate the information provided using public content from Microsoft.

Governance insights

The Licensing Navigator demonstrates how content governance, retrieval grounding, and transparent citations can significantly improve reliability and trust in AI-assisted knowledge systems.



A governance feedback loop

Using AI responsibly provides an important feedback loop for Microsoft's broader AI governance program and informs the guidance we provide to customers. Operational insights from deployment help refine risk assessment practices, shape updates to the Responsible AI Standard, and guide research and engineering investments across the company.

Our experience as a deployer also reinforces our commitment to earning and sustaining trust through responsible AI practices—not just in what we build, but in how we use it. By governing AI use across technology, people, and processes, Microsoft aims to scale AI in ways that are trusted by users, resilient to risk, and aligned with long-term value creation.

Section 4

How we contribute to and learn from the ecosystem

Responsible AI depends not only on strong internal governance but also on advances in science, ecosystem infrastructure, customer enablement, and collaboration across researchers, industry, governments, civil society, and affected communities. At Microsoft, we see advancing the capabilities of customers and contributing to the broader responsible AI ecosystem as part of acting responsibly ourselves. Over the last year, we built new agentic AI governance capabilities across our platforms and through open-source frameworks, enabling customers to develop and deploy agents with greater confidence. In addition, we advanced research where evidence is still emerging, strengthened collaboration where there are common risks and opportunities, and helped translate new knowledge and community insights into shared practices, tools, and technical standards.

Advancing research and technical progress

One of our central priorities is strengthening the underlying evidence base for how increasingly capable AI technologies behave, where they can fail, and how they can be designed to better support people. This spans evaluation of AI model capabilities, development of new techniques to detect misuse threats, and study of how real-world use of AI technologies shapes human reasoning and judgment. In 2025 and 2026, Microsoft advanced research on:

- **The Hawthorne effect in reasoning models:** Researchers discovered reasoning-focused large language models (LLMs) change behavior when they detect they are being evaluated rather than deployed in real-world contexts. Results revealed models were more likely to comply with harmful requests when prompts were framed as hypothetical or consequence-free.[48]
- **Detecting backdoored models at scale:** Researchers explored techniques to identify backdoored language models, where hidden behaviors are embedded in a model during training. These behaviors activate when the model encounters a specific trigger phrase (e.g., “deploy”). Researchers found backdoored models may activate even if fragments of the phrase are detected (e.g., “dep”) and developed a scanner to help identify potential triggers at scale.[49]
- **Identifying and mitigating new threat classes:** Researchers identified a new class of threats, which they termed “AI recommendation poisoning.” This is when malicious or opportunistic actors attempt to manipulate AI outputs by injecting hidden instructions into user interactions through specially crafted links—often embedded in “summarize with AI” buttons. For example, prompt-based commands can instruct an AI system to persistently favor certain sources, products, or services in future recommendations without the user’s awareness.[50]
- **Designing for augmented cognition through generative AI:** Researchers examined how generative AI impacts critical thinking in real-world knowledge work. Results revealed that the use of AI to complete a task is associated with reduced perceived enactment of critical thinking. In response, the Microsoft Tools for Thought team developed strategies for leveraging generative AI in ways that augment cognition rather than replace it. Research on augmented cognition is important for that reason: it focuses not only on what AI can automate, but also on how AI can be designed to support reflection, creativity, and better decision-making.[51]

Taken together, this research reflects a common objective: improving the scientific and practical foundations of measuring the reliability and impact of AI technologies. Our research investments also help responsible AI grow stronger as an evidence-based discipline that can adapt as AI capabilities, deployment patterns, and risk profiles evolve.

This progress is especially important as AI systems become more agentic and interact directly with tools, external content, and multi-step workflows. In those contexts, risks such as indirect prompt injection, data exfiltration, and unintended actions can emerge in new ways. Microsoft has invested over the last year in efforts to enhance defenses, including through:

- **TaskTracker:** In collaboration with academia, researchers created TaskTracker, a new technique for detecting indirect prompt injection attacks on LLM-powered systems by inspecting the internal activation states of an LLM.
- **A public dataset of attack prompts:** Microsoft hosted the Adaptive Prompt Injection Challenge, focused on identifying novel prompt injection attacks against a simulated AI-powered email client. More than 800 participants submitted over 200,000 attack prompts that were compiled into a public dataset to support continued research on prompt injection defenses.[52] This kind of work strengthens not only our own safeguards but also the field's broader ability to understand and address new attack surfaces as AI technologies become more connected and more autonomous.

Government research partnerships

Over the last year, Microsoft has expanded our partnerships with government research institutions worldwide to strengthen the science of evaluating and securing AI systems and enhancing their resilience.

US Center for AI Standards and Innovation: Microsoft entered into a Cooperative Research and Development Agreement (CRADA) with the National Institute of Standards and Technology (NIST) and a Memorandum of Understanding (MoU) with the Center for AI Standards and Innovation (CAISI). These collaborations focus on improving methodologies for evaluating the resilience of AI technologies under real-world adversarial pressure. They aim to improve how the field tests AI systems for issues such as prompt injection risks, resistance to "jailbreak" attempts, and whether systems follow safety-related instructions in high-stakes areas.

UK AI Security Institute: With the UK AI Security Institute (AISI), Microsoft entered into a Collaboration Agreement, which includes a research track focused on enhancing cybersecurity and the reliability of agentic AI systems.

Australia AI Safety Institute: Microsoft entered into an MoU with the Australian Government to engage in collaborative research with Australia's AI Safety Institute on human-AI interaction risks. Such risks include overreliance and emotional dependency on AI companion chatbots and conversational systems.

Singapore Infocomm Media Development Authority: With Singapore's Infocomm Media Development Authority (IMDA), Microsoft entered into an MoU to support research collaboration on projects related to agentic AI security, multilingual safety across Southeast Asian languages, and trusted access to frontier models.[53]

Supporting and scaling the broader research ecosystem

Beyond direct research collaborations, Microsoft supports the global research community through academic partnerships and programs to fund research and support access to models and infrastructure.

Agentic AI Research & Innovation Program: In 2025, Microsoft launched the Agentic AI Research & Innovation (AARI) program, building on the Accelerating Foundation Model Research (AFMR) initiative.[54] AARI supports twelve research projects exploring topics such as multi-step reasoning, improving reliability and uncertainty handling, enabling autonomous and resilient supply chains, and cultural misalignment in AI agents.

External Red Team Alliance: In 2026, Microsoft launched the External Red Team Alliance (EXTRA), a global initiative that expands support for independent AI safety and security research. EXTRA provides unrestricted funding to 18 university labs across six continents, supporting researchers investigating emerging questions in AI safety and security and helping strengthen research capacity in regions and disciplines with wide-ranging expertise.

Trustworthy AI Research Fellowship for early-career scholars: Over the past year, Microsoft supported the inaugural cohort of the Computing Research Association's Trustworthy AI Research Fellowship, investing in early-career scholars to help strengthen capacity for responsible AI research. Interdisciplinary fellows from computing and the social sciences receive structured training, mentoring, and sustained peer engagement and contribute to shared frameworks and language for trustworthy AI. The inaugural cohort advanced research that informs AI governance, evaluation, and deployment practices in real-world contexts.[55]

Research from Microsoft's new AI Economy Institute

In 2025, Microsoft launched the AI Economy Institute to study how AI can diffuse responsibly across education systems and labor markets. Multidisciplinary research cohorts[56] examine how AI adoption is reshaping teaching, learning, workforce preparation, and access to opportunity. Microsoft aims to leverage insights from this work to help inform evidence-based strategies for preparing students and workers for an AI-enabled economy.

US National AI Research Resource: Microsoft supports the US National AI Research Resource (NAIRR) Pilot by providing access to Azure infrastructure, AI models, and technical expertise. Last year, Microsoft and NAIRR launched a Deep Partnership track, supporting "grand challenge" scientific research projects, where Microsoft provided researchers with access to Microsoft Discovery, an agentic research acceleration platform. Supported projects announced in early 2026 span fields including materials science, sustainability, semiconductor design, and biomedical research.[57]

Frontier Model Forum biosafety research grants: Over the last year, Microsoft has also supported research addressing the biological risks of advanced AI systems through grantmaking efforts led by the Frontier Model Forum (FMF). These grants fund "wet-lab" research (i.e., hands-on laboratory experiments) that explore biological risk scenarios and evaluate safeguards that complement computational and policy-based approaches. By contributing alongside other FMF members, Microsoft helps support independent, expert-led research that strengthens preparedness and informs practical risk mitigation strategies for high-consequence biological threats.

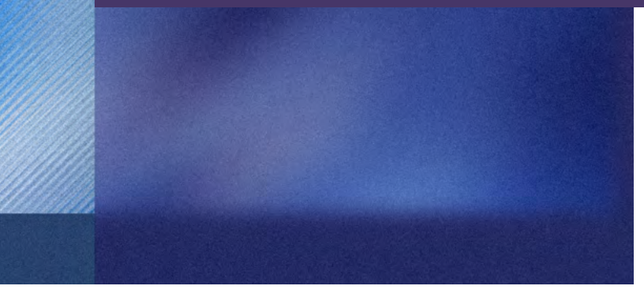
Building shared practices, tools, and standards

As research advances an evidence-based foundation for implementing responsible AI, shared practices, standards, and tools help translate research into consistent ways to assess and verify effective governance across the ecosystem. We actively contribute to industry and multistakeholder forums to rapidly develop the reference points and tools needed to define practices and enable interoperability across a globally interconnected value chain.

Best practices for advancing trustworthy AI

Over the last year, we have contributed to developing consensus best practices for frontier risks and agentic AI with industry peers and multistakeholder communities.

- **Frontier Model Forum (FMF)** brings together frontier AI developers and deployers. With FMF partners, we have defined best practices for evaluating and mitigating risks related to biosecurity, cybersecurity, and harmful AI system behavior.[58] In addition, FMF members have leveraged an information sharing agreement announced in 2025 to advance shared knowledge relevant to effective risk mitigation as well as further develop state-of-the-art practices.
- **Partnership on AI (PAI)** brings together industry, civil society, and academia. Microsoft contributed to its September 2025 report, Prioritizing Real-Time Failure Detection in AI Agents,[59] which examines how AI systems' risk profiles have evolved and recommends layered approaches to robust failure detection.



Tools for demonstrating trustworthy AI

We have likewise contributed to developing shared tools for AI transparency and measurement with industry peers and multistakeholder communities.

- **Hiroshima AI Process Reporting Framework** version 2.0 was developed by an OECD-led informal task force involving contributors from industry and the research community. In 2025, Microsoft contributed a voluntary report under version 1.0 along with others.[60] Relative to version 1.0, version 2.0, released in May 2026, streamlines reporting, expands and clarifies applicability across the AI value chain, and integrates new topics like agentic AI.[61] More than 50 organizations across sectors and geographies have pledged to provide a voluntary report under version 2.0, contributing to transparency and defining clearer shared expectations for what effective AI governance looks like in practice.
- **MLCommons** is an engineering consortium that works with industry, academia, and civil society to build tools for AI measurement. Over the last year, Microsoft stepped up our support for MLCommons projects to advance reliability measurement,[62] including:
  - AILuminate's multilingual, multicultural, and multimodal expansion that is building from the ground up to increase support for languages and culturally adapted framing—with progress underway for 6 locales in Asia—enabling more reliable measurement of AI safety and security beyond Western English, starting in Indic and East Asian contexts.[63]
  - The Jailbreak benchmark that quantifies how a safety system degrades under adversarial attack, using a consistent "resilience gap" metric. It reflects a shift from static safety testing toward stress testing AI systems under increasingly realistic and adversarial conditions.[64]
  - The Psychosocial Task Force, which aims to develop benchmarks covering multiple risk areas, such as crisis prevention and unqualified AI advice.

Standards for driving interoperability in trustworthy AI

As best practices and tools help build common understanding and support consistent and comparable implementation of AI governance, standards and specifications play complementary roles to enable and reinforce interoperability and trust. Some standards establish technical governance infrastructure that functions as a shared foundation for responsible innovation; others align the trust architecture that supports broader assurance needs. Over the last year, across key issues, we have contributed to industry and multistakeholder efforts to advance both types of standards, the organizations that facilitate their development, and how they are used for impact.

- **Governing agentic AI systems:** As AI systems increasingly operate through networks of agents that can access tools, interact with other agents, and act on behalf of users, there is a growing need for common approaches to interoperability, observability, security, and governance. The Agentic AI Foundation (AAIF), a Linux Foundation initiative, was established to advance open standards and shared infrastructure for agentic systems, anchored by founding contributions like Model Context Protocol.[65] It is complemented by initiatives such as OpenTelemetry (OTel), a Cloud Native Computing Foundation graduated project, which maintains an open, vendor-neutral observability standard that defines how telemetry data—traces, metrics, and logs—is generated, structured, and exported across distributed systems. We contributed to an agentic extension of OTel, enabling correlation across model calls, tool invocations, and downstream services and helping advance it as a practical foundation for tracing, understanding, and monitoring complex multi-step and multi-agent workflows.[66]
- **Establishing content provenance:** The ability to understand the origin and history of digital content is increasingly important as generative AI makes it easier to create, modify, and distribute synthetic media at scale. Industry collaboration is essential to ensuring provenance information can be created, preserved, and verified across platforms and services. The Coalition for Content Provenance and Authenticity (C2PA) continues to bring together technology and media organizations to advance an open, interoperable standard for verifying the origin and history of digital content, including whether it is AI-generated or -modified. Over the last year, C2PA made progress advancing the interoperability and real-world applicability of content provenance, updating Content Credentials and committing to support adoption and accessibility across additional languages.[67]
- **Operationalizing AI assurance across the value chain:** As organizations seek practical ways to assess and demonstrate that AI systems are trustworthy, there is growing need for assurance approaches that are credible, scalable, and interoperable across sectors and jurisdictions. The Appia Foundation, a Linux Foundation initiative, brings together companies from across the AI value chain and regions to help identify and address gaps in the global AI assurance architecture and advance coherence among standards, conformity assessment approaches, and evidence models.[68] The MLCommons AILuminate Global Assurance Program[69] aims to contribute reusable evidence, translating benchmark results into operational assurance through capabilities such as Benchmarking-as-a-Service, consistently defined objects of evaluation, and reliability labels. Together, these initiatives help progress consistent, interoperable, and applied assurance pathways that can inform governance, procurement, and compliance across AI supply chains, consistent with international standards such as ISO 42001.

Equipping developers and customers

As organizations across industries adopt AI technologies, customers and partners increasingly seek practical guidance and assurance that systems can be deployed responsibly and in alignment with evolving regulatory expectations. Microsoft invests in industry-leading compliance, with a broad and growing portfolio of certifications, as well as tools and training to help developers and customers innovate with AI while strengthening governance, security, and accountability.

Leading assurance across our ecosystem

Microsoft has contributed to the development of foundational international standards for AI governance, including ISO/IEC 42001:2023, the first global management system standard for AI.

Microsoft has achieved ISO 42001 certification across a broad set of AI systems, including:

- Microsoft 365 Copilot
- Microsoft 365 Copilot Chat
- Microsoft Foundry
- Microsoft Security Copilot
- Microsoft Copilot Studio
- GitHub Copilot
- Dragon Copilot
- Dragon Copilot (Radiologist)
- Copilot Health[70]

Microsoft is also one of the only companies to certify under ISO 42001 AI systems spanning productivity, developer tools, healthcare, security, and platform services through a single, converged audit framework operationalized over the last year. This marks a significant milestone in scaling responsible AI governance across diverse product environments.

Earning trust at the forefront of compliance leadership

To support our customers in a rapidly evolving environment, Microsoft engages proactively on regulatory developments and makes clear compliance commitments. In July 2026, we signed the EU Code of Practice on Transparency of AI-Generated Content. Over the last year, we have also been among the first companies to test our interpretation of novel requirements. Ahead of the effective date for California's AB 2013 *(Generative Artificial Intelligence: Training Data Transparency Act)*, we completed data summaries for in-scope models. Likewise, we were among the first providers to release a general-purpose AI (GPAI) model under the EU AI Act and GPAI Code of Practice and to publish a data summary and model card under the Code's Transparency Chapter.

To stress test our re-engineered Responsible AI Standard and strengthen our ongoing readiness as a compliance leader, we analyzed over 100 enacted and pending regulations. We aimed to validate our Standard's architecture amidst evolving external expectations so that it proves more durable for our product engineering and customer solutions teams.

Delivering technical tools in our platforms and open source

In addition to investing as a leader in certifications and compliance, Microsoft directly enables developers and customers to build, evaluate, and operate AI systems responsibly through technical capabilities and tools delivered open source and in our platforms (e.g., Foundry and Copilot Studio). In delivering these capabilities and tools, we also integrate and align with best practices, tools, and standards where applicable.

Our technical capabilities and tools support customers and developers in risk identification, testing, and monitoring throughout the AI lifecycle—from pre-deployment evaluation to continuous post-deployment assurance.

As AI systems become more agentic, effective governance increasingly requires a technical stack that can establish which agent is acting, constrain what it is allowed to do, and make its behavior observable and continuously assessable over time. Over the last year, Microsoft has expanded these capabilities across our platforms and open-source contributions to help developers and organizations build, deploy, and govern agentic systems with greater rigor.

In assessing behavior of AI over time, robustness against adversarial misuse is especially important to measure. Layered defenses and evaluation—enabled by our latest research and tooling—can help ensure greater reliability even as attack patterns continue to evolve.

Identity and runtime governance for agents

Entra Agent ID: To govern agentic systems, developers must be able to bind actions to a distinct agent identity and apply enforceable policies at runtime. In April 2026, **Entra Agent ID**, a framework for designing, securing, and governing AI agent identities in Microsoft Entra, became generally available. It allows organizations to plan agent deployments, manage credentials, enforce access policies, and monitor agent activity in a more consistent way.[71] Entra Agent ID uses agent identity blueprints as templates for groups of similar agents, letting organizations define required settings, permissions, metadata, and governance controls up front so agent instances inherit a common security posture rather than being created ad hoc. It also helps organizations apply lifecycle and security controls, such as unique identities per agent, Conditional Access policies, least-privilege permissions, sign-in and audit logging, and ongoing monitoring for anomalies or orphaned agents.

Agent Control Specification (ACS): In June 2026, we announced the **Agent Control Specification (ACS)**, an open, vendor-neutral specification for applying runtime governance across the agent lifecycle, independent of framework, runtime, or policy engine.[72] ACS is designed to provide a portable, auditable policy layer so that controls do not need to be reimplemented separately in prompts or application code. It allows organizations to define deterministic controls at key checkpoints in an agent's lifecycle—such as before model calls, before tool execution, and before outputs are returned—so that policies can be versioned, reused, and enforced more consistently across environments. This helps address a core challenge of agentic systems: the same access token or tool permission may be low-risk in one context and high-risk in another, depending on what the agent has seen, what it is trying to do, and what downstream effects an action could trigger.

Guided Guardrail Setup: Microsoft is providing platform capabilities that help customers operationalize runtime controls in practice. In Foundry, Toolboxes provide a single managed endpoint for tools, skills, MCP clients, and governance to better manage how agents access enterprise data, call tools, and operate inside connected systems. **Guided Guardrail Setup** recommends controls based on an agent's audience, data access, and use case—including protections such as personally identifiable information filters, jailbreak protections, and task-adherence safeguards—helping teams tailor controls to deployment context.[73]

Evaluation and observability for agents

Governing agentic systems requires the ability to evaluate and monitor how they behave—not only before release but also after deployment in changing real-world conditions. This is especially important for agents because risks often arise from multi-step behavior: whether an agent used the right tool, followed constraints across several turns, handled retrieved content appropriately, or drifted from task boundaries over time. As a result, Microsoft has invested in evaluation and observability capabilities that are more specific to agentic workflows and closely connected to production operations.

Adaptive Spec-driven Scoring for Evaluation and Regression Testing (ASSERT): ASSERT, released in June 2026, is an open-source framework for turning natural-language behavior requirements (e.g., ACS runtime controls) into executable evaluations for AI models and agents.[74] ASSERT treats written requirements and policies as first-class inputs to evaluation and, based on an approach developed by Microsoft Research, systematizes those requirements into an inspectable taxonomy, generates test cases, runs them against a target system, and scores outcomes against the behavior or policy statement that produced them. This makes it easier to test whether an agent behaves as intended in context-specific ways—for example, respecting approval boundaries, using tools appropriately, or resisting prompt injection embedded in external content. ASSERT as integrated into Foundry Control Plane can be run pre-deployment and continuously in production, enabling:

- **Pre-deployment benchmarking:** Developers run ASSERT test suites before model deployment, documenting baseline safety and capability measurements;
- **Continuous evaluation:** ASSERT harnesses run automated tests against live agent deployments, detecting drift, safety degradation, and emergent behavior that pre-deployment testing did not anticipate.
- **Reproducible verification:** ASSERT uses versioned test datasets and scoring methods, enabling auditors or third parties to re-run evaluations and verify results.
- **Records:** Results are logged in Foundry with timestamps and test suite versions, supporting ongoing monitoring, compliance reporting, and incident investigation.

In Foundry, developers can also use evaluators to readily assess end-to-end system outcomes and step-by-step behaviors within complex workflows—and tailor evaluation criteria.

- Built-in agent evaluators can measure whether an agent correctly understands a user's intent, follows instructions, selects appropriate tools, and completes tasks efficiently. Additional evaluators assess factors such as tool-call accuracy and whether agents successfully use external data or services to produce reliable results. Together, these tools help developers test the final output of an agentic system and the process by which the system arrives at its result.[75]
- Capabilities such as Rubric help generate evaluation criteria tailored to a specific agent and use case.[76]

Observability is also a key part of this shift. Foundry's hosted agent runtime includes built-in OpenTelemetry (OTel) tracing, and newer tracing capabilities extend observability to agents built on other frameworks as well. This helps create a production-grade trace of what an agent did, which tools it called, what other agent it delegated tasks to, what policies were enforced (e.g., ACS-based runtime controls), or what ASSERT evaluation results were recorded—supporting debugging, performance monitoring, policy review, and ongoing improvement.

Taken together, these investments reflect an important shift in how governance works for agentic systems. Rather than relying primarily on pre-release reviews or high-level instructions to the model, Microsoft is building infrastructure that supports a more continuous lifecycle: define requirements, apply runtime controls, evaluate behavior, observe production performance, and improve iteratively over time.

Robustness to adversarial misuse

To help address risks that models, applications, or agents are manipulated by adversarial inputs and behave in ways that are unsafe or unintended, Microsoft has expanded layered defenses that operate both before and during deployment.

- **Enhanced classifiers and prompt shields** scan inputs and outputs in real time to detect harmful content and prompt injection attempts, including indirect prompt injection.
- **Spotlighting** protects agents from indirect prompt injection attacks by signaling and neutralizing untrusted inputs before they influence model behavior. Developed by Microsoft researchers, Spotlighting helps agents distinguish trusted system and user instructions from potentially adversarial embedded content. This enables agents to detect, surface, and optionally block attempts to override policies or exfiltrate data while preserving task performance when working with third-party data sources.[77]
- **Task Adherence** focuses on action-level alignment at runtime, evaluating whether an agent's planned actions remain consistent with both the user's intent and the developer's intended design before those actions are executed. By analyzing the full interaction context, including user requests, agent reasoning, and proposed tool calls, Task Adherence can detect misalignment such as premature, unintended, or out-of-scope actions. When such discrepancies are identified, it returns structured signals and explanations that developers can use to block execution, log events, or trigger human-in-the-loop review.[78]

By integrating protections that operate both before information reaches the model and before actions are carried out, Foundry supports the development of AI agents that are more reliable, controllable, and trustworthy as autonomy and complexity increase.

Continuous evaluation also helps ensure robustness. In May 2026, Microsoft introduced **RAMPART**, an open-source agent test framework for encoding adversarial and benign scenarios as repeatable tests that can run in continuous integration pipelines, helping teams turn red-team findings and incidents into durable regression coverage. RAMPART builds on Microsoft's earlier open-source contributions to AI safety and security testing, including **PyRIT**, which supports AI red teaming and has also been integrated into Foundry's **AI Red Teaming Agent** workflows.

Together, these projects reflect Microsoft's view that trustworthy agentic AI will depend not only on product-specific safeguards but also on more portable and inspectable policies, evaluators, and traceable records that researchers, developers, auditors, and enterprises can use across the ecosystem.

As agentic AI becomes more capable and more widely deployed, Microsoft's goal is to help customers govern these systems with infrastructure that identifies and applies policies to agents, runtime controls that enforce boundaries, and evaluation and observability tools that make behavior measurable over time.

Ramping up capacity through training

Microsoft invests in training and education programs that help customers and the technology community develop the knowledge and expertise needed to deploy AI with trust, turning lessons learned from real-world operations into reusable patterns. Over the past year, our teams delivered responsible AI training and knowledge-sharing programs that have reached more than 20,000 engineers, policymakers, and customers, helping organizations understand emerging AI risks and practical strategies for secure deployment.

Expanding access to AI security education

In August 2025, the AI Red Team launched "AI Red Teaming 101," a free, self-paced training series designed to introduce security professionals and developers to techniques for evaluating generative AI systems. The ten-episode series explains how generative AI systems operate, introduces direct and indirect prompt injection attacks, and explores strategies for detecting and mitigating vulnerabilities.

Helping ready enterprises to deploy trustworthy AI

Microsoft also provides hands-on briefings and technical workshops for enterprise customers, helping security and engineering teams understand emerging threats. These engagements focus on translating AI principles into operational practices, including logging, monitoring, and layered security controls. In addition, our security teams trained more than 150 Global Black Belts—specialized experts who support enterprise customers in applying Microsoft technologies to complex technical challenges—equipping them to address AI security concerns and share responsible AI best practices across industries and regions.

Investing in people and communities

Microsoft works with educators, journalists, researchers, community organizations, and governments around the world to support AI literacy, strengthen information resilience, and help advance the needs and perspectives of diverse communities as AI develops.

Advancing AI literacy

Microsoft invests in AI literacy to help people of all ages understand how to use and appropriately rely on AI systems. These efforts span education initiatives for young people, training and resources for older adults, and support for journalists and civil society organizations.

Delivering AI education at scale globally

In July 2025, we launched Microsoft Elevate, which aims to reach 20 million people by 2030 with upskilling opportunities ranging from foundational AI fluency to advanced technical training—working across Microsoft, including with LinkedIn and GitHub, and with partners.[79]

Supporting AI literacy for young people

Microsoft's approach to AI literacy for young people both aims to generally equip them to interact with and use AI effectively and responsibly and tailors age-specific resources to address human-AI sensitive interaction risks. Tailored resources supporting child safety build on our longstanding commitments to digital safety, privacy, and responsible AI.

- To support AI literacy in education, Microsoft offers several integrated learning resources through Minecraft Education, a classroom version of Minecraft designed to support curriculum-aligned learning—including Reed Smart: AI Detective, which teaches students how AI works while helping them develop critical thinking skills for evaluating AI-generated content, and AI Adventurers, an animated series introducing similar concepts through short narrative stories designed for younger learners.
- Microsoft created an AI classroom toolkit to help educators introduce generative AI and responsible use. The toolkit supports classroom discussions on fabricated content, privacy considerations, gaps in linguistic or cultural context, and mental wellbeing. Additionally, Microsoft has partnered with the nonprofit Childnet to develop educational materials addressing emerging risks, such as AI-enabled intimate image abuse.[80]
- As AI chatbots and other generative tools become more widely used, ensuring that younger users are prepared to engage with these technologies safely is increasingly important. In 2026, Microsoft co-developed a "by teens, for teens" guide to AI companions, informed by direct youth input. Partnering with Cyberlite, an online safety education leader, we worked with 131 students aged 12–18 across India and Singapore during AI Companionship Youth Co-Design Sessions. We learned that co-designed, hands-on engagements proved incredibly impactful for building AI literacy, resources for parents and trusted adults, and classroom materials for educators.[81]

We are also committed to continuing to learn more about how best to reach youth and address their needs through research and partnerships. Through our support for the Paris Peace Forum, we have been engaged in efforts of the International Research-driven Alliance for AI Serving Every Child (iRAISE), contributing to the publication of *Adolescents & Anthropomorphic AI, Rethinking Design for Wellbeing*.[82]

Partnering to reach digitally underrepresented communities with AI literacy resources

In April 2026, Microsoft announced a partnership with Ampere to advance rural, remote, and Indigenous AI fluency in Canada. This partnership is focused on three key objectives: embed AI skilling into Canada's regional education system; deepen workforce readiness by delivering AI fundamentals courses; and equip 650+ youth and 2,000+ community members with long-term, sustainable digital literacy skills to navigate the AI economy confidently.

In Australia, Microsoft partnered with Deadly Coders—an Indigenous-owned nonprofit providing STEM education to Aboriginal and Torres Strait Islander students—to launch culturally relevant AI and Minecraft Education sessions for 5,000 students in 2025.[83]

Equipping older adults to navigate AI-enabled risks

As AI-enabled fraud and scams become more sophisticated and target older adults, Microsoft has partnered with the American Association of Retired Persons (AARP) and Older Adults Technology Services (OATS), a charitable affiliate of the AARP, to provide support and technical guidance for:

- OATS's AI Guide for Older Adults, which helps people understand how AI technologies are used and how they may affect fraud and scams.[84] Since May 2025, Microsoft has amplified the guide through in-kind advertising across its platforms, generating more than 1 billion impressions.
- OATS's new scam and fraud hub, which includes videos and interactive learning resources to help older adults recognize AI-enabled scams.[85] Within its first month, the hub reached more than 3,000 participants through virtual and in-person trainings.
- AARP's national awareness campaign promoting its fraud helpline during the 2025 holiday season, generating more than 10 million impressions.

Strengthening news and information literacy in the AI era

As AI tools reshape how information is created and shared, Microsoft partners with organizations that support journalists, creators, election officials, and civil society in navigating these changes, including:

- The Elections Group, International Institute for Democracy and Electoral Assistance (IDEA), and Ready for Tuesday to launch the AI Elections Accelerator, a global training initiative for election management bodies to understand the risks and beneficial applications of AI in election administration. The program enabled participants to navigate emerging AI capabilities in alignment with democratic values, election integrity, and public trust—regardless of their technical background—and reached approximately 1,000 election officials across 60+ countries and 40+ US states.
- The Poynter Institute to strengthen AI literacy across both newsrooms and the growing creator ecosystem. The collaboration produced two major initiatives: an AI Toolkit for Newsrooms, launched to more than 430 participants from 50 countries and downloaded more than 1,800 times within its first two months, and a pilot News Creator Training program developed with Microsoft Australia, bringing journalists and digital creators together for programming on how to explain AI technologies.[86]

Centering worker voices in AI adoption

In January 2026, Microsoft Australia and the Australian Council of Trade Unions (ACTU) signed a framework agreement to prioritize skilling and elevate workers' voices in how AI is designed, developed, and deployed across Australian workplaces. The agreement centers on three priorities: sharing information and learning, including AI training developed with the Australian Trade Unions Institute for union leaders and staff; embedding the worker voice in technology development through channels for workers to share experiences that inform how AI systems are built and deployed; and collaborating on public policy and skills to expand upskilling and reskilling opportunities. Building on Microsoft's broader work to put workers at the center of responsible AI diffusion—including its collaboration with the AFL-CIO in the United States—Microsoft and the ACTU convened a first-of-its-kind Workers' Summit in Sydney in April 2026 to advance practical, worker-centered approaches to AI adoption.

Microsoft has also continued to progress our AI labor partnership with the American Federation of Labor and Congress of Industrial Organizations (AFL-CIO), the largest federation of labor unions in the United States representing workers across a wide range of industries, which aims to create an open dialogue to discuss how AI must anticipate the needs of workers and include their voices in its development and implementation. Through this partnership, we have sought feedback from workers and labor leaders to better understand concerns and opportunities arising from AI's use in the workplace and inform how we adapt our approach accordingly.

Supporting users of digitally underrepresented languages

Language remains a major barrier to AI access and adoption for many communities. For billions worldwide, AI systems perform less consistently in the languages they rely on most than in English. To address this challenge, Microsoft announced in February 2026 investments we've made across the AI lifecycle—from data and models to evaluation and deployment—to strengthen multilingual and multicultural AI capabilities.[87]

First, we are investing upstream in language data and model capability through LINGUA Europe and LINGUA Africa—a $5.5 million open call led by the Masakhane African Languages Hub, Microsoft's AI for Good Lab, and the Gates Foundation.[88] Likewise, our AI for Good Lab developed a reproducible pipeline[89] for adapting open-weight large language models to low-resource languages, demonstrating measurable gains for languages such as Chichewa, Inuktitut, and Māori.[90]

Second, we are advancing multilingual and multicultural evaluation tools, including benchmarks and community-centered methods.

- We are co-chairing an MLCommons working group that is expanding multilingual and multicultural capacities of AILuminate benchmarking, with early progress based on languages and cultural contexts from 6 locales across Asia. The working group aims to establish repeatable methods, scale across additional regions, and operationalize research in benchmarking tools, enabling more reliable measurement of AI safety and security beyond English.[91]
- We have advanced Samiksha, a community-centered method for evaluating AI behavior in real-world contexts, in collaboration with Karya and The Collective Intelligence Project in India.[92]
- Through Project Gecko, we have co-designed AI technologies with local communities in East Africa and South Asia to support agriculture, including the Paza family of automatic speech recognition models that can operate on mobile devices across six Kenyan languages, multilingual Copilots, and a Multimodal Critical Thinking (MMCT) Agent that can reason over community-generated video, voice, and text.[93] We also launched PazaBench—the first automatic speech recognition leaderboard, with initial coverage of 39 African languages—and developed two playbooks for multilingual and multicultural capabilities: Paza and Vibhasha.[94]

Third, we are working to scale content provenance for linguistic diversity. With partners in the Coalition for Content Provenance and Authenticity (C2PA), Microsoft is helping extend content provenance standards beyond an English-ready baseline.[95] This includes forthcoming support for multiple Indic languages across metadata, specifications, and UX guidance, alongside efforts to support mobile-first deployment.

Building community-first AI infrastructure

In January 2026, Microsoft announced a new approach to building and operating US AI datacenters that explicitly prioritizes the needs of local communities. Recognizing that large-scale AI infrastructure can place pressure on electricity grids, water resources, and local services, Microsoft committed to a five-part framework focused on being a responsible long-term neighbor. This includes paying the full cost of electricity and associated grid upgrades so datacenter demand is not passed on to residents; minimizing water use and replenishing more water than is consumed within the same watershed; paying full local property taxes without seeking abatements; investing in local workforce development through partnerships and training programs; and expanding AI education and community investment in host regions.[96]

A key example of Microsoft's longstanding community-first approach to datacenter operations can be seen in Quincy, Washington, where we broke ground on our first company-owned datacenter 20 years ago. Since the facility was established, Quincy's poverty rate has declined from 29.4% to 13.1%, local property tax rates have decreased for residents, and a strengthened tax base has helped fund critical community infrastructure, including a new hospital, city hall, library, high school, and public safety facilities. Further, recognizing the relationship between datacenters and water usage, Microsoft partnered with the City of Quincy to build the first-of-its-kind cooling water reuse facility, helping achieve an average 97% reduction in potable water consumption. Today, the datacenter employs approximately 400 people, with employment expected to grow to 700 by the end of 2026.[97]

Broadening feedback and learning

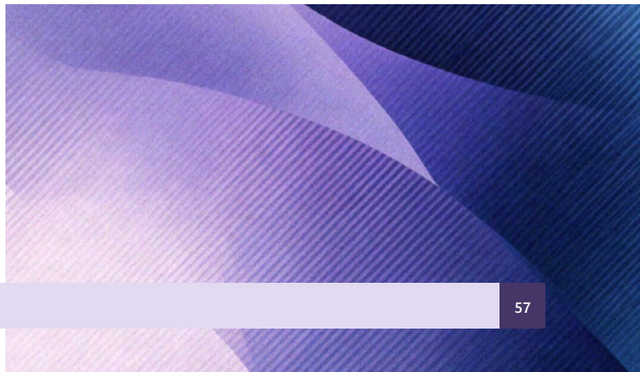
Through ongoing dialogue with everyday technology users, enterprise adopters, and subject matter experts, we continuously refine our AI governance practices based on real-world experiences, cross-disciplinary insights, historical lessons, and evolving societal expectations.

Engaging with external subject matter experts

As part of our AI governance program, Microsoft regularly monitors the AI landscape to identify emerging risks and develop early guidance for product teams. As we contemplate issues raised by rapidly evolving risks, we consult external experts to share and test our thinking and learn new approaches. As part of this approach, over the last year, Microsoft co-hosted workshops and roundtables with:

- The Center for Democracy and Technology (CDT) and other experts in academia, civil society, and industry to explore evolving capabilities of AI memory and personalization, including how memory is constructed, stored, and operationalized in AI products.
- Ethical Resolve to gather insights from diverse subject matter experts to help shape Microsoft's internal policies and practices and re-engineered Responsible AI Standard.

We also convened a workshop on AI companions and youth safety with external subject matter experts from industry, academia, and civil society, aiming to help strengthen collaboration across a community working to rapidly advance understanding of risks and effective mitigations.



Strategic foresight discussions with Global Perspectives Fellows

In partnership with the Stimson Center, Microsoft is sponsoring the third year of the Global Perspectives on Responsible AI Fellowship. As part of the program, this year's fellows participated in global convenings alongside policymakers and technology leaders, including a summit in Abu Dhabi, where fellows emphasized that AI systems depend on physical infrastructure unevenly available across regions; that sustainable AI adoption requires education and workforce pathways resilient to AI-driven economic change; and that public trust in AI is ultimately rooted in inclusion, access, and meaningful civic participation.[98] This year, the program also shifted toward a forward-looking approach, supporting fellows as they explore alternative futures for scaled AI adoption and design practical, locally grounded pathways maximizing social and economic benefit while minimizing risk.

Integrating lessons from other domains

To help advance more effective and interoperable AI governance, Microsoft has been working since 2023 with independent experts across a range of established domains.[99] In 2025, we examined the role of evaluation in governance, convening experts from numerous fields with decades of history managing cross-border technology risk: civil aviation, cybersecurity, financial services, genome editing, medical devices, nanotechnology, nuclear technology, and pharmaceuticals. We found that evaluation has worked best when driven by scientific rigor, standards development, and public-private coordination.[100] A key takeaway—that general-purpose technologies like genome editing, nanotechnology, and software also present similar challenges with evaluation and information-sharing across the value chain—has also informed our ongoing research about what more we can learn from other domains about effective transparency in practice.

Tuning into new voices

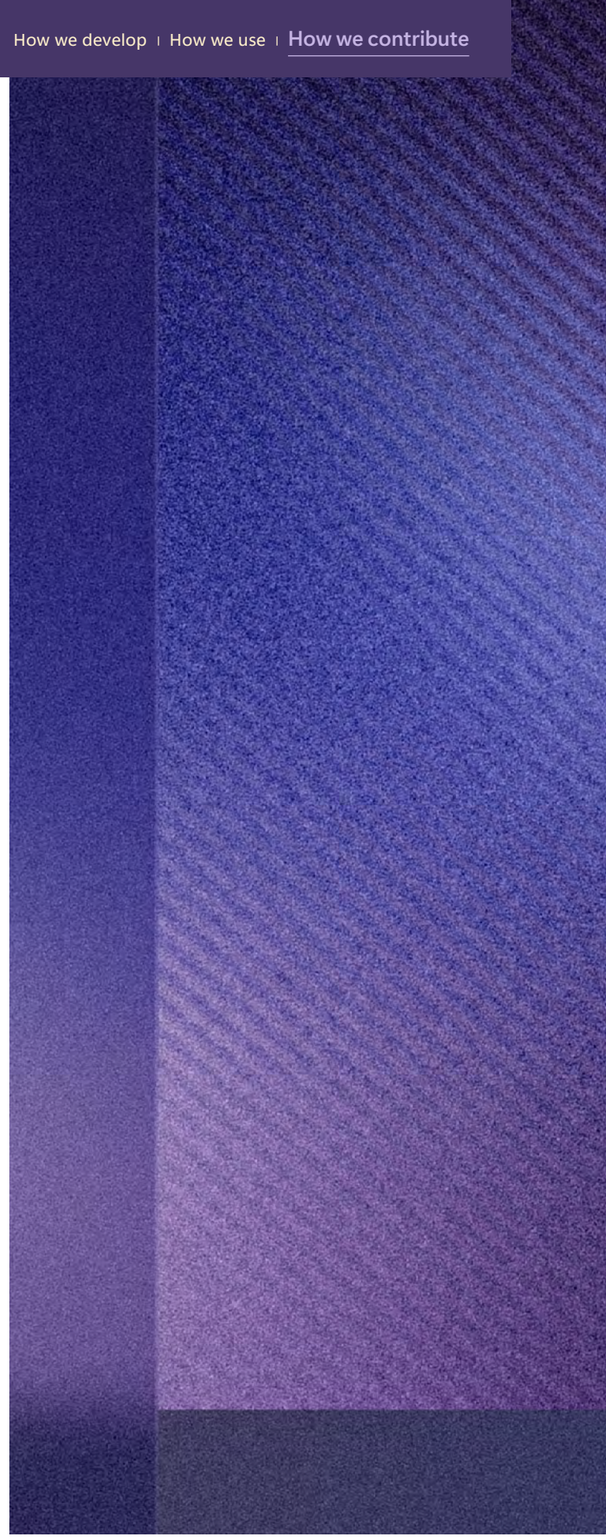
Beyond the signals we get through customers, partners, and experts, we recognize the value of a broader feedback loop on a range of issues. Over the last year, we've worked through new and evolving partnerships to consult with a range of AI technology users in both small-group settings and at-scale surveys, aiming to learn what people expect from AI and responsible practices and how today's approaches address their interests or leave gaps, including:

- **Youth:** We learned from students during the AI Companionship Youth Co-Design Sessions with Cyberlite that loss of critical thinking was a top concern for students, with 40% fearing AI dependency will make them "stop thinking for themselves." Another top concern involved losing control of their personal narrative. This year, we also launched our first **AI Futures Youth Council**, bringing teens aged 13–17 from the US and Europe together to share their ideas and experiences.[101]

> "AI should be incorporated [...] to help students like brainstorm, understand, and learn. But when [...] it's time to use like actual critical thinking, I think it's best for the students to actually do it."
>
> **Boy**
> 17, California
> Participant in our inaugural AI Futures Youth Council

- **Everyday users:** Microsoft participated in the "Industry-Wide Forum," convened by Stanford University's Deliberative Democracy Lab, to better understand expectations of users from the US and India regarding AI agent transparency and human oversight and where deployment of more autonomous capabilities is more widely supported.[102]
- **AI builders:** Microsoft supported research by UC Berkeley's AI Research (BAIR) Lab that informed a playbook on transparency in generative AI systems.[103] As part of this project, BAIR mapped and assessed existing transparency tools and resources and surveyed how different stakeholder groups—including AI builders, end users, and policymakers—experience transparency in practice, and sought feedback on playbook prototypes from study groups, yielding insights that Microsoft incorporated into our own transparency practices.

Microsoft partners across and beyond the responsible AI ecosystem, investing in a feedback loop that is core to Microsoft's AI governance program and our commitment to advancing practices and tools. We aim to empower and learn from a global community as we shape the future of AI together.



Section 5

Looking forward

As we look to the year ahead, we are optimistic about AI's opportunity and determined to meet the measure of this moment. It calls for clarity, rigor, and coherence across the AI ecosystem and within a company like Microsoft. Expectations for accountability understandably continue to rise as AI technologies become more capable, interconnected, and unevenly deployed around the world.

For our part, we are focused on building on our foundation and strengthening our capabilities across the following three areas.

1 | Continuously evolving our governance framework as AI advances

Governance must do more than set expectations; it must continuously incorporate new evidence about how systems behave in practice and shape how they are operated across the AI lifecycle. We will continue updating our Responsible AI Standard to reflect emerging risks, regulatory developments, and customer expectations while also embedding responsible AI requirements more deeply into the technology stack itself. This includes strengthening identity and access controls for agentic AI systems, advancing runtime safeguards, and expanding observability so governance can move closer to the point of execution. It also includes tightening the feedback loop between research, policy, and practice—identifying emerging capabilities and risks, understanding how they evolve through deployment and adaptation, and rapidly translating those insights into governance improvements.

2 | Strengthening measurement, evaluation, and accountability in practice

The next generation of responsible AI governance requires more than static assessments; it requires ongoing evidence about how systems perform pre- and post-deployment as capabilities evolve and real-world uses emerge. We will continue to invest in advancing AI evaluation science and strengthening tools to support dynamic evaluation internally, for customers, and across the ecosystem. This includes internal red teaming and measurement; tools embedded in Foundry; research partnerships with global AISIs and US CAISI; shared benchmarks for reliability developed via MLCommons; and assurance pathways developed by the Appia Foundation. Robust and dynamic evaluation and assurance are essential to translating policy into practice, operationalizing responsible AI across the lifecycle, and creating clearer expectations for accountability across the ecosystem.

3 | Building resilience and shared trust across the AI ecosystem

AI governance is a shared responsibility, and as emerging capabilities introduce novel risk surfaces, resilience will depend on both strong internal response mechanisms and effective external collaboration. We will continue advancing our ability to detect, understand, and respond to AI-related incidents; strengthening appropriate information sharing with defenders; and enhancing our transparency practices so customers and policymakers understand the capabilities, limitations, and evolving performance of our AI technologies. At the same time, we will deepen engagement with customers, policymakers, researchers, and civil society to help shape governance approaches that are practical, interoperable, and responsive to the realities of AI deployment at scale.

We also recognize that AI governance outcomes depend on collaboration. We will continue to expand our partnerships across industry, academia, civil society, and governments to advance best practices and shared learning. These collaborations are essential not only to aligning governance frameworks with the pace and scale of AI innovation but also to developing shared evidence, common evaluation approaches, and practical governance practices that no single organization can establish alone.

As our governance practices continue to evolve, we'll continue to share what is working, where challenges remain, and where we are focused on driving greater readiness. We invite you to engage with the insights we have shared this year, to challenge our assumptions, and to collaborate with us. Ultimately, building justified trust and confidence in AI through rigorous and effective governance is both a joint endeavor and an essential prerequisite to ensuring that AI's extraordinary promise becomes reality for everyone.

Endnotes

1 Global AI Diffusion in Q1 2026 - AI Economy Institute.
microsoft.com/en-us/corporate-responsibility/topics/ai-economy-institute/reports/global-ai-adoption-2026-q1/?=1

Global AI Adoption in 2025 - A Widening Digital Divide.
microsoft.com/en-us/research/wp-content/uploads/2026/01/Microsoft-AI-Diffusion-Report-2025-H2.pdf

2 AILuminate Multimodal – MLCommons.
mlcommons.org/ailuminate/ailuminate-multimodal/

We need to act with urgency to address the growing AI divide - Microsoft On the Issues.
blogs.microsoft.com/on-the-issues/2026/02/17/acting-with-urgency-to-address-the-growing-ai-divide

Pluralis v0.1: Towards a Multicultural, Multimodal, Multilingual Benchmark for AI Risk and Reliability.
arxiv.org/abs/2607.06196

3 LINGUA: Expanding Europe's Voices in AI - Microsoft Research.
microsoft.com/en-us/research/academic-program/lingua-expanding-europes-voices-in-ai

We need to act with urgency to address the growing AI divide - Microsoft On the Issues.
blogs.microsoft.com/on-the-issues/2026/02/17/acting-with-urgency-to-address-the-growing-ai-divide

4 Project Gecko - Microsoft Research.
microsoft.com/en-us/research/project/project-gecko

Paza - Build Robust Speech Models.
microsoft.github.io/Paza

Vibhasha - Build Multilingual & Multicultural AI Systems.
microsoft.github.io/Vibhasha

BYOL: Bring Your Own Language Into LLMs.
arxiv.org/abs/2601.10804

PazaBench - ASR Leaderboard for Low Resource Languages.
huggingface.co/spaces/microsoft/paza-bench

Behind the Chat: A Human Guide to Safe AI Conversations.
cyberlite.org/behind-the-chat

5 Agentic systems also increase the importance of efficient and reliable inference infrastructure since they operate through repeated cycles of planning and action rather than single prompts. This shift has heightened the need for efficient, reliable, and low-latency inference infrastructure, including hardware optimized for agentic workloads and integrated into control planes that deliver security and observability.

Maia 200: The AI accelerator built for inference - The Official Microsoft Blog.
blogs.microsoft.com/blog/2026/01/26/maia-200-the-ai-accelerator-built-for-inference/

6 A Roadmap For Responsible Approaches to AI Memory - Center for Democracy and Technology.
cdt.org/insights/a-roadmap-for-responsible-approaches-to-ai-memory/

Advancing AI evaluation with the Center for AI Standards (US) and Innovation and the AI Security Institute (UK).
blogs.microsoft.com/on-the-issues/2026/05/05/advancing-ai-evaluation-with-the-center-for-ai-standards-us-and-innovation-and-the-ai-security-institute-uk//

7 Agent Evaluators for Generative AI - Microsoft Foundry.
learn.microsoft.com/en-us/azure/foundry/concepts/evaluation-evaluators/agent-evaluators

How to configure guardrails and controls in Microsoft Foundry - Microsoft Foundry.
learn.microsoft.com/en-us/azure/foundry/guardrails/how-to-create-guardrails?tabs=python

Build agents you can trust across any framework with open evals and a control standard - Microsoft Foundry Blog.
devblogs.microsoft.com/foundry/build-2026-open-trust-stack-ai-agents

8 Prompt Shields in Microsoft Foundry.
learn.microsoft.com/en-us/azure/foundry/openai/concepts/content-filter-prompt-shields

9 Advancing digital content transparency and authenticity.
c2pa.org/

Content Credentials - C2PA Technical Specification.
spec.c2pa.org/specifications/specifications/2.3/specs/C2PA_Specification.html

10 Project Glasswing: Securing critical software for the AI era.
anthropic.com/glasswing

Get my drift? Catching LLM Task Drift with Activation Deltas.
arxiv.org/pdf/2406.00799

11 Cross-border collaboration: International law enforcement and Microsoft dismantle transnational scam network targeting older adults - Microsoft On the Issues.
blogs.microsoft.com/on-the-issues/2025/06/05/microsoft-dismantle-transnational-scam/

Disrupting a global cybercrime network abusing generative AI - Microsoft On the Issues.
blogs.microsoft.com/on-the-issues/2025/02/27/disrupting-cybercrime-abusing-gen-ai/

12 Industry Accord Against Online Scams & Fraud.
services.google.com/fh/files/newsletters/industryaccord.pdf

13 It's About Time: The Copilot Usage Report 2025 | Microsoft AI.
microsoft.ai/news/its-about-time-the-copilot-usage-report-2025/

Health Check: How People Use Copilot for Health. These insights are derived through privacy-preserving, automated research workflows designed to understand emerging patterns while protecting user data.
microsoft.ai/news/health-check-how-people-use-copilot-for-health/

14 Microsoft Family Safety.
account.microsoft.com/family/windows/coldstart

15 Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies - Microsoft Research.
microsoft.com/en-us/research/publication/fostering-appropriate-reliance-on-large-language-models-the-role-of-explanations-sources-and-inconsistencies/

"Helping Me Versus Doing It for Me": Designing for Agency in LLM-Infused Writing Tools for Science Journalism - Microsoft Research.
microsoft.com/en-us/research/publication/helping-me-versus-doing-it-for-me-designing-for-agency-in-llm-infused-writing-tools-for-science-journalism/

Tools for Thought - Microsoft Research.
microsoft.com/en-us/research/project/tools-for-thought/

What Teens Say About AI Companionship - Cyberlite.
cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/bade/documents/products-and-services/en-us/digitalsafety/What-Teens-Say-About-AI-Companionship-Microsoft-x-Cyberlite.pdf

16 Data Summary for MAI-Image-1.
aka.ms/MAI-Image-1-data-summary

MAI-Image-1 Model Card.
aka.ms/MAI-Image-1-modelcard

Bing Image Creator - Microsoft Bing.
microsoft.com/en-us/bing/features/bing-image-creator/

17 Publications - Frontier Model Forum.
frontiermodelforum.org/publications/

AILuminate – MLCommons.
mlcommons.org/ailuminate/

A New Standard for AI Risk: How the AILuminate Global Assurance Program Is Reshaping Reliability - MLCommons.
mlcommons.org/2026/02/ailuminate-global-assurance/

18 OECD launches Hiroshima AI Process Reporting Framework 2.0.
oecd.ai/en/haip-2-launch

19 MAI-Image-2e - Microsoft Foundry.
ai.azure.com/catalog/models/MAI-Image-2e

AI Model Catalog | Microsoft Foundry Models (MAI- Image-2).
ai.azure.com/catalog/models/MAI-Image-2.5?search=mai-image-2

20 Introducing the Agent Governance Toolkit: Open-source runtime security for AI agents - Microsoft Open Source Blog.
opensource.microsoft.com/blog/2026/04/02/introducing-the-agent-governance-toolkit-open-source-runtime-security-for-ai-agents

What is Microsoft Foundry Control Plane? - Microsoft Foundry.
learn.microsoft.com/en-us/azure/foundry/control-plane/overview

21 Microsoft Agent 365: The Control Plane for Agents.
microsoft.com/en-us/microsoft-agent-365/

22 Microsoft Global Human Rights Statement.
microsoft.com/en-us/corporate-responsibility/human-rights-statement

23 Fiscal Year 2025 Microsoft Generative AI HRIA Executive Summary.
cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/Generative-AI-HRIA-Executive-Summary-2025.pdf

24 Microsoft Vulnerability Severity and Content Classifications for AI Systems.
microsoft.com/en-us/msrc/aibugbar

25 Manipulating AI memory for profit: The rise of AI Recommendation Poisoning - Microsoft Security Blog.
microsoft.com/en-us/security/blog/2026/02/10/ai-recommendation-poisoning/

Detecting backdoored language models at scale | Microsoft Security Blog.
microsoft.com/en-us/security/blog/2026/02/04/detecting-backdoored-language-models-at-scale/

26 Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge.
arxiv.org/abs/2502.00561

Taxonomizing Representational Harms using Speech Act Theory.
aclanthology.org/2025.findings-acl.202.pdf

Comparison requires valid measurement: Rethinking attack success rate comparisons in AI red teaming.
arxiv.org/abs/2601.18076

Validating LLM-as-a-Judge Systems under Rating Indeterminacy.
arxiv.org/abs/2503.05965

Understanding and Meeting Practitioner Needs When Measuring Representational Harms Caused by LLM-Based Systems.
aclanthology.org/2025.findings-acl.947.pdf

GitHub - Responsible AI - ASSERT.
github.com/responsibleai/ASSERT

Turn specs into evals for any agent with ASSERT - Command Line.
commandline.microsoft.com/assert-written-intent-executable-evals/

27 Microsoft Frontier Governance Framework.
cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft-Frontier-Governance-Framework.pdf

28 Advancing AI evaluation with the Center for AI Standards (US) and Innovation and the AI Security Institute (UK) - Microsoft On the Issues.
blogs.microsoft.com/on-the-issues/2026/05/05/advancing-ai-evaluation-with-the-center-for-ai-standards-us-and-innovation-and-the-ai-security-institute-uk/

29 Defending Against Indirect Prompt Injection Attacks With Spotlighting - Microsoft Research.
microsoft.com/en-us/research/publication/defending-against-indirect-prompt-injection-attacks-with-spotlighting

30 To address circumstances in which indirect prompt injection attacks bypass prevention and detection safeguards, Microsoft takes a defense-in-depth approach, designing systems with the assumption that some attacks may succeed and focusing on minimizing their potential security impact. A key pillar of this approach is strong data governance. Since indirect prompt injection relies on the LLM-based application having the same access permissions as the user, this risk can be deterministically mitigated through fine-grained permissions and access controls. For example, Microsoft 365 Copilot respects sensitivity labels and Microsoft Purview policies, which can restrict the AI's access to specific data or prevent it from summarizing sensitive files. Additionally, once a specific attack vector—such as a data exfiltration technique—is identified, Microsoft can implement deterministic blocks to neutralize its impact. Finally, as a last line of defense, human-in-the-loop mechanisms can be used to require user confirmation before the AI proceeds with sensitive actions or outputs, further reducing the risk of unintended consequences.

31 Verifying Media Content Sources missing to include this link as well - C2PA.
c2pa.org/

32 Data summary for MAI- Image-1.
huggingface.co/microsoft/phi-4/blob/main/data_summary_card.md

MAI-Image-1 Model Card.
publicdoc.blob.core.windows.net/public/MAI-Image-1-Model-Card-13-Nov-25.pdf

MAI-Image-2e (Efficient) Model Card.
microsoft.ai/pdf/MAI-Image-2e-Model-Card.pdf

33 MAI-Image-2e - Microsoft Foundry Models.
ai.azure.com/catalog/models/MAI-Image-2e

MAI-Image-2 - Microsoft Foundry Models.
ai.azure.com/catalog/models/MAI-Image-2

34 Overview of Microsoft Foundry Models - Microsoft Learn.
learn.microsoft.com/en-us/azure/foundry/concepts/foundry-models-overview

35 Bing Image Creator.
microsoft.com/en-us/bing/features/bing-image-creator

36 Copilot Cowork: Now available in Frontier - Microsoft 365 Blog.
microsoft.com/en-us/microsoft-365/blog/2026/03/30/copilot-cowork-now-available-in-frontier

37 Application card: Microsoft 365 Copilot Cowork.
learn.microsoft.com/en-us/microsoft-365/copilot/responsible-ai/copilot-cowork-application-card

38 Public use of a generalist LLM chatbot for health queries.
nature.com/articles/s44360-026-00117-x

It's About Time: The Copilot Usage Report 2025.
microsoft.ai/news/its-about-time-the-copilot-usage-report-2025/

39 About Copilot Health - Microsoft Support.
support.microsoft.com/en-us/microsoft-copilot/copilot-health

40 It's About Time: The Copilot Usage Report 2025.
microsoft.ai/news/its-about-time-the-copilot-usage-report-2025/

Health Check: How People Use Copilot for Health | Microsoft AI.
microsoft.ai/news/health-check-how-people-use-copilot-for-health/

To ensure user privacy, the usage studies employed rigorous deidentification protocols, including automated scrubbing of personally identifiable information (PII) and the use of machine-based "eyes-off" classifiers, meaning no human researchers directly viewed the content of conversations.

41 Actionable issues remain concentrated in Information Disclosure (21.3%) and AI-Derived Harm (13%), highlighting priorities around privacy safeguards and robust AI governance. Additional risks include Spoofing (7.8%), Elevation of Privilege (5.6%), and Remote Code Execution (5%), reinforcing the need for layered defenses and proactive alignment with bug bars to strengthen resilience and responsible AI development. Cases classified as "Not a Vulnerability" reflect ongoing efforts to clarify boundaries of what falls within bug bar scope and our required criteria.

42 This chart is based on analysis of case data that is inherently noisy and subject to limitations in classification, reporting, and completeness. While individual records may contain inaccuracies or omissions, the dataset is believed to be sufficiently representative to illustrate overall trends and relative distributions of cases by impact. Results should be interpreted as directional rather than precise measurements.

43 GitHub Copilot.
learn.github.com/learning-pathways/github-copilot

44 The Impact of AI on Developer Productivity: Evidence from GitHub Copilot.
microsoft.com/en-us/research/publication/the-impact-of-ai-on-developer-productivity-evidence-from-github-copilot

Agentic DevOps: Evolving software development with GitHub Copilot and Microsoft Azure.
azure.microsoft.com/en-us/blog/agentic-devops-evolving-software-development-with-github-copilot-and-microsoft-azure/

45 Microsoft uses Copilot Studio to reshape customer experience and drive higher engagement.
microsoft.com/en/customers/story/26166-microsoft-microsoft-copilot-studio

46 Run evaluations from the Microsoft Foundry portal - Microsoft Foundry | Microsoft Learn.
learn.microsoft.com/en-us/azure/foundry/how-to/evaluate-generative-ai-app

47 Microsoft provides faster answers to complex licensing questions with an agent built on Azure AI Services and Microsoft Copilot Studio | Microsoft Customer Stories.
.microsoft.com/en/customers/story/25413-microsoft-azure-ai-search

48 The Hawthorne Effect in Reasoning Models: Evaluating and Steering Test Awareness.
arxiv.org/html/2505.14617v3

49 Detecting backdoored language models at scale | Microsoft Security Blog.
microsoft.com/en-us/security/blog/2026/02/04/detecting-backdoored-language-models-at-scale

The Trigger in the Haystack: Extracting and Reconstructing LLM Backdoor Triggers.
arxiv.org/pdf/2602.03085

50 Manipulating AI memory for profit: The rise of AI Recommendation Poisoning | Microsoft Security Blog.
microsoft.com/en-us/security/blog/2026/02/10/ai-recommendation-poisoning/

51 The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers.
microsoft.com/en-us/research/wp-content/uploads/2025/01/lee_2025_ai_critical_thinking_survey.pdf

52 LLMail-Inject: A Dataset from a Realistic Adaptive Prompt Injection Challenge.
arxiv.org/pdf/2506.09956

53 IMDA and Microsoft commit to advancing AI safety and security | IMDA.
imda.gov.sg/resources/press-releases-factsheets-and-speeches/imda-and-microsoft-commit-to-advancing-ai-safety-and-security

54 Agentic AI Research and Innovation (AARI).
microsoft.com/en-us/research/academic-program/agentic-ai-research-and-innovation/

Accelerating Foundation Models Research.
microsoft.com/en-us/research/collaboration/accelerating-foundation-models-research/projects/

55 CRA Trustworthy AI Research Fellow Spotlight: Diana Freed - CRN.
cra.org/crn/2025/11/cra-trustworthy-ai-research-fellow-spotlight-diana-freed/

CRA Trustworthy AI Research Fellow Spotlight: Calvin Liang - CRN.
cra.org/crn/2026/01/cra-trustworthy-ai-research-fellow-spotlight-calvin-liang/

CRA Trustworthy AI Research Fellow Spotlight: Jayshree Sarathy - CRN.
cra.org/crn/2026/02/cra-trustworthy-ai-research-fellow-spotlight-jayshree-sarathy/

56 Education in the AI Economy: the AI Economy Institute's 2025 Fall Cohort - Microsoft Research.
microsoft.com/en-us/research/articles/education-in-the-ai-economy-the-ai-economy-institutes-2025-fall-cohort/

57 Transforming R&D with agentic AI: Introducing Microsoft Discovery - Microsoft Azure Blog.
azure.microsoft.com/en-us/blog/transforming-rd-with-agentic-ai-introducing-microsoft-discovery

National AI Research Resource (NAIRR) Pilot - Microsoft Research: Research Inspiration.
microsoft.com/en-us/research/project/national-ai-research-resource-nairr-pilot/research-inspiration/

58 Preliminary Taxonomy of AI-Bio Misuse Mitigations.
frontiermodelforum.org/issue-briefs/preliminary-taxonomy-of-ai-bio-misuse-mitigations/

Managing Advanced Cyber Risks in Frontier AI Frameworks.
frontiermodelforum.org/technical-reports/managing-advanced-cyber-risks-in-frontier-ai-frameworks/

ISSUE BRIEF: Adversarial Distillation.
frontiermodelforum.org/issue-briefs/issue-brief-adversarial-distillation/

59 Prioritizing Real-Time Failure Detection in AI Agents - Partnership on AI.
partnershiponai.org/resource/prioritizing-real-time-failure-detection-in-ai-agents/

60 Browse Reports - OECD.AI.
oecd.ai/en/transparency/reports

61 OECD launches Hiroshima AI Process Reporting Framework 2.0.
oecd.ai/en/haip-2-launch

62 AI Risk & Reliability - MLCommons.
mlcommons.org/working-groups/ai-risk-reliability/ai-risk-reliability/

63 AILuminate Multimodal.
mlcommons.org/ailuminate/ailuminate-multimodal/

64 MLCommons Unveils New Jailbreak Benchmark, Quantifying AI's "Resilience Gap" to Adversarial Attacks - MLCommons.
mlcommons.org/2025/10/ailuminate-jailbreak-v05/

65 From open source to agentic systems: Microsoft at Open Source Summit North America 2026 | Microsoft Open Source Blog.
opensource.microsoft.com/blog/2026/05/18/from-open-source-to-agentic-systems-microsoft-at-open-source-summit-north-america-2026/

66 Members of the OpenTelemetry Project | OpenTelemetry.
opentelemetry.io/community/members/

Companies Table - Grafana.
opentelemetry.devstats.cncf.io/d/5/companies-table?orgId=1&var-period_name=Last month&var-metric=contributions

Azure AI Foundry: Advancing OpenTelemetry and delivering unified multi-agent observability.
techcommunity.microsoft.com/blog/azure-ai-foundry-blog/azure-ai-foundry-advancing-opentelemetry-and-delivering-unified-multi-agent-obse/4456039

Observability for Multi-Agent Systems with Microsoft Agent Framework and Azure AI Foundry | Microsoft Community Hub.
techcommunity.microsoft.com/blog/azure-ai-foundry-blog/observability-for-multi-agent-systems-with-microsoft-agent-framework-and-azure-a/4469090

Observability | Microsoft Learn.
learn.microsoft.com/en-us/agent-framework/agents/observability?pivots=programming-language-csharp

67 Advancing digital content transparency and authenticity.
https://c2pa.org/

The C2PA Launches Content Credentials 2.3 and Celebrates 5 Years of Impact Across the Digital Ecosystem – Coalition for Content Provenance and Authenticity (C2PA).
c2pa.org/the-c2pa-launches-content-credentials-2-3-and-celebrates-5-years-of-impact-across-the-digital-ecosystem/

68 Linux Foundation Launches Appia Foundation to Establish Standardized Conformity Specifications Across the AI Value Chain – Appia Foundation.
appiafoundation.org/linux-foundation-launches-appia-foundation-to-establish-standardized-conformity-specifications-across-the-ai-value-chain/

69 A New Standard for AI Risk: How the AILuminate Global Assurance Program Is Reshaping Reliability - MLCommons.
mlcommons.org/2026/02/ailuminate-global-assurance/

70 ISO/IEC 42001:2023 Artificial Intelligence Management System Standards - Microsoft Compliance | Microsoft Learn.
learn.microsoft.com/en-us/compliance/regulatory/offering-iso-42001

71 Microsoft Entra Agent ID documentation | Microsoft Learn.
learn.microsoft.com/en-us/entra/agent-id/

Best practices for Microsoft Entra Agent ID - Microsoft Entra Agent ID | Microsoft Learn.
learn.microsoft.com/en-us/entra/agent-id/best-practices-agent-id

72 Build agents you can trust across any framework with open evals and a control standard | Microsoft Foundry Blog.
devblogs.microsoft.com/foundry/build-2026-open-trust-stack-ai-agents

Introducing Agent Control Specification: Portable runtime governance for AI Agents.
commandline.microsoft.com/agent-control-specification-runtime-governance/

73 What's new in Microsoft Foundry | Build Edition | Microsoft Foundry Blog.
devblogs.microsoft.com/foundry/whats-new-in-microsoft-foundry-build-2026/

74 Build agents you can trust across any framework with open evals and a control standard - Microsoft Foundry Blog.
devblogs.microsoft.com/foundry/build-2026-open-trust-stack-ai-agents/

Turn specs into evals for any agent with ASSERT - Command Line.
commandline.microsoft.com/assert-written-intent-executable-evals/

75 Agent evaluators.
learn.microsoft.com/en-us/azure/foundry/concepts/evaluation-evaluators/agent-evaluators

76 What's new in Microsoft Foundry | Build Edition | Microsoft Foundry Blog.
devblogs.microsoft.com/foundry/whats-new-in-microsoft-foundry-build-2026/

77 Defending Against Indirect Prompt Injection Attacks with Spotlighting.
arxiv.org/pdf/2403.14720

microsoft.com/en-us/msrc/blog/2025/07/how-microsoft-defends-against-indirect-prompt-injection-attacks?msockid=3da790040c776d6f2b5485e40de56c06

78 Task Adherence in Azure AI Content Safety - Foundry Tools | Microsoft Learn.
learn.microsoft.com/en-us/azure/ai-services/content-safety/concepts/task-adherence

79 Microsoft Elevate: Putting people first - Microsoft On the Issues.
blogs.microsoft.com/on-the-issues/2025/07/09/elevate/

80 Unlock generative AI safely and responsibly—toolkit | Microsoft Learn.
learn.microsoft.com/en-us/training/educator-center/instructor-materials/classroom-toolkit-unlock-generative-ai-safely-responsibly

New resources looking at AI misuse and nudification launched today.
childnet.com/blog/new-resources-looking-at-ai-misuse-and-nudification-launched-today/

81 What-Teens-Say-About-AI-Companionship-Microsoft-x-Cyberlite.pdf.
cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/bade/documents/products-and-services/en-us/digitalsafety/What-Teens-Say-About-AI-Companionship-Microsoft-x-Cyberlite.pdf

Behind the Chat: AI Companionship Resources for Teens | Cyberlite.
cyberlite.org/behind-the-chat

82 Adolescents & Anthropomorphic AI: Rethinking Design for Wellbeing.
everyone.ai/wp-content/uploads/2026/02/Adolescents-Anthropomorphic-AI-Rethinking-Design-for-Wellbeing-.pdf

83 Deadly Coders | Minecraft Education.
education.minecraft.net/en-us/blog/deadly-coders

84 Virtual AI Resources.
seniorplanet.org/ai

85 Understanding Scams and Fraud.
seniorplanet.org/scam-prevention

86 Talking about AI: Newsroom Toolkit.
poynter.org/mediawise/programs/talking-about-ai-newsroom-toolkit/

87 We need to act with urgency to address the growing AI divide - Microsoft On the Issues.
blogs.microsoft.com/on-the-issues/2026/02/17/acting-with-urgency-to-address-the-growing-ai-divide/

88 LINGUA: Expanding Europe's Voices in AI.
microsoft.com/en-us/research/academic-program/lingua-expanding-europes-voices-in-ai

LINGUA Africa Open Call for Inclusive AI Language Projects - Microsoft Research.
microsoft.com/en-us/research/academic-program/lingua-africa-open-call-for-inclusive-ai-language-projects/

89 BYOL: Bring Your Own Language Into LLMs.
arxiv.org/abs/2601.10804

90 Giving Every Language a Voice.
unlocked.microsoft.com/llms-nunavut/

91 AILuminate Multimodal.
mlcommons.org/ailuminate/ailuminate-multimodal/

92 Building Benchmarks from the Ground Up: Community-Centered Evaluation of LLMs in Healthcare Chatbot Settings.
arxiv.org/pdf/2509.24506

93 Paza: Introducing automatic speech recognition benchmarks and models for low resource languages.
microsoft.com/en-us/research/blog/paza-introducing-automatic-speech-recognition-benchmarks-and-models-for-low-resource-languages/

Advancing AI to meet needs of the global majority.
microsoft.com/en-us/research/story/advancing-ai-to-meet-needs-of-the-global-majority/

94 PazaBench.
huggingface.co/spaces/microsoft/paza-bench

Paza Playbook.
microsoft.github.io/Paza/

Vibhasha Playbook.
microsoft.github.io/Vibhasha/

95 Advancing digital content transparency and authenticity.
c2pa.org/

96 Building Community-First AI Infrastructure - Microsoft On the Issues.
blogs.microsoft.com/on-the-issues/2026/01/13/community-first-ai-infrastructure/

97 Partnering with the City of Quincy to open Washington's first industrial water reuse center - Microsoft Local.
local.microsoft.com/blog/partnering-with-the-city-of-quincy-to-open-washingtons-first-industrial-water-reuse-center/

Growing with community – Microsoft Unlocked.
unlocked.microsoft.com/quincy-growth/

98 The power of context: Embracing global perspectives to foster meaningful AI adoption.
microsoft.com/en-us/corporate-responsibility/topics/responsible-ai/stories/global-ai-perspectives

99 Global governance: goals and lessons for AI - Microsoft On the Issues.
blogs.microsoft.com/on-the-issues/2024/09/23/global-governance-goals-and-lessons-for-ai/

100 Learning from other domains to advance AI evaluation and testing - Microsoft Research.
microsoft.com/en-us/research/blog/learning-from-other-domains-to-advance-ai-evaluation-and-testing

AI Testing and Evaluation: Learnings from Science and Industry - Microsoft Research.
microsoft.com/en-us/research/story/ai-testing-and-evaluation-learnings-from-science-and-industry/

Learning from other domains to advance AI evaluation and testing.
microsoft.com/en-us/research/wp-content/uploads/2025/08/Learning-from-other-Domains-to-Advance-AI-Evaluation-and-Testing_-v3-1.pdf

101 Uplifting and empowering young people for an AI future - Microsoft On the Issues.
blogs.microsoft.com/on-the-issues/2025/11/12/uplifting-and-empowering-young-people-for-an-ai-future/

102 People Shaping AI: Groundbreaking Industry-Wide Forum Invites Public Input for the Future of AI Agents Through Public Deliberation | FSI.
cddrl.fsi.stanford.edu/news/people-shaping-ai-groundbreaking-industry-wide-forum-invites-public-input-future-ai-agents

103 Transparency in Generative AI | Responsible AI Initiative.
re-ai.berkeley.edu/projects/transparency-generative-ai