

OECD AI Transparency Report

Organization: Microsoft (USA)

Reporting Period: Q3 2026

Published: October 2, 2026

Section 0 – Your role in the AI system lifecycle

Responding as (please select all that apply):

- Model developer (/provider)
- Application developer (/provider)
- Deployer

If any of the fields above are selected, please indicate your organization size:

Large enterprise: 250 or more employees

Section 1 – Risk Identification and Evaluation

1. How does your organization define and/or classify different types of risks related to AI?¹

- Maintain an internal taxonomy, classification system, and/or set of risk area definitions
- Align with specific external frameworks, standards, reports, and/or guidance documents such as taxonomies, classification systems, and/or sets of definitions for risk areas (e.g. the NIST AI Risk Management Framework, the ISO/IEC 42001, the OECD AI Principles, the EU General-Purpose AI Code of Practice, the International AI Safety Report). (please specify): Microsoft supports internationally aligned approaches to AI governance, including those that provide taxonomies, classification systems, and/or sets of definitions for risk areas. Microsoft's Responsible AI Principles are aligned with the OECD AI Principles and are operationalized through the Microsoft Responsible AI Standard. We align with recognized standards and best practices, including the NIST AI Risk Management Framework, and have achieved ISO/IEC 42001 certification for Microsoft 365 Copilot and other products, providing independent validation of our AI management system. Microsoft signed the EU General-Purpose AI Code of Practice, a framework to support compliance with the EU AI Act's requirements for general-purpose AI (GPAI) models, and published Microsoft's Frontier Governance Framework which demonstrates alignment with the Code's scope of risks to be assessed for GPAI models with systemic risk. Microsoft also contributes to efforts to strengthen the shared evidence base for defining advanced AI capabilities and risks, including the International AI Safety Report.
- Consult with external stakeholders (e.g., from academia, industry groups, civil society, or standard development organisations) when developing or updating risk taxonomies
- Define enhanced safety measures for specific dangerous capability thresholds, that, once reached, require safeguards (e.g. capability threshold for dangerous Chemical, Biological, Radiological and Nuclear

(CBRN), cybersecurity attacks, dangerous advanced autonomy capabilities, or other capabilities).

How does your organization define thresholds at which severe risks, unless adequately mitigated, may be considered "unreasonable"? Please describe the framework(s) most relevant to your organization's role and context. If your organisation does not use formal thresholds, you may describe instead the policies, decision criteria, or governance processes you apply when assessing severe AI risks. You may also describe:

- Specific examples of contexts in which your models or systems could, in your assessment, pose unreasonable or intolerable risk;
- How mitigations are graduated by capability or propensity level;
- How your organisation considers the specificities of different AI technologies, such as generative AI and agentic AI, including any specific risk assessment or governance mechanisms applied to agentic AI capabilities.

Microsoft's [Frontier Governance Framework](#) (FGF) defines specific tracked high-risk capabilities at which severe risks — unless adequately mitigated — may warrant pausing development and deployment. These are: (i) chemical, biological, radiological, and nuclear (CBRN) weapons capabilities providing significant uplift; (ii) offensive cyberoperations capabilities providing uplift for highly disruptive or destructive cyberattacks, including on critical infrastructure; (iii) advanced autonomy, including expert-level autonomous completion of AI research and development tasks; (iv) loss of control, a model's ability to undermine effective human control through adaptive, deceptive, or self-reinforcing mechanisms such that, when deployed, it can no longer be reliably directed, modified, or shut down; and (v) harmful manipulation, or a model's ability to strategically distort human behavior or beliefs at scale.

Loss of control and harmful manipulation were added as tracked capabilities in the February 2026 FGF update. Microsoft continues to advance research on evaluation methods and risk acceptance criteria for these two newest tracked capabilities. The FGF applies to AI models developed by Microsoft where the training compute exceeds 10^{25} FLOPs. It also applies when Microsoft substantially fine-tunes a first- or third-party model and the fine-tuning compute exceeds one-third of the compute used to train the base model. Where the training compute of a third-party base model is unknown, the FGF applies a default threshold of one-third of 10^{25} FLOPs.

For models within scope, Microsoft first monitors for leading indicators of high-risk capabilities and conducts deeper capability assessments where warranted. Each tracked capability is assessed against qualitative thresholds corresponding to low, medium, high, or critical risk levels, detailed in Appendix I of the FGF. Models assessed as low or medium risk may be deployed with appropriate safeguards.

Models assessed as high or critical risk are subject to further review and additional safety and security mitigations before deployment, followed by post-mitigation assessment to ensure that residual risk has been reduced to a low or medium level. Examples of circumstances that could indicate elevated risk include:

- *CBRN weapons - high risk*: the model provides meaningful uplift to an expert's (PhD-level) ability to develop and deliver a novel CBRN threat, or to a medium-skilled actor's ability to develop and deliver a known CBRN threat. *Critical*: meaningful uplift to an expert's ability to develop a highly dangerous novel threat, or significantly lowering barriers for a low-skilled actor to develop and deliver a known CBRN threat.
- *Offensive cyberoperations - high risk*: the model meaningfully uplifts a low-to-medium-skilled actor's ability to conduct highly disruptive or destructive attacks on critical infrastructure (e.g., novel zero-day exploit chains, complex malware). *Critical*: it lets a low-skilled actor identify and exploit major

vulnerabilities, or enables an expert, well-resourced actor to mount novel, effective operations against hardened targets.

Mitigations are proportionate to the assessed risk. If during FGF implementation Microsoft identifies a risk that cannot be sufficiently mitigated, development and deployment are paused until mitigation practices evolve to meet the risk [FGF p.8]. “High” and “critical” risk levels on any tracked capability trigger escalating security measures - from restricted access, defense-in-depth, and advanced security red teaming for “high-risk” levels, up to high-trust developer environments, hardened tamper-resistant workstations with enhanced logging, and physical bandwidth limitations between weight-containing devices and the outside world for “critical” risk levels [FGF p.7].

Risk determinations do not rely exclusively on model capability. Microsoft also considers system-level mitigations, societal and institutional factors that affect whether a hazard could materialize, and the marginal capability uplift the model provides relative to currently available tools and information, including open-weights models [FGF p.6].

The FGF sits within Microsoft’s broader AI governance program, which addresses risks across the AI models, platforms, and applications Microsoft develops and deploys. This program defines three policy categories reflecting graduated concern: Restricted uses (subject to specific AI development or deployment restrictions, defined by the Office of Responsible AI); Sensitive uses (scenarios with significant impact on life opportunities, physical safety, or human rights, requiring notification to the Office of Responsible AI and additional governance steps); and Unsupported uses (uses not designed or evaluated for, for which customer use is discouraged).

Microsoft also draws on external expertise from independent evaluators, researchers, industry groups and other stakeholders to inform and help improve its risk identification and governance approaches, including as our approach to frontier risk evolves. Following the February 2026 update to the FGF, SaferAI increased Microsoft’s safety framework score from 18% to 33%, ranking Microsoft third among frontier AI developers in its [assessment](#) of risk management industry practices. SaferAI also assessed Microsoft as best in class for overall risk identification, as well as for risk governance decision-making and transparency.

Microsoft considers the specificities of different AI technologies in designing and implementing governance. Previously, generative AI and autonomous agentic AI requirements were formalized as standalone policies, and both have been integrated into Microsoft’s re-engineered Responsible AI Standard as “scenario-specific” (versus “core”) requirements. Scenario-specific requirements are tailored to specific technologies and risks, and specific practices are defined in response to those risks (e.g., specific threat modeling practices for agentic AI). From July 2025 to June 2026, Microsoft assessed nearly 2,500 generative AI releases, 30% of which were related to agentic AI, and reviewed over 450 Sensitive Use cases, of which over 30% were agentic related.

Frontier models are addressed as another “scenario” in our Responsible AI Standard, with scenario-specific requirements mapping to our FGF. The FGF’s “Advanced autonomy” capability thresholds directly address agentic risks by assessing autonomous task completion, tool use, and the ability to correct for complex error conditions [FGF p.13]. FGF capability elicitation explicitly tests models with multi-agent setups, prompt optimization, and tool/plugin connection to maximize performance on tracked high-risk capabilities [FGF p.6].

2. Both before and after deployment, what practices does your organization use to identify and evaluate vulnerabilities, incidents, emerging risks and misuse throughout the AI system lifecycle?

- Internal red teaming³
- Independent third-party red teaming
- Adversarial testing to detect technical vulnerabilities
- Publicly available benchmarks and measurement tools
- Internal benchmarks, standard metrics (e.g., frequency, severity scores)
- Structured techniques to explore the behavior or performance of a model and/or application, such as model elicitation or bottleneck assessments
- Qualitative assessments (e.g., expert judgment, narrative scoring)
- Specific risk identification practices for vulnerable groups
- Impact assessments such as personal data protection assessment and fundamental rights impact assessments
- Internal model, system, or application-independent research initiatives
- External model, system, or application-independent research initiatives
- Vulnerability reports
- Incident reports, including reports shared by other organizations
- Automated monitoring / remote status reporting (e.g., telemetry pipelines, automated alerts when risk thresholds are breached)
- Specific risk identification practices for agentic AI capabilities, e.g., autonomous action-taking, tool use, multi-step planning (please specify): Microsoft has developed new threat modelling techniques for agentic AI and integrated them into its Security Development Lifecycle and Responsible AI Standard. These techniques reflect the distinctive risks created when AI systems combine models with tools, data sources, memory, and external environments to execute multi-step actions. They help teams map system architectures and trust boundaries, assess risks associated with tool permissions and agent actions, model potential misuse and failure scenarios, and design mitigations such as least-privilege access, human approval points, logging, and ongoing monitoring. To support implementation at scale, Microsoft has conducted internal deep-dive workshops involving more than 3,000 engineers, helping product teams apply these new threat-modeling techniques to agentic AI systems and integrate appropriate safeguards throughout the development lifecycle. The FGF's "advanced autonomy" capability thresholds directly address agentic risks by assessing autonomous task completion, tool use, and the ability to correct for complex error conditions [FGF p.13]. FGF capability elicitation explicitly tests models with multi-agent setups, prompt optimization, and tool/plugin connection to maximize performance on tracked high-risk capabilities [FGF p.6].
- Assessment of risks arising from interactions among multiple AI agents
- Multilingual and cross-cultural evaluation, including collaboration with local ecosystems (e.g., national AI institutes, language-specific evaluation consortia) for under-represented languages and cultural contexts (please specify): Microsoft is committed to advancing multilingual and context-relevant AI because language and cultural context are fundamental both to expanding access to AI's benefits and to ensuring that AI systems are safe, secure, and reliable wherever they are deployed. As Microsoft highlighted at the India AI Impact Summit, language remains a significant barrier to AI diffusion, with AI systems often

performing less consistently in digitally underrepresented languages than in English. Addressing this gap requires investment across the AI lifecycle - from locally relevant data and model capabilities to evaluation methods grounded in the languages, cultural norms, and real-world contexts in which AI is used. Expanding the geographic and linguistic coverage of AI research and evaluation can therefore help broaden the diffusion of beneficial AI while surfacing safety, security, and performance risks that English-first or translation-based testing may fail to capture. Microsoft is advancing these objectives through work on multilingual and multicultural evaluation, frontier safety and security, and community-centered AI research:

- Microsoft is co-leading the MLCommons AILuminate benchmark's multilingual, multicultural, and multimodal expansion in partnership with Google DeepMind, enabling more reliable measurement of AI safety and security beyond English and Western-dominant cultural contexts. The effort is currently focused on building the capability to assess performance of benchmarks in multilingual and multicultural contexts, working with partners to implement that capability in major Indic and Asian languages, with plans to develop an open source playbook that would accelerate expansion to other regions and additional APAC languages.
- Microsoft contributed to the Frontier Model Forum's issue brief *Emerging Practices for Multilingual Evaluations for CBRN and Advanced Cyber Risks* (July 2026), which addresses the gap in safety and security evaluations for the highest-risk capabilities beyond English, and sets out best practices to build native evaluations across languages and cultures.
- Microsoft is also advancing Samiksha, a community-centered method for evaluating AI behavior in real-world contexts, in collaboration with Karya and The Collective Intelligence Project in India. This work complements benchmark-based evaluation by helping incorporate the perspectives and lived experiences of the communities in which AI systems are used.
- Microsoft signed a partnership with Singapore's Infocomm Media Development Authority (IMDA) in June 2026 to advance joint research and evaluation methodologies, including in multilingual AI safety.

Through Project Gecko, Microsoft is co-designing 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. 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. This work also advances the New Delhi Frontier AI Impact Commitments, which Microsoft signed alongside other leading AI developers and Indian innovators at the India AI Impact Summit in February 2026. The commitments reinforce the importance of strengthening multilingual and context-relevant AI evaluations and collaborating with local ecosystems to develop and apply evaluations for under-represented languages and cultural contexts.

Are any national AI safety / security institutes of the AISI network involved?

Yes

How does your organization monitor the interactions of AI agents under your control with other agents / assess risk from multiple agents interacting?

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.

Microsoft has evolved its risk management practices to account for risks increasingly shaped by access, permissions, memory, use over time, and interactions. As part of the new threat modelling techniques that Microsoft has developed for agentic AI and integrated into our [Security Development Lifecycle](#) and [Responsible AI Standard](#) [see response to above question], teams identify risks in their scenario, including risks from multiple agents interacting, such as the compromise or failure of one agent propagating through shared context, delegated tasks or interconnected tool chains.

Addressing these risks requires runtime controls as well as monitoring whether those controls are being enforced and how agents are behaving. More specifically, Microsoft has been developing and implementing measures that include assigning agents distinct identities and role-based access controls; applying runtime policies; evaluating agent performance on tasks and across end-to-end workflows; monitoring agent actions, communications and tool use, including through OpenTelemetry(OTel)-based observability traces and other metrics, event logs, and security logs; detecting anomalous behavior; and maintaining mechanisms to interrupt, restrict or disable agents when necessary.

Microsoft has also implemented capabilities across applications with agentic features, developer platforms (e.g., Foundry), and administrator tools (e.g., Microsoft Agent 365) and provided open-source tools, empowering our customers and the ecosystem to assess risks and monitor the agents and multi-agent systems they develop and deploy. For example, integrating with other Microsoft services, such as Entra and Defender, these capabilities can help developers and organizations:

- Manage agent identities and access: Establish distinct identities for agents and apply risk-based and role-based access controls (Microsoft Entra and Foundry Agent Identity).
- Establish runtime policies, including to enable human and administrative intervention: Help customers establish and enforce controls, including to enable users and administrators to interrupt, restrict or disable an agent at runtime (capabilities available in Foundry, such configuration to require human approval before specified agent tool calls are executed, and the open-source Agent Governance Toolkit, including the Agent Control Specification (ACS), which enables policy guardrails across agent actions, tool calls, and outputs).
- Protect against untrusted content and tool misuse: Limit access to untrusted resources and apply layered mitigations against risks such as indirect prompt injection (capabilities available in Foundry).
- Run evaluations pre- and post-deployment: Assess task adherence, tool selection, and tool-call accuracy, end-to-end outcomes, and the effectiveness of safeguards (capabilities in Foundry and open-source tools, including ASSERT and Clarity).
- Observe agent behavior and interactions, identify anomalies, and support investigation: Trace and log behaviors, events and metrics, such as agent actions, communications, tool calls; identify anomalous activity or divergence from an agent's intended role or objective; and investigate unexpected behavior (capabilities available in Foundry and Azure Monitor).

Please briefly describe your organization's decision-making process for deploying an AI model or system including: who has decision authority; what evidence and criteria are required (e.g., performance, safety/security, robustness, fairness, privacy, compliance, independent external testing, evaluation, validation and verification (TEVV) or assurance); contextual caveats or limitations required in interpreting risk evaluation metric results; what approvals/sign-offs are needed; and what safeguards (e.g., mitigations, monitoring, or phased release) must be in place to move into deployment.

Microsoft's deployment decision-making process varies by the risk profile, consistent with Microsoft's Responsible AI Standard, which was updated in 2026 to better reflect the roles Microsoft plays in developing or deploying the models, platform services, or applications with different risk profiles.

Decision authority. For frontier models subject to the Frontier Governance Framework (FGF), as described in our response to question 1A, the decision to deploy can only be made by Executive Officers or their delegates, based on a Recommendation for Secure and Trustworthy Deployment. For technologies or use scenarios triggering Sensitive Use criteria (including consequential impacts on legal status or life opportunities, risks of significant physical or psychological injury, or risks to human rights), the Sensitive Uses and Emerging Technologies team within the Office of Responsible AI works with relevant internal teams to assess risks, develop additional mitigations where needed, and determine whether Microsoft can proceed consistent with our principles and Responsible AI Standard. For other generative AI products, Microsoft implements risk-based launch-readiness reviews that may include centralized review processes or reviews by experts within our responsible AI governance community with product teams. From July 2025 to June 2026, we conducted pre-release reviews for nearly 2,500 generative AI features or products.

Evidence and criteria. Depending on a technology's or use scenario's risk profile, Microsoft may require pre-deployment evaluations across performance, safety, security, robustness, fairness, and privacy. For frontier models, a leading indicator assessment using state-of-the-art benchmarks screens for early indicators of tracked high-risk capabilities; triggering indicators then lead to a deeper assessment involving capability evaluation, capability elicitation, and holistic risk assessment, resulting in a risk classification of low, medium, high, or critical against capability thresholds.

Under Microsoft's Frontier Governance Framework, models within scope undergo independent red teaming by Microsoft's AI Red Team (AIRT). Where specified gating criteria are met, including based on the model's assessed capabilities, risk profile, and relevant domain-specific risks, Microsoft also engages qualified and independent third-party evaluators to conduct additional testing.

Contextual caveats. Evaluation results should be treated as decision-relevant evidence. Microsoft's own [public research](#) recognizes that AI evaluation remains an emerging science: results can be sensitive to how constructs, tasks, contexts and metrics are defined, and performance observed in controlled benchmarks may not generalize reliably. Microsoft therefore draws on multiple, complementary sources of evidence across the lifecycle, including pre-deployment benchmarks, capability and safeguard testing, red teaming and qualitative expert assessment, complemented after deployment by ongoing monitoring, incident and user feedback, and re-evaluation as systems, capabilities or deployment contexts change, as reflected in Microsoft's [2026 Responsible AI Transparency Report](#).

Approvals and safeguards. As described in our response to question 1A, frontier model deployment requires tailored state-of-the-art safety and security mitigations applied before deployment and re-assessment post-mitigation. AI technologies and use scenarios that qualify for Sensitive Use reviews or generative AI release reviews may be required to implement various safeguards depending on risk profile, including staged release, content safety guardrails and prompt injection defenses, off-by-default settings, human review in agentic workflows, in-product disclosures, and broader transparency and feedback mechanisms [e.g., [Smart Impression](#) and [Copilot Voice](#) case studies].

Following deployment, how do you use incident reports — including those shared by other organizations — to identify emerging issues, update risk assessments, and trigger corrective actions.

Microsoft implements multiple channels for collecting user feedback (including through in-product mechanisms), reports of emerging issues or potential vulnerabilities or incidents, and other signals. Microsoft employees are also provided with multiple avenues to raise concerns, including an anonymous reporting channel.

Initial concerns reported or routed to the [Microsoft Security Response Center \(MSRC\)](#), including via the [researcher portal](#), are triaged and assessed by expert teams. If these concerns are assessed to warrant an investigation and potential incident response, appropriate teams are assembled and coordinated by response specialists, who manage root-cause analysis, mitigation, and communication. After an incident is mitigated, a postmortem analysis is conducted to assess and absorb the learnings from the event and convert them into long-term improvements in system robustness.

We incentivize the responsible disclosure of issues of concern, including risks and potential vulnerabilities and incidents, through our commitment to [Coordinated Vulnerability Disclosure](#); [bug bounty programs](#), which we have extended to [AI products](#); and efforts to build community and offer public recognition through conferences like [BlueHat](#) and the [Microsoft Researcher Recognition Program](#). Microsoft also draws on lessons shared through the Frontier Model Forum (FMF), where members exchange information on frontier-model vulnerabilities, threats and capabilities of concern; these insights can inform internal evaluations, safeguards and threat detection (more details in our response to question 3D).

3. Does your organization have mechanisms to receive and/or share reports of risks, incidents or vulnerabilities by/with third parties?

Yes, both receiving and sharing

If yes, please indicate which stakeholders reporting mechanisms may be accessible to, recognising specifics will depend on the nature of the report and risk level. For each, indicate whether access is on a standing basis (Always), activated based on context (Ad hoc/as relevant), or not currently accessible.

	Always	Ad hoc / as relevant	Not currently accessible
Relevant employees (security, legal/compliance, etc.)	☐		
Public authorities (e.g., public reporting offices for security incidents or data breaches)		☐	
Supervisory authorities (e.g., data protection authorities)		☐	
Industry partners or industry fora		☐	
Users (end users, enterprise customers, developers using APIs)	☐		
Business customers / deploying organizations (if different from “users”)		☐	
Supply-chain partners (cloud, data providers, etc.)		☐	

	Always	Ad hoc / as relevant	Not currently accessible
Academia / researchers		☐	
Independent auditors		☐	
Standards bodies / multi-stakeholder initiatives (OECD.AI, ISO/IEC, IEEE, etc.)		☐	
Specific civil society organizations		☐	
Media / general public (people or communities indirectly impacted)		☐	
Other			

Please describe how your organisation determines when and how reporting mechanisms are made accessible to different stakeholders (e.g. based on legal or regulatory requirements, internal policies, or risk level). Please also provide public URLs or details on how to submit reports. If none, write “Not applicable”

Microsoft makes reporting mechanisms available according to the nature and severity of the issue, the stakeholders best placed to identify it, and applicable legal, regulatory, and internal policy requirements. Customers and the public can report concerns relating to Microsoft’s Responsible AI principles through the Microsoft Security Response Center’s Report Abuse Portal, while security researchers can submit and track potential vulnerabilities through the [MSRC Researcher Portal](#) under Microsoft’s Coordinated Vulnerability Disclosure process.

Where vulnerabilities fall within defined program scopes and demonstrate significant security impact, researchers may also be eligible for [Microsoft’s Bug Bounty Programs](#), which have been extended to cover AI products. For trusted security providers, the [Microsoft Active Protections Program](#) supports coordinated protection by sharing actionable vulnerability information and detection guidance ahead of public security updates. These differentiated mechanisms allow reports to be directed to the appropriate review, investigation, remediation, and disclosure processes based on the type of concern, potential impact, and reporting stakeholder.

How does your organization share relevant information as appropriate?

- Publicly
- Information-sharing partnerships (with confidentiality agreements as appropriate)
- Reporting to regulatory authorities as relevant

Please describe what type of information is shared.

Microsoft shares information across a spectrum of scenarios, including incident reports to regulatory authorities as required by law or regulation in the markets we operate in; operational or technical information with relevant peers, partners, or government authorities (e.g., jailbreak techniques to impacted industry organizations); and publicly and with customers to support understanding of our responsible AI practices and compliance. For example, we publish an annual Responsible AI Transparency Report (third annual version here: [2026 Responsible AI Transparency Report](#)), and as

part of our commitment to security, we launched our [Secure Future Initiative](#) and provide public reports on our progress every quarter ([see our latest progress report from July 2026](#)).

Microsoft is part of the Frontier Model Forum's first-of-its-kind [information-sharing agreement](#), which helps build a shared evidence base across three defined areas: frontier model vulnerabilities, including jailbreaks, data poisoning, and attempts to bypass safeguards; threats involving unauthorized access or manipulation, including relevant threat actors and attack vectors; and capabilities of concern in advanced cyber, CBRN, and model autonomy. The agreement provides a legally reviewed, controlled channel for exchanging sensitive technical information and supporting stronger evaluations, red teaming, threat detection, safeguards, incident preparedness, and safety research.

Through the [Microsoft Active Protections Program \(MAPP\)](#), Microsoft shares security vulnerability information with members in advance of monthly security updates, enabling faster downstream protection. Microsoft shares common threats and research with other AI model providers as appropriate. Where ecosystem-level risks are identified, Microsoft publishes technical write-ups of emergent attack techniques and mitigations: Crescendo and Skeleton Key are representative examples. When products of other organizations are affected by risks Microsoft identifies, Microsoft shares information directly with impacted organizations consistent with Coordinated Vulnerability Disclosure principles, including through the [Microsoft Security Vulnerability Research program \(MSVR\)](#). Microsoft is also one of twelve launch partners in [Project Glasswing](#), a cross-industry initiative focused on using advanced AI capabilities to identify and help remediate vulnerabilities in important software and participating organizations' codebases. Similarly, OpenAI and Microsoft work closely through [Trusted Access for Cyber program](#), and we support OpenAI's use of scoped, early deployments for safety and security testing. Microsoft currently supports structured access for certain cybersecurity, safety and research use cases through related [exception mechanisms](#). These arrangements may enable approved participants to conduct defensive vulnerability analysis, controlled safety and security research, and internal defensive cybersecurity operations within defined technical and governance conditions.

Together, these channels enable information arising from incidents, vulnerabilities and emerging threats to inform corrective action within Microsoft while also strengthening evaluations, mitigations and preparedness across the broader AI ecosystem.

4. If your organization has programs to incentivize risk(s) disclosure, incidents and systemic vulnerabilities, please share what type of programs you use?

- Paid rewards ("bug bounty")
- Vulnerability Disclosure Programs (VDPs) with safe harbor / liability protection
- Recognition and reputation incentives
- Early access and collaboration incentives
- Research grants, contracts, and hiring red teamers
- Operational incentives for customers and partners
- Internal incentives and culture

Please provide public URLs and/or details on your incentive program(s). If none, write "Not applicable."

- [Microsoft bug bounty programs](#) - Microsoft operates a broad portfolio of bug bounty programs that provide financial incentives and recognition for external researchers who responsibly identify and disclose security vulnerabilities. These programs have been extended to AI products, including through the Microsoft Copilot Bounty Program, which rewards eligible findings demonstrating a direct security impact. Submissions are handled under Microsoft's Coordinated Vulnerability Disclosure framework, enabling researchers to test systems responsibly while giving Microsoft an opportunity to investigate and remediate vulnerabilities before public disclosure.
- [Microsoft Active Protections Program \(MAPP\)](#) - Led by the Microsoft Security Response Center (MSRC), MAPP is a trusted information-sharing partnership through which qualified security providers receive vulnerability and detection guidance ahead of Microsoft's monthly security updates. This enables partners to develop and deploy protections through endpoint security products, intrusion-prevention systems and managed services before vulnerabilities can be exploited at scale. Partners can also share telemetry and threat information back with Microsoft, creating a two-way mechanism for strengthening collective threat visibility and defensive preparedness.
- [External Red Team Alliance \(EXTRA\)](#) - In 2026, Microsoft launched EXTRA as a global extension of Microsoft's AI Red Team, designed to bring independent academic, practitioner and regional expertise into AI safety and security research and testing. Microsoft provided unrestricted funding to 18 university labs across six continents to advance independent research, while a second component is building a distributed network of specialists who can participate directly in red teaming where particular technical, linguistic, cultural or domain expertise is required. The initiative is intended both to broaden the range of risks that can be identified through external scrutiny and to strengthen evaluation methodologies and testing practices for increasingly capable AI systems.

5. Does your organization use or contribute to the development of international technical standards, best practices or tools for the identification, assessment, and evaluation of risks?

Contributes and uses

If contributes or uses, please select all that apply:

- ISO/IEC standards (e.g., ISO/IEC 27001, ISO/IEC 42001, ISO/IEC 23894, ISO/IEC 42005)
- NIST AI Risk Management Framework (AI RMF)
- Industry-led best practices or codes of conduct
- Framework Convention on AI and Human Rights, Democracy and the Rule of Law (Council of Europe)
- OECD Due Diligence Guidance for Responsible AI
- General Purpose AI Code of Practice

6. Does your organization collaborate with relevant stakeholders across sectors to assess and adopt risk mitigation measures to address risks that could have broad and potentially cascading socio-economic impacts? Please indicate which stakeholders you collaborate with, and whether collaboration is on a standing basis (Always-on), activated based on context (Ad hoc/as relevant), or there is no current collaboration.

	Always-on	Ad hoc / as relevant	No current collaboration
Relevant employees (security, legal/compliance, etc.)	☐		
Public authorities / regulators / AI safety institutes / other public sector bodies and inter-governmental coordination structures		☐	
Industry partners or industry fora		☐	
Users (end users, enterprise customers, developers using APIs)		☐	
Business customers / deploying organizations (if different from “users”)		☐	
Supply-chain partners (cloud, data providers, etc.)		☐	
Academia / researchers		☐	
Independent auditors		☐	
Standards bodies / multi-stakeholder initiatives (OECD.AI; ISO/IEC; IEEE; AIID etc.)		☐	
Independent civil society organizations (e.g., for evaluations, audits, or policy review)		☐	
Media / general public (people or communities indirectly impacted)		☐	
Other(s) (please add any other relevant stakeholder groups):			
Not applicable			

Please describe examples of how your organization collaborates. Examples may include: participating in multi-stakeholder or industry initiatives; partnering with academia, civil society, or the public sector; establishing internal or joint governance structures with external experts or in partnership with other efforts to share knowledge and jointly address risks.

Microsoft collaborates with industry peers, academia, civil society, standards bodies, governments and other public-sector institutions through a range of bilateral partnerships, multi-stakeholder initiatives and institutional forums. These engagements help Microsoft draw on external expertise, contribute to shared approaches to AI risk management, and strengthen collective capacity to identify, assess and mitigate emerging risks, including those that may have broader societal impacts. Examples include:

- **Frontier Model Forum (FMF).** As a member of the Frontier Model Forum (FMF), Microsoft works with other frontier AI developers to advance shared approaches to managing frontier AI safety and security risks. Microsoft has contributed to FMF issue briefs and technical reports on topics including frontier AI safety evaluations, foundational security practices, risk taxonomies and thresholds, third-party assessments, advanced cyber risk, multilingual evaluation, AI agents, biosecurity, information

sharing, incident reporting and incident response. In March 2025, Microsoft joined other FMF members in signing a first-of-its-kind agreement to share information about vulnerabilities, threats and capabilities of concern unique to frontier AI.

- As a member of **Partnership on AI (PAI)**, Microsoft contributes substantive feedback to PAI research and guidance and participates in multi-stakeholder convenings involving industry, civil society, academia and other experts. Over the past year, this has included engagement on emerging psychosocial risks and AI's impact on human connection, including PAI's [report](#) on how AI companies are handling suicide and self-harm and agentic report on [real-time failure detection in AI agents](#) and [AI agent governance](#). This builds on earlier Microsoft contributions to [PAI's Guidance for Safe Foundation Model Deployment](#), which translates shared safety principles into practical and adaptable recommendations for model providers across the deployment lifecycle.
- Microsoft co-founded the **Coalition for Content Provenance and Authenticity (C2PA)** and has held a co-chair position since the organization's founding. Over the last year, Microsoft participated in the C2PA's ongoing efforts to advance the interoperability and real-world applicability of content provenance.
- Microsoft actively participates in **MLCommons** work on AI risk and reliability evaluation, including benchmarks that form the AILuminate suite. Recent efforts have focused on developing tools to evaluate robustness against jailbreaks, psychosocial risk, privacy of agentic systems, and multilingual and multicultural performance. Microsoft is also contributing to the [MLCommons AILuminate Global Assurance Program](#), which aims to translate benchmark results into assurance evidence 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.
- Microsoft contributes to the development of **international AI standards**, including ISO/IEC 42001:2023 and related standards, such as ISO/IEC 42005, ISO/IEC 42006, and ISO/IEC 42007. Microsoft has achieved ISO/IEC 42001 certification across a broad set of systems, including Microsoft 365 Copilot, Microsoft 365 Copilot Chat, Microsoft Foundry, Microsoft Security Copilot, Microsoft Copilot Studio, GitHub Copilot, Dragon Copilot, and Copilot Health.
- Microsoft is sponsoring the third year of the [Global Perspectives on Responsible AI Fellowship](#), implemented in partnership with the **Stimson Center**, to strengthen our understanding of how AI practices can better reflect Global South priorities and help elevate fellows' perspectives in conversations on responsible AI development. 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. 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.
- In 2025, Microsoft Research launched the [Agentic AI Research & Innovation \(AARI\) program](#), building on the strong foundation of the [Accelerating Foundation Model Research \(AFMR\)](#) initiative.

AARI supports twelve research projects led jointly by academic partners and Microsoft researchers around the world. The projects explore topics such as multi-step reasoning, improving reliability and uncertainty handling, enabling autonomous and resilient supply chains, physics-guided 4D world modelling, and cultural misalignment in AI agents.

- Microsoft supports public research through **the US National AI Research Resource (NAIRR)** Pilot by providing access to Azure infrastructure, AI models, and technical expertise. In the program's second year, Microsoft and NAIRR launched a Deep Partnership track supporting “grand challenge” research projects, including many 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.
- Microsoft is a founding member of the **Appia Foundation**, a Joint Development Foundation project administered by the Linux Foundation, launched as a cross-sector international initiative to develop practical, scalable conformity specifications grounded in international standards. The specifications are being designed to be applicable globally and able to support procurement scenarios and multiple regulatory regimes; adaptable across sectors, organizational sizes and AI use cases; clear about roles and responsibilities across the AI value chain; and suitable for conformity assessment with practical evaluation criteria and audit-ready evidence. Appia's membership includes industry across sectors, and Appia's governance includes an Advisory Committee of independent leaders from academic and civil society.
- Microsoft's engagements with **AI Safety and Security Institutes (AISIs)** from multiple regions and **the US Center for AI Standards and Innovation (CAISI)** are structured around specific evaluation and research areas. Further details are available in our responses to questions 16.A and 24.A.

Any further comments and implementation documentation.

No answer provided

Section 2 – Risk Management and Information Security

7. Across the AI system lifecycle, what steps does your organization take to manage risks and vulnerabilities? Please indicate which specific practices, tools or metrics your organization uses at each lifecycle stage. You may refer to tools or metrics present in the OECD.AI Catalogue of Tools and Metrics or suggest additions. Respondents may complete only the lifecycle stages relevant to their organisation's role. Stages that do not apply may be left blank.

Microsoft InterpretML

Counterfit

PyRIT

EUREKA: Evaluating and Understanding Large Foundation Models

Microsoft's Responsible AI Toolbox

Microsoft Agent Governance Toolkit

Plan and design

Microsoft InterpretML

Collect and process data

Build, integrate and interpret model

Verify and validate

Deploy

Operate and monitor

Please provide a consolidated description of your organisation's approach to risk management across the AI system lifecycle, including how the practices identified above operate together and any cross-cutting governance, review or escalation processes. Please also indicate how often these practices are reviewed and what types of events trigger reviews and updates. Where relevant, please distinguish between different types of AI systems (e.g. generative or agentic AI).

Microsoft applies a comprehensive and layered risk management approach across the AI lifecycle, grounded in the Responsible AI Standard and supported by scenario-specific requirements for particular technologies, capabilities, and use cases. This approach is designed to uphold Microsoft's AI principles consistently across the development lifecycle by defining policies, assigning roles and responsibilities, enabling proactive risk management, and establishing processes that reinforce responsible AI by design.

How the practices operate together. Microsoft draws on the Govern, Map, Measure, and Manage functions of the [NIST AI Risk Management Framework](#). Consistent with this approach, teams identify reasonably foreseeable risks and benefits; use red teaming, evaluations, and other measurement methods to assess risks and the effectiveness of mitigations; implement layered technical and operational safeguards; and document the evidence supporting deployment decisions. Measurement and mitigation are iterative: evaluation results inform mitigations, and systems are reassessed following mitigation to determine whether residual risks have been appropriately addressed.

The Responsible AI Standard serves as Microsoft's internal playbook for building AI systems in alignment with its AI principles. Core requirements apply across Microsoft's AI technologies, while additional requirements and review processes are applied according to the technology's capabilities, intended uses, and risk profile. These include requirements for generative AI systems and models; the Frontier Governance Framework for advanced model capabilities that could pose national security or at-scale public safety risks; and the Sensitive Uses and Emerging Technologies program for higher-impact uses and novel technologies or risks.

Pre-deployment review connects these practices and supports consistent release decisions. Product teams document how applicable requirements have been met, including the risks identified, evaluations undertaken, mitigations applied, and any residual risks. Relevant experts review higher-risk or higher-impact uses and generative AI releases, and teams may be required to conduct further evaluation or implement additional safeguards before release. Depending on the risk profile, safeguards may include post-training interventions, system prompts, input and output classifiers, human review, transparency documentation, user feedback mechanisms, and phased or staged release. Post-deployment monitoring, incident reporting, user feedback, and real-world deployment experience then provide signals that can prompt reassessment and feed into improvements to products, engineering practices, and policy.

Governance and escalation. Product teams manage risk for the technologies they build. The Office of Responsible AI maintains the Responsible AI Standard and leads the development of new policies and guidance, informed by regulatory developments, pre-deployment reviews, red teaming, product

roadmaps, research, ongoing monitoring, and incident response. The Responsible AI engineering function develops supporting tools and technical guidance so that policy requirements can be integrated into engineering workflows. Deployment of a frontier model rests with Executive Officers or their delegates, as described in our response to question 2.A. After release, escalation follows the severity assigned under the AI Security and Safety Bug Bar (see 23.A). Oversight structures, including the Responsible AI Council and reporting to the Board, are described in our response to question 21.A.

Review and updates. Microsoft reviews and updates its responsible AI policies, practices, tools, and individual risk assessments on an ongoing and event-driven basis. Relevant triggers include new or materially changed system capabilities; major product or deployment changes; evaluation, red-teaming, or monitoring results; reported vulnerabilities or incidents; evidence from real-world use and user feedback; advances in risk measurement or mitigation; research and horizon-scanning findings; and changes in applicable laws, standards, or regulatory expectations.

The [Frontier Governance Framework](#) includes additional review points for models within its scope. Leading-indicator assessments are conducted during pre-training, after pre-training, after post-training, and before deployment, and at least every six months thereafter for models within scope. A further capability assessment is conducted where leading indicators warrant deeper inspection, with subsequent assessments triggered by material changes to a deployed model's risk profile. Microsoft also commits to an explicit discussion at least every twelve months on how the FGF should be improved, with material revisions published within 30 days of adoption.

Generative and agentic AI. Microsoft formalized specific internal requirements for generative AI systems in 2023 and has continued to update them as capabilities, risks, and regulatory requirements evolve. Generative AI releases are subject to pre-deployment review, red teaming, evaluation, mitigation, and documentation requirements tailored to their capabilities and deployment context. Agentic AI requires additional attention because systems may connect multiple models, tools, data sources, permissions, and memory, and may take actions with increasing autonomy over time. Microsoft therefore addresses agentic risks through system-level threat modelling, evaluation of autonomy, planning and tool use, controls governing access and permissions, and safeguards designed around the system's deployment context. Where agentic capabilities may indicate advanced autonomy or other tracked high-risk capabilities, the [Frontier Governance Framework](#) provides for leading-indicator assessment, deeper capability assessment, proportionate mitigation, and risk-informed deployment decisions.

Please describe any measures used to ensure data quality and fairness across the data collection and curation stage (e.g., dataset audits, filtering training data to remove harmful content or personal information, fine-tuning on curated data, use of moderation tools, assessment of limitations to model generalisability).

Microsoft addresses data quality and fairness through layered practices spanning data selection and curation, safety post-training, evaluation, red teaming, documentation, and monitoring. The specific measures applied depend on the model, modality, intended use, and risks identified.

Dataset assessment and limitations. Teams assess relevant characteristics and limitations of the data used to develop and evaluate an AI technology. Depending on the model or system and its intended use, this can include assessing the suitability and inclusiveness of datasets, whether intended uses are adequately represented in training and test data, and whether applicable data collection and

processing requirements have been met. Identified limitations inform evaluation priorities, mitigation measures, deployment decisions, and documentation for downstream users.

Safety post-training and content safeguards. Training-data interventions are complemented by safety post-training and product-level safeguards. For the MAI-Image models, Microsoft used safety post-training to shape model behaviour and applied safety classifiers to prompts and outputs to identify and restrict harmful content. The product team also used controls designed to reduce the generation of content that could infringe third-party intellectual property rights. In operation, meta-prompts, classifiers and output controls limit potentially harmful or infringing generation, including refusal to provide song lyrics or book extracts, and customer data is processed to analyze prompts, completions and images for content or patterns of use that may violate Microsoft's Code of Conduct. The same content safety and moderation tooling is made available to customers through Azure AI Content Safety. These safeguards were assessed through pre-release testing and continued monitoring after deployment.

Filtering of training data. Safety-focused filtering is applied to training datasets to reduce a model's exposure to harmful content. For the MAI-Image models, for example, categories including explicit violence and child sexual exploitation and abuse imagery were removed from training data, so that the models were less likely to learn detailed or realistic representations of such content. Microsoft also respects the Robots Exclusion Protocol and Bing's webmaster HTTP head and HTML metatag control. Microsoft is also investing in broader linguistic and cultural coverage for AI evaluation, including work to expand AI reliability measurement across additional languages and cultural contexts. These efforts strengthen the evidence available to assess performance and safety for populations and contexts that may be underrepresented in existing evaluation datasets. Additional details on such initiatives are described in our responses to question 2 and 28.A

8. What security measures are used in testing AI systems?

- Isolated and secure test environments
- Network isolation from external access
- Locally hosted testing environments
- Restrictions on outbound Internet or API access, including a list of permitted domains or APIs
- Isolation of staging environments (from production / operational environments)
- Controlled environments for red-teaming and adversarial testing
- Containment measures for prompt-injection and jailbreak attempts
- Controlled security vulnerability testing
- Rate limits and resource caps
- Human-in-the-loop controls for higher-risk test cases
- Access controls, including roles-based permissions
- Disabling, restricting or simulating high-impact capabilities during testing
- Audit logging and monitoring
- Use of "AI-safe" test-data, as appropriate
- Use of synthetic, anonymized, or dummy data, as appropriate

8.A. Please indicate, if relevant, the team or role responsible for overseeing secure testing and whether this responsibility is shared across multiple teams and/or functions.

Secure testing responsibility is shared across product engineering, responsible AI, and security, with testing protocols depending on the risk profile of the testing scenario, including the technology capability and evaluation approach. Microsoft's AI Red Team leads testing for the highest-risk systems and works with other security teams to assess testing risks and apply appropriate security measures for the profile of the testing scenario.

Internal testing responsibilities. Product and engineering teams conduct testing and evaluation as part of the development and pre-release review process. These teams are responsible for evaluating relevant risks, assessing the effectiveness of safeguards, documenting results, and addressing identified issues before release. Microsoft's AI Red Team provides specialist adversarial testing of AI models and systems, including human-led testing intended to identify potential misuse, harmful outputs, security vulnerabilities, and weaknesses in existing safeguards. In 2025, the AI Red Team conducted more than 100 red-teaming operations, representing a 1.5-fold increase over the previous year. Testing combines human-led and automated approaches. Microsoft uses PyRIT, the [Python Risk Identification Tool](#) for generative AI, which supports structured and repeatable red teaming across AI models and systems. PyRIT supports testing across different modalities and interfaces and has also been integrated into Microsoft Foundry through an AI red-teaming agent. Findings from testing inform mitigations and can be converted into repeatable evaluations to assess whether identified weaknesses have been addressed.

Responsible AI oversight. Microsoft's central pre-release oversight process verifies that identified risks have been appropriately evaluated and addressed before release. The Office of Responsible AI establishes applicable policy requirements and review processes, while responsible AI engineering teams develop tools and technical guidance to support implementation. Where a use meets the criteria for review under the Sensitive Uses and Emerging Technologies program, specialist reviewers provide hands-on consultation to identify relevant risks, define additional requirements, and assess whether further testing or mitigation is needed.

External testing partnerships. Microsoft supplements internal testing with external evaluators where independent or specialized expertise is needed. Over the last year Microsoft worked with ten firms holding general AI risk and red teaming expertise and with eight firms and government research institutes with deep expertise in high-severity risk areas, including biological, chemical and cyber weapons (see 24.A). External evaluators, including the UK AI Security Institute (AISi) and the US Center for AI Standards and Innovation (CAISI) under the agreements described at question 2, require credentials to send API requests for testing access. In 2026, Microsoft launched the [External Red Team Alliance](#) (EXTRA), which provides funding to 18 university labs across six continents and through which Microsoft is building a distributed network of researchers, practitioners and regional specialists who can contribute directly to red teaming where specialized expertise is needed.

Containment controls. Where possible, the Microsoft AI Red Team conducts testing in secure virtual machines and manages testing reports and results in an access-controlled environment that requires a customized access request for each operation. Testing environments are tailored to the purpose and risk profile of the assessment. Certain cyber evaluations require controlled access to cloud services, enterprise infrastructure, identity systems, or other external services to assess how AI systems operate in realistic environments and identify risks that would not emerge in a fully isolated setting. Where an assessment requires the system under test to retain legitimate access to external services, containment

is complemented by monitoring designed to identify activity beyond the engagement's authorized scope.

9. How does your organization protect intellectual property, including copyright-protected content?

9.A. Safeguards that promote lawful data use and transparency in AI training:

- Maintain a copyright policy that addresses data sourcing, content provenance, safeguards, and organizational roles and responsibilities
- Publishing "model cards" or contextual information on training data and model limitations
- Apply provenance controls to guard against training data being sourced from websites or repositories that have been determined to engage in large-scale piracy
- Refrain from the unlawful circumvention of technological protection measures in data sourcing
- Maintain policies or practices to respect copyright holders' preferences through, where appropriate/applicable, mechanisms like Robots.txt "do not train" tags or other machine-readable protocols developed through voluntary, industry-led standards processes
- Conduct data cleaning, qualification, and/or tokenization activities as part of a transformative data preparation process
- Engage in AI training that comprises a non-consumptive use of any underlying data
- Consult and/or retain an intellectual property attorney to protect, enforce, and comply with, intellectual property rights.

9.B. Safeguards that reduce the risk of dissemination of outputs that may infringe copyright:

- Restricting outputs that raise ethical, human rights, or legal concerns—such as non-consensual intimate imagery, unauthorized deepfakes, or outputs that are substantially similar to copyrighted works.
- Filtering results using keyword or image-based restrictions to reduce risks of misuse

10. How do you protect privacy and prevent your systems from disclosing confidential or sensitive data across the AI lifecycle?

- Make information about data privacy, practices, and controls accessible to a diverse set of stakeholders
- Use security controls, including encryption at rest and in transit
- Provide organizational customers with controls to configure and implement privacy choices (e.g., data retention, logging, access, or processing settings)
- Apply Privacy-Enhancing Technologies (PETs) where applicable (e.g. differential privacy, federated learning, synthetic data generation, homomorphic encryption)

Describe key privacy measures and data-protection standards applied.

Microsoft applies privacy and data-protection measures across the AI lifecycle, from development and pre-release review through deployment, monitoring, and research. The measures applied are tailored to Microsoft's role as developer or deployer, the relevant layer of the AI technology stack, the types of data involved, and the risks presented by the system.

Privacy-by-design governance. Privacy and security are among Microsoft's core AI principles and are operationalized through the Responsible AI Standard and related privacy and security requirements. The updated Standard distinguishes Microsoft's responsibilities as an AI developer and deployer and tailors requirements across models, platform services, and applications. Privacy risks are considered throughout the lifecycle, including during system design, risk assessment, pre-release review, deployment, and post-release monitoring.

Security controls. Training data and fine-tuned models are available exclusively to the customer, are encrypted at rest by default using AES-256 with the option of a customer-managed key, and can be deleted by the customer at any time. Model weights and related sensitive assets are encrypted at rest and in transit, with more stringent controls for highly capable models (see 11.E). Microsoft applies international information security standards including ISO/IEC 27001, to which Microsoft online services are certified through regular independent audit, and ISO/IEC 27017, 29147 and 30111. Microsoft Foundry supports geographic data-residency choices.

Customer and user controls. AI systems operating over organizational data are bound by the permissions already in place: Microsoft 365 Copilot respects sensitivity labels and Microsoft Purview policies, which can restrict its access to specific data or prevent it from summarizing sensitive files (see 7). Microsoft Purview provides further data security and compliance controls for generative AI applications. Through the Microsoft Privacy Dashboard, account holders can access, manage and delete their personal data and stored conversation history, and Copilot honors users' privacy choices made through cookie banners and in-product controls.

Privacy-enhancing research techniques. Where AI usage data informs research, deidentification protocols are applied, including automated scrubbing of personally identifiable information and machine-based "eyes-off" classification, so that no human researcher directly views the content of conversations; the analyses of Copilot usage described at 20.A were produced this way. Post-release monitoring and automated threat detection operate on privacy-reviewed observability platforms.

Transparency and informed use. Microsoft provides privacy statements, transparency documentation, model cards, application cards, and other notices to explain how relevant AI technologies use and protect data, their intended uses and limitations, and the controls available to users and customers. The applicable privacy policies and product-specific notices covering personal data, user inputs, and AI-generated outputs are listed in the response to question 17.A.

Transparency resources. Microsoft maintains publicly accessible materials including the Microsoft Privacy Statement, M365 Copilot FAQs, The New Bing: Our Approach to Responsible AI, Azure OpenAI Service data-privacy documentation, and Dynamics 365/Power Platform Copilot data security and privacy FAQs.

11. How does your organization assess cybersecurity risks and implement policies to enhance the cybersecurity of advanced AI systems across the AI system lifecycle? Respondents may complete the checkbox subsections in 11.A–11.D or describe their overall approach under 11.E.

11.A. Governance, policy and assessments

- Conduct regular cybersecurity risk and vulnerability assessments tailored to AI models or systems
- Perform threat modeling that includes AI-specific risks (e.g., data poisoning, prompt injection)

- Maintain or update security policies aligned with recognized standards (e.g., NIST CSF, ISO/IEC 27001, SOC 2, CIS Critical Security Controls), and review them regularly in response to emerging risks or regulatory guidance
- Maintain vulnerability management processes (scanning, channels for external reporting, patching, remediation)
- Implement incident response, disaster recovery, or business continuity plans, including escalation procedures for high-severity incidents
- Train staff on cybersecurity best practices and AI-specific threats, including how to recognize and respond to misuse or unintended outcomes of advanced AI systems
- Integrate cybersecurity AI risks into broader cybersecurity risk management activities
- Regularly review security measures and update them in response to emerging risks, incidents, or major releases

11.B. Technical and operational controls

- Apply multi-factor authentication (MFA) and role-based access controls for advanced AI systems
- Encrypt AI assets (datasets, models, configurations) in storage and in transit as appropriate
- Restrict physical access to servers, GPUs, or devices hosting AI models, systems, and/or applications
- Monitor system activity and log access to detect anomalies or misuse
- Conduct secure AI development and/or deployment practices
- Conduct security testing focused on AI-specific risks (e.g., red-teaming, adversarial testing, stress tests)
- These responsibilities are delegated contractually to model or infrastructure providers

11.C. Protection of sensitive or high value AI assets

- Limit access to proprietary AI assets based on role or need-to-know
- Store model weights, datasets, and algorithms in secure, access-controlled environments
- Use hardware security modules (HSMs) or trusted execution environments (TEEs) where feasible
- Track and log access to sensitive AI resources
- Require NDAs or confidentiality agreements for staff and contractors with access
- These responsibilities are delegated contractually to model or infrastructure providers

11.D. Insider risk protection detections

- Conduct pre-employment screening or role-sensitive vetting
- Monitor for unauthorized internal access or suspicious activity
- Provide staff training on cybersecurity best practices and AI-specific threats
- Implement insider-risk escalation or response procedures
- These responsibilities are delegated contractually to model or infrastructure providers

11.E. Please describe the main security measures your organization applies and indicate which are specific to particularly sensitive AI assets, where relevant.

Microsoft's approach to AI cybersecurity builds on its established security-engineering practices and adds requirements, testing methods, and safeguards tailored to the distinctive risks of generative and

agentic AI systems. These measures operate across the AI lifecycle, from threat modelling and secure development through pre-release testing, deployment, monitoring, and incident response.

Security engineering. Microsoft's foundational security policy is the [Security Development Lifecycle](#) (SDL), which is regularly updated in response to emerging risks, techniques and technologies. New AI-focused requirements have been implemented within the SDL, covering threat modeling, agent identity management, access control, memory safeguards, logging and reliable shutdown mechanisms. Secure AI development workflows govern the handling of access credentials and sensitive data through identity and access management controls and automated scanning. These practices operate within the [Secure Future Initiative](#), Microsoft's multiyear company-wide security program launched in November 2023.

Defenses against AI-specific threats. Microsoft applies layered defenses that operate both before information reaches a model and before an agent acts. At the product level, enhanced classifiers, including [Prompt Shields](#), scan inputs and outputs in real time for harmful content and prompt injection attempts, including indirect injection through content a model reads. Spotlighting, developed by Microsoft researchers and available in Microsoft Foundry, helps models distinguish trusted system and user instructions from potentially adversarial embedded content. For agents, [Task Adherence](#) evaluates whether an agent's planned actions remain consistent with the user's intent and the developer's design before those actions are executed, returning signals that developers can use to block execution, log the event or trigger human review. At the organizational policy and practice level, fine-grained permissions and access controls limit the impact of any injection attempt that succeeds, alongside threat modeling and other security practices. Question 7 provides further detail on threat modelling for agentic AI.

Microsoft's AI Red Team and researchers have deepened expertise on novel attack vectors, including adversarial memory poisoning, techniques to detect backdoors in open-weight models, and AI recommendation poisoning, a threat class identified by Microsoft researchers in which hidden instructions embedded in links bias an AI system's future recommendations without the user's awareness. Alongside ISO/IEC 27001 and ISO/IEC 27017, Microsoft uses ISO/IEC 23894, NIST SP 800-218 and MITRE ATLAS as practical references for prompt injection, poisoned context, unsafe tool use and data access risks.

Agent identity and runtime control. At the product level, [Microsoft Entra Agent ID](#), generally available since April 2026, binds each agent's actions to a distinct identity and allows organizations to manage credentials, enforce Conditional Access and least-privilege policies, maintain sign-in and audit logging, and monitor for anomalies or orphaned agents. At the organizational policy level, runtime policies can be configured to define which agents may invoke which tools, access which resources and perform which actions under specified conditions, including approval gates for sensitive or irreversible actions and administrative controls to activate, block or revoke agents. Observability and monitoring capabilities for agents are described at question 2 and 14.A.

Protection of sensitive AI assets. Model weights and other relevant elements are encrypted at rest and in transit to mitigate model theft risk. More stringent security controls are applied based on risk, including for protecting highly capable models. Robust physical, operational and network security measures protect cloud infrastructure, including AI datacenters enabling training and hosting of models; supplier security and privacy are governed by [Microsoft's Supplier Security and Privacy Assurance program](#). Access to physical datacenter facilities is tightly controlled with outer and inner perimeters, subject to a least-privileged-access policy with time-limited access for personnel with approved business need; access requests are logged and retained, with data analyzed for anomalies.

Detection, response and threat landscape. AI systems are developed and deployed with capabilities for observability and security and event logging, supporting privacy-reviewed post-release monitoring, incident response and auditability. In 2025, the [Microsoft Security Response Center](#) received over 11,000 submissions, of which more than 600 escalated to full AI safety or security cases. The most frequent subjects were malicious use, CBRN risks and command injection. Prompt injection remained frequently reported, with notable growth on coding-assistant agents and agentic code-generation workflows and a rise in multimodal techniques using audio, images and QR-encoded inputs. Insights from these cases informed updates to guardrail tuning. Incident response processes and escalation are described at 23.A. Microsoft's Threat Intelligence community comprises more than 10,000 security researchers, analysts and threat hunters, and security standards are reviewed at least annually, with individual requirements updated between reviews. Microsoft products undergo third-party security assessments reflected in compliance offerings for ISO/IEC 27001, ISO/IEC 27017, SOC 2 and other programs.

Insider risk. The Digital Security and Resilience Insider Threat Team (DSR ITT) focuses on insider threat detection and prevention, and Microsoft's SDL threat modeling assumes insider threats, informing privileged-access requirements. All red team operations assume insider access.

12. What is your organization's vulnerability management process? Does your organization take actions to address identified risks and vulnerabilities, including in collaboration with other stakeholders?

- Internal vulnerability disclosure receipt and tracking processes
- Bug bounty or incentive programs for external reporters
- Investigation of reported or identified risks and vulnerabilities
- Development of patches or other mitigations for confirmed risks or vulnerabilities
- Testing of patches or other mitigations
- Scanning for and mitigation of similar risks or vulnerabilities
- Collaborative vulnerability sharing (e.g., with partners, industry groups)
- Vulnerability disclosure program that is aligned with internationally recognized standards, e.g. ISO/IEC 30111 and 29147.
- Reporting to regulators / authorities when relevant
- Recognition of external finders of confirmed risks or vulnerabilities
- Defined escalation criteria and follow-on actions for identified vulnerabilities (e.g., for external reporting or information sharing)

13. What post-deployment practices does your organization use to monitor and respond to vulnerabilities, incidents, emerging risks and misuse of advanced AI systems?

- Conduct post-deployment monitoring
- Perform root-cause analysis and apply lessons learned to improve future development and/or deployment practices

- Monitoring and emergency controls specific to agentic AI systems (e.g., ability to interrupt autonomous action sequences, override tool-use permissions, or revoke delegated authority)

14. Do you develop or deploy agentic AI systems (systems that can autonomously plan/act or use tools)?⁴

Yes

14.A. If yes, please indicate which categories of controls your organization applies. Please select all.

- Constraints on the agent's action space (e.g., tool-use permissions, environment sandboxing, allowlists/blocklists of resources, scope limits per task)
- Human oversight or interruptibility mechanisms (approval requirements for sensitive actions, escalation paths, interruption interfaces)
- Monitoring (anomaly detection, multi-agent interaction monitoring)
- Logging of agent actions

Please provide any further comments and implementation documentation.

Additional information supporting Question 9

Microsoft protects intellectual property, including copyright-protected content, through measures applied across the AI lifecycle: governing how training data is sourced and documented; evaluating and mitigating the risk that models reproduce or generate infringing content; and considering feedback from rightsholders when they report concerns.

Training data sourcing. Microsoft maintains a copyright policy and supporting guidance, covering data sourcing, content provenance, safeguards, and organizational roles and responsibilities. The policy establishes requirements for models entering the EU market, while the underlying data-sourcing controls are applied more broadly across relevant data-collection processes. Provenance controls guard against sourcing training data from websites or repositories determined to engage in large-scale piracy, and Microsoft does not circumvent technological protection measures when sourcing data. Microsoft also respects the Robots Exclusion Protocol and Bings webmaster HTTP head and HTML metatag control.

Training data transparency. Microsoft publishes information about the data used to train relevant models through data summaries and model cards. During the reporting period, Microsoft published data summary cards for relevant general-purpose AI models placed on the EU market in accordance with Article 53(1)(d) of the EU AI Act, including for MAI-Image-2, MAI-Image-2.5, and MAI-Code-1-Flash. Microsoft also published disclosures for in-scope models, including MAI-Image-1 and Phi-4-Reasoning, ahead of the effective date of California Assembly Bill 2013, and published information for MAI-Image-1 in accordance with the Transparency Chapter of the EU General-Purpose AI Code of Practice (see 26.A). These disclosures provide rightsholders and other stakeholders with greater visibility into the categories and sources of content used in model development.

Evaluation and safeguards against potentially infringing outputs. Microsoft applies layered safeguards at the model and product levels to reduce the risk that its AI technologies generate content that may infringe third-party intellectual property rights. These include training-data measures, pretraining data processing, post-training interventions, system-level instructions, and input and output controls designed to identify and restrict requests or generated content that violates applicable intellectual property policies. For MAI-Image-1

and MAI-Image-2, Microsoft assessed risks related to potentially infringing outputs through pre-release evaluation and applied safeguards across the model and product lifecycle. This layered approach allows safeguards to operate at multiple points, from the composition of training data and the behavior of the model to the processing of user requests and the delivery of generated content.

Monitoring and rightsholder recourse. Microsoft continues to assess relevant safeguards after release through monitoring, user feedback, and reports of potentially problematic generated content. Consumer generative AI experiences provide mechanisms through which users and rightsholders can report concerns. These signals can inform investigation, corrective action, and improvements to model behavior, classifiers, product controls, and applicable policies.

Section 3 – Transparency Reporting on Advanced AI Systems

15. Does your organization publish public reports and/or technical documentation related to the capabilities, limitations, and domains of appropriate and inappropriate use of advanced AI systems? Are these clear and understandable by non-technical audiences?

Yes

15.A. If yes, which of the following information is included in your organization's publicly available documentation for any of your AI products?

- Risk management policies
- Summary of results from evaluations of safety, security, and societal risks including potential adverse impacts on individuals or groups (e.g. impacts on human rights, privacy or differential outcomes across user groups).
- Summary of results from red-teaming and other testing (e.g., adversarial evaluations, stress tests, and other pre-deployment safety assessments).
- Capabilities, behaviours, and limitations, including domains of appropriate and inappropriate use or indicative thresholds associated with systemic risks.
- Technical documentation and instructions for use if relevant.
- Plain-language or non-technical summaries

Please provide links or references showing where this information is made public.

Examples of publicly available documentation include the 2026 Responsible AI Transparency Report, model and data summary cards, and application and platform cards. Documentation on AI agents includes the [Microsoft Foundry Agent Service Transparency Note](#). Microsoft also publicly documents agent observability and tracing, including tracing of agent-to-agent interactions and multi-agent orchestration in Microsoft Foundry. See [Microsoft Foundry's agent tracing documentation](#) and Question 19 for additional examples and links.

15.B. If yes, how often are such reports usually updated?

Other (e.g., triggered by material changes, incidents, or new risks)

16. How does your organization share information with a diverse set of stakeholders (other organizations, governments, civil society and academia, etc.) regarding the outcome of evaluations of risks and impacts related to an advanced AI system?

- Public documentation
- Protected information sharing
- Regulatory reporting as relevant
- Direct engagement with AI safety institutes, civil society or academic partners

Describe how evaluation results are shared and with which stakeholders (e.g. other organizations, governments, civil society and academia, etc). You may also indicate whether non-technical or plain-language summaries are provided for broader audiences.

Microsoft shares information about AI risk and impact evaluations through a combination of public documentation, protected exchanges with government and independent evaluators, direct engagement with civil society and academia, and regulatory reporting where applicable. The level of detail and method of sharing depend on the sensitivity of the information, the stage of the AI lifecycle, and the expertise and role of the recipient.

Pre-release evaluation sharing. Microsoft supplements internal evaluation and red teaming with testing by government institutes and qualified third parties where independent scrutiny or specialized expertise is needed. This includes collaboration with the US Center for AI Standards and Innovation (CAISI) and the UK AI Security Institute (AISi) on frontier model testing and evaluation. Under these arrangements, Microsoft can provide model evaluation results, pre-release access and relevant technical information to support external assessment of model capabilities, safeguards, and national-security or public-safety risks.

Public information at release. Microsoft publishes transparency documentation to help developers, customers, users, and other stakeholders understand relevant AI technologies. Depending on the technology, this can include model cards, data summaries, and application or platform cards describing capabilities, limitations, intended uses, safeguards, and responsible use considerations. Since January 2026, Microsoft has published 20 application and platform cards using updated documentation templates. Microsoft also provides product-level disclosures and guidance designed for broader audiences. The format and level of technical detail are tailored to the audience, with public documentation complementing more detailed information shared through protected evaluation or regulatory channels.

Post-release information sharing. Microsoft continues to engage government institutes, external evaluators, researchers, civil society organizations, and industry partners such as the Frontier Model Forum (FMF) after release. Post-release monitoring, red teaming, vulnerability reporting, user feedback, and incident response can generate new information about system performance, safeguards, vulnerabilities, and emerging risks.

Annual synthesis. The [Responsible AI Transparency Report](#) provides a public account of evaluation practices and results across the year, including case studies tracing how our risk management framework has been applied to individual models and applications. The [Frontier Governance Framework](#), updated in February 2026, added public disclosures on Microsoft's red teaming and risk modeling practices.

17. Does your organization publish public privacy policies addressing the use of personal data, user prompts, and/or the outputs of advanced AI systems?

Yes

If yes, provide links to privacy policies or notices covering personal data, user inputs, and AI-generated outputs.

Foundational and cross-company privacy disclosures

[Microsoft Privacy Statement](#) Consumer-facing privacy notice

[Change history for the Microsoft Privacy Statement](#) Versioned record of notice changes

[Privacy at Microsoft \(Trust Center\)](#) Privacy principles hub.

[Microsoft Data Access Management](#) Data access controls and entry point to the subprocessor list.

[Microsoft Products and Services Data Protection Addendum](#) Contractual processing terms covering AI services.

[Microsoft EU Data Boundary](#) EU processing commitments and AI-related exclusions.

[Data for AI Training](#) Categories of data used to train Microsoft models.

Copilot and enterprise experiences

[Data, Privacy, and Security for Microsoft Copilot](#) Core description of enterprise Copilot data processing.

[FAQ for Copilot data security and privacy for Dynamics 365 and Power Platform](#) Business applications Copilot data handling.

[Anthropic models in Microsoft Online Services](#) Anthropic as a Microsoft subprocessor: retention, regions, and admin controls.

[Copilot in Microsoft 365 apps with Anthropic models](#) Anthropic as the default model setting in the EU, EFTA, and the UK.

[Data, privacy, and security for models sold by Azure in Microsoft Foundry](#) Prompt and completion handling, retention, and abuse monitoring.

[Microsoft Copilot for individuals: your data, privacy, and responsible AI](#) Consumer Copilot data use overview.

[Microsoft Copilot privacy controls](#) Model training opt-out and conversation controls.

[Privacy FAQ for Microsoft Copilot](#) Consumer questions on Copilot data.

[Privacy and control over your Recall experience](#) On-device processing for Windows Recall.

[LinkedIn: update to our terms and data use](#) Member data use for generative AI on LinkedIn.

18. How does your organization provide information about the sources of data used for the training of advanced AI systems, as appropriate, including information related to the sourcing of data annotation and enrichment?

- Public documentation
- Provide information or tools to downstream actors on data quality
- Protected information sharing
- Regulatory reporting as relevant

Please provide links to any publicly available information.

Microsoft provides information about the data used to train relevant advanced AI systems through complementary company-wide and model-specific disclosures. At the organisational level, Microsoft describes the principal categories of training data it uses, including publicly available, acquired, first-party and synthetic data, as well as human feedback from AI trainers, red teamers, Microsoft employees and external annotation partners. It also describes relevant sourcing and publisher controls.

For relevant first-party models, Microsoft publishes model cards, technical reports and training-data summaries that provide more detailed information, as appropriate, on data modalities and source categories; major public datasets; commercially acquired, web-crawled, first-party and synthetic data; collection and cut-off dates; linguistic coverage; data-selection rationale; and filtering, curation, deduplication, annotation and enrichment practices. The level of detail varies by model and reflects applicable transparency requirements, the nature of the model and data, and confidentiality obligations associated with individual datasets or commercial arrangements.

Microsoft complements these disclosures through its annual Responsible AI Transparency Report, which describes its broader documentation, evaluation and governance practices. Further information on data quality and curation is provided in response to question 7B, and information on copyright, provenance and rightsholder controls is provided in response to question 9.

Company-wide documentation

- [Data for AI Model Development](#) (Microsoft Trust Center) — categories of training data (publicly available, acquired, first-party, synthetic, human feedback), publisher opt-out controls and excluded sources
- [2026 Responsible AI Transparency Report](#) — documentation and evaluation practices across the AI stack

Model-level: public summaries of training content

- MAI-Code-1 Flash [data summary card](#)
- MAI-Image-2 [model card](#)
- MAI-Image-1 [data summary card](#)
- MAI-Thinking-1 [data summary card](#)
- MAI-Image-2.5 [data summary card](#)
- MAI-Image-2.6 / 2.6-Flash [data summary card](#)
- MAI-Code-1.1-Flash [data summary card](#)

Model-level: model cards and data summary cards for open-weight models

- Phi-4-reasoning [model card](#) and [technical report](#)
- Phi-4 [model card](#)
- Phi-4 [data summary card](#)
- Phi-4-reasoning [data summary card](#)
- Phi-4-mini-instruct, Phi-4-mini-reasoning and Phi-4-mini-flash-reasoning [data summary card](#)
- MAI-DS-R1 [data summary card](#)

19. Does your organization demonstrate transparency related to advanced AI systems through any other methods?

Yes

If yes, describe additional method(s), including where relevant transparency measures related to AI agents

Microsoft publishes technology-specific documentation, such as application cards, that explain how AI technologies, including agents, operate beyond the capabilities of the underlying model. These artefacts may describe, as applicable, an application's or platform service's intended uses and limitations; its agentic capabilities; the models, knowledge sources, tools and connected services it uses; the data and resources it may access; the permissions and security boundaries under which it operates; the actions it may take; and the human-approval, monitoring and administrative controls that govern those actions. Examples include:

- **Microsoft Copilot Cowork Application Card:** <https://learn.microsoft.com/en-us/microsoft-365/copilot/responsible-ai/copilot-cowork-application-card>. This explains how Cowork carries out multi-step tasks across Microsoft 365 within the user's existing permissions, makes its actions visible in the conversation, and requires explicit approval for sensitive actions.
- **Microsoft Foundry Agent Service Transparency Note:** <https://learn.microsoft.com/en-us/azure/foundry/responsible-ai/agents/transparency-note>. This describes the service's capabilities and limitations; connections to models, knowledge sources and tools; expectations for human oversight; and considerations for consequential uses.
- **Microsoft Copilot Studio Application Card:** <https://learn.microsoft.com/en-us/microsoft-copilot-studio/system-service-card-copilot-studio>. This explains the platform's intended uses, capabilities and limitations and the choices developers and deployers make when configuring agents, tools and connectors.
- **Microsoft Security Copilot Agents Application Card:** <https://learn.microsoft.com/en-us/copilot/security/security-copilot-application-card-agents>. This describes how Security Copilot agents generate outputs and take actions according to customer-configured logic, permissions and triggers, including the boundaries placed on autonomous action.
- **Microsoft Fabric Real-Time Intelligence Operations Agent Transparency Note:** <https://learn.microsoft.com/en-us/fabric/real-time-intelligence/operations-agent-transparency-note>. This explains the agent's knowledge sources, tools, stored threads, playbooks, monitored entities, rules and recommended actions.
- **Microsoft Discovery Platform Card:** <https://learn.microsoft.com/en-us/azure/microsoft-discovery/concept-platform-card>. This describes Microsoft Discovery's multi-agent architecture, intended uses, capabilities and limitations, deployment modes and responsible-deployment considerations.

Microsoft also provides broader guidance for designing and governing agents responsibly:

- **Responsible AI for agent design:** <https://learn.microsoft.com/en-us/agents/design-guidelines/responsible-ai>
- **Responsible AI considerations for Microsoft 365-hosted agents:** <https://learn.microsoft.com/en-us/agents/architecture/responsible-ai>

These artefacts complement the 2026 Responsible AI Transparency Report, which also provides information about how Microsoft is governing systems with rapidly evolving agentic capabilities.

20. Does your organization publish statistical insights derived from anonymised, aggregated usage data to advance understanding of AI adoption and its economic impacts in areas including occupations and tasks by region, sector and country? To this end, do you contribute to international cooperation efforts?

Yes

If yes, please provide links or describe the scope of insights published, including where relevant contributions to understanding AI adoption in specific sectors (e.g. public administration, education systems).

Microsoft publishes statistical insights derived from aggregated, anonymised or de-identified data to advance understanding of AI adoption, patterns of use and economic impact across countries, regions, sectors, occupations and user contexts. Microsoft also contributes relevant data and research to international organisations, governments, researchers and public interest partners.

Global AI adoption. Microsoft's Global AI Diffusion Reports, developed by the Microsoft AI Economy Institute, use aggregated and anonymised Microsoft telemetry to estimate the share of the working age population using generative AI. The methodology adjusts for differences in operating system and device market share, internet penetration and national population, enabling comparisons across economies and over time. The Q2 2026 report covers 147 economies and examines the gap in adoption between the Global North and Global South, alongside differences in how early users apply AI for learning, schoolwork, content creation, shopping, advice and information retrieval.

- Global AI Diffusion Report, Q2 2026: <https://www.microsoft.com/en-us/corporate-responsibility/topics/ai-economy-institute/reports/global-ai-diffusion-report/>
- Downloadable data: <https://github.com/microsoft/ai-diffusion-report>
- Technical methodology: <https://doi.org/10.48550/arXiv.2511.02781>

Patterns of AI use across sectors and user contexts. Microsoft also publishes privacy-preserving research based on large-scale, de-identified Copilot usage data. These studies use taxonomies to analyse what people use AI for, how use varies by device and time, and how patterns differ across countries and demographic or institutional contexts.

- *Public use of a generalist LLM chatbot for health queries* (Nature Health, April 2026) examines more than 500,000 de-identified Copilot health conversations and provides evidence about how people use general-purpose AI for health information, advice and personal symptom assessment: <https://www.nature.com/articles/s44360-026-00117-x>
- *Global analysis of country-level factors associated with chatbot usage for health* (Nature Health, July 2026) examines approximately 1.7 million de-identified health-related conversations across 109 countries and relates differences in usage to country-level factors, including income, demographics and trust in health systems: <https://www.nature.com/articles/s44360-026-00174-2>
- *It's About Time: The Temporal and Modal Dynamics of Copilot Usage* (December 2025) analyses 37.5 million de-identified Copilot conversations from January to September 2025. It identifies differences in use by device, time of day and context, including the prominence of health-related use on mobile devices and work- and technology-related use on desktop devices during business hours: <https://arxiv.org/abs/2512.11879>

Occupational, sectoral and labour-market impacts. LinkedIn's Economic Graph Research Institute publishes analysis based on aggregated LinkedIn profile, skills and job-posting data to examine how AI is affecting employment, skills, occupations and sectors. Its AI labour-market research tracks measures such as AI talent, demand for AI skills, AI literacy, hiring, occupational change and entry-level employment across countries and sectors.

- AI Labour Market Update, August 2026, and other AI-at-work research: <https://economicgraph.linkedin.com/research/ai-skills-resources>
- LinkedIn Economic Graph publications: <https://economicgraph.linkedin.com/workforce-data/publications>
- Workforce data and methodology: <https://economicgraph.linkedin.com/>

LinkedIn also publishes sector-specific research on the workforce supporting AI infrastructure. A [Deep Dive into the Global Data Center Workforce](#) examines changes in the occupations and skills supporting the expansion of data-centre infrastructure. While this is not itself an analysis of AI-product usage, it provides complementary evidence on AI's wider employment and infrastructure impacts.

International cooperation and data partnerships. Through its [Data for Impact programme](#), LinkedIn shares anonymised, aggregated Economic Graph datasets with governments, multilateral organisations and selected public-benefit entities at no cost. The programme supports partners in producing research and informing policies on jobs, skills, migration, digitalisation and economic opportunity. LinkedIn is a founding data partner of the Development Data Partnership and works with organisations including the World Bank, IMF, OECD, IDB, UNDP, CAF and EBRD.

LinkedIn data on AI skills and employment has supported AI ecosystem and labour-market analysis by OECD.AI, Stanford HAI and Chile's National AI Center, as well as partner-led research with international organisations and public research institutions. Further information on the datasets, safeguards and partnership process is available at: <https://economicgraph.linkedin.com/data-for-impact>.

Please estimate the total staff time and annual expenses associated with conducting compliance activities and include any further comments and implementation documentation (optional).

No answer provided

Section 4 – Organizational Governance, Incident Management, and Transparency

21. How has AI risk management been embedded in your organization's governance framework?

- Board of Directors oversight
- Executive oversight
- Senior management oversight
- Governance council or committee
- Independent advisory body
- Other (e.g., distributed / multi-institutional governance arrangement) (please specify): research-informed advisory input from Microsoft Research and external subject matter experts, without decision rights

Please describe how AI risk management is integrated into governance structures, including: which roles or bodies have advisory versus decision-making authority; which roles or bodies have final accountability for AI risk-related decisions; and how often policies are reviewed or updated (e.g., following major incidents, regulatory changes, or material updates to systems).

Governance framework. Microsoft embeds AI risk management through a layered governance framework combining company-wide requirements, specialist review, executive decision-making and Board oversight. The Responsible AI Standard translates Microsoft's Responsible AI principles into requirements for teams developing and deploying AI models, platform services and applications. The Office of Responsible AI (ORA), led by the Chief Responsible AI Officer, manages this framework, defines roles and responsibilities, establishes governance processes and provides guidance on legal and regulatory requirements. Day-to-day implementation is supported by engineering partners and the Responsible AI Governance Community embedded across product groups.

Specialist and advisory review. Specialist review complements requirements. Under the Sensitive Uses program, ORA works with product engineering and customer and partner solutions teams where foreseeable use or misuse meets Sensitive Use criteria, with reviews resulting in additional requirements where appropriate. ORA also provides early guidance on emerging capabilities and risks not yet addressed by formal policy, while Microsoft Research and external subject-matter experts provide research-informed advisory input. For qualifying generative AI products and features, ORA, a responsible AI engineering team, and experts from the Responsible AI Governance Community facilitate centralized or expert-led review processes. These review mechanisms operate at scale: between July 2025 and June 2026, Microsoft conducted nearly 2,500 generative AI product and feature reviews and more than 450 Sensitive Use reviews.

Executive oversight and decision-making. At executive level, the Responsible AI Council, co-led by Microsoft's Vice Chair and President Brad Smith and the Chief Technology Officer and Executive Vice President of AI Kevin Scott, meets quarterly and oversees company-wide progress. It includes Corporate Vice Presidents from across the business, supported by division-level leaders responsible for operationalizing the responsible AI program. At Board level, Microsoft's responsible AI leaders engage regularly with the Board of Directors through internal enterprise risk management reviews as well as engagement with the Environmental, Social and Public Policy Committee. Under the [Frontier Governance Framework](#), for frontier models, the Executive Officers responsible for Microsoft's AI governance program, or their delegates, make the final decision on whether to approve the recommendation for secure and trustworthy deployment. Along with other responsible AI policies and programs, the Frontier Governance Framework is subject to internal audit and Board oversight.

Review and adaptation. Policies and requirements are reviewed and updated on an ongoing basis in response to technological developments, regulatory changes, identified risks and incidents, lessons from after-action reviews and implementation feedback. The Responsible AI Standard has been progressively updated as AI capabilities, risks and regulatory expectations have evolved. Most recently, it was re-engineered to include dedicated Developer and Deployer Chapters and to distinguish core requirements from scenario-specific requirements. As part of this work, Microsoft analysed more than 100 enacted and proposed laws and refreshed its process for continuous regulatory analysis and adaptation. The Frontier Governance Framework also requires an explicit review at least every twelve months. Changes to practices are reviewed by the Chief Responsible AI Officer before adoption, and material revisions are published within 30 days of adoption.

22. Are relevant staff trained on your organization's governance policies and risk management practices?

Yes

If yes, describe training methods (e.g., onboarding modules, regular sessions, and confidential channels to report concerns).

Microsoft combines company-wide responsible AI training with role-based and technical learning, targeted training for relevant internal communities, and confidential channels for employees to raise concerns.

Company-wide training. Microsoft employees are required to complete the Trust Code, Microsoft's standards of business conduct training, which applies to the broadest cross-company audience. Responsible AI is one of the most prominent topics in the latest edition, which 99 percent of employees have completed.

Role-based and technical learning. This company-wide baseline is complemented by role-based and technical training. In 2025, Microsoft offered over 500 AI-related educational resources for employees, spanning live sessions, recorded training and written reference materials. In 2026, the internal responsible AI learning programme was updated to introduce role-based learning paths, providing consistent, centralized education tailored to different roles and responsibilities across engineering, sales, policy, legal and other internal audiences. Since January 2026, Microsoft's Machine Learning, AI, and Data Science programme (MLADS+) has delivered more than 45 learning sessions related to responsible AI, reaching over 4,000 attendees. The internal responsible AI community also receives targeted training and community-based learning, including over the past year on prompt injection defenses and threat modelling for agentic systems. New agentic AI threat modelling techniques, integrated into the Security Development Lifecycle and the Responsible AI Standard, were scaled through internal deep-dive workshops that drew more than 3,000 engineers.

Reporting concerns. Employees have multiple avenues to raise concerns, including an anonymous reporting channel. Over the past year, Microsoft updated its whistleblower policy and process, including to address AI concerns, with the goal of supporting employee confidence in reporting AI-related issues.

23. Are steps taken to address reported incidents documented and maintained internally?

Yes

If yes, describe how incident records are kept and used for follow-up (e.g., incident logs, root-cause analysis reports, internal tracking systems). You may indicate whether and how you define escalation thresholds for when incidents must be surfaced to senior leadership or external parties.

Issue records and classification. Issues reported or discovered, including through the Microsoft Security Response Center (MSRC), are classified against [Microsoft's Vulnerability Severity and Content Classifications for AI Systems](#), the AI Security and Safety Bug Bar, which defines categories and severity levels for common AI security vulnerabilities and content-related issues. Each MSRC case is assigned a unique case number, and records are retained indefinitely. The Bug Bar is correlated with Microsoft's central AI risk taxonomy, so that assessment, mitigation, detection, and response use a common vocabulary before and after release.

Externally, vulnerabilities at any severity level are surfaced publicly as CVEs (“Common Vulnerabilities and Exposures”) whenever customer action is required to patch or update software. Since [2024](#), Microsoft also issues CVEs for cloud service vulnerabilities assessed as Critical severity even where no customer action is required (and marks them as such in the Security Update Guide and in the CVE record). Where identified risks or vulnerabilities affect other organizations, information is shared with them consistent with Coordinated Vulnerability Disclosure principles.

Response and follow-up. In 2025, Microsoft formalized incident response playbooks across 13 topics, supporting MSRC, product, and legal teams in following a unified process when AI incidents arise; the playbooks are reinforced through tabletop exercises and readiness drills. At the conclusion of a response to an issue or incident, a review is typically conducted to understand the nature of the issue or incident, ensure the full scope of impact has been assessed, and track items to remediation. Reviews are presented to key executive leaders; themes are synthesized across reported issues and used to educate product teams, update engineering processes, and update policies where applicable; and findings feed back into product safeguards, including updates to guardrail tuning.

24. Does your organization share research and best practices on addressing or managing risk?

Yes

If yes, describe how research or good practices are shared (e.g., publications, conferences, collaborations with other sectors).

Public research and publications. Microsoft Research and subject matter experts from engineering and security teams publish their AI safety and security research, including work on evaluation-aware behavior in reasoning models, techniques for detecting backdoored models at scale, AI recommendation poisoning, and research on how generative AI use affects critical thinking in knowledge work. The Microsoft Research Sociotechnical Alignment Center has published a social-science-based measurement framework for generative AI evaluation and identified validity gaps in commonly used red teaming and LLM-as-judge metrics. Microsoft's MAI Futures team brings together multidisciplinary expertise to study the societal impacts of AI, publishing peer-reviewed research on both near- and long-term effects, drawing on real-world AI usage data.

Open-source releases and shared datasets. Microsoft contributes safety and security tooling to the open ecosystem so that others can apply and inspect it, including PyRIT for AI red teaming, RAMPART and Clarity for encoding adversarial and benign scenarios as repeatable tests, ASSERT for turning written behavior requirements into executable evaluations, TaskTracker for detecting indirect prompt injection developed with academic partners, and the Agent Control Specification for applying runtime governance across the agent lifecycle. Following the [Adaptive Prompt Injection Challenge](#), more than 200,000 attack prompts submitted by over 800 participants were compiled into a public dataset to support continued research on prompt injection defenses.

Industry and multistakeholder fora. Through the [Frontier Model Forum](#), Microsoft works with peers to develop and share practices for evaluating and mitigating risks related to biosecurity, cybersecurity, agentic risks and harmful AI system behavior; members have used an information-sharing agreement announced in 2025 to advance shared knowledge relevant to risk mitigation, and Microsoft co-funds the Forum's wet-lab biosafety research grants. Microsoft has also served as an industry reviewer for the 2025 and 2026 International AI Safety Reports, providing feedback to the independent scientific

assessment of general-purpose AI capabilities and risks. Microsoft contributed to the Partnership on AI report [Prioritizing Real-Time Failure Detection in AI Agents](#), published in September 2025. With MLCommons, Microsoft co-chairs the working group expanding the AILuminate benchmark across languages and cultures, supports the Jailbreak benchmark and the Psychosocial Task Force, and contributes to the development of the AILuminate [Global Assurance Program](#). Microsoft is a founding member of the [Appia Foundation](#), which works on AI conformity assessment across the AI supply chain and across sectors, and is a co-founder of the Coalition for Content Provenance and Authenticity.

Government research partnerships. Microsoft has research partnerships with government research institutions in the US, UK, Australia and Singapore. In the US and UK, Microsoft has agreements with the US Center for AI Standards and Innovation (CAISI) and the UK AI Security Institute (AIS) to advance AI testing and evaluation, including collaborative work to test Microsoft's frontier models, assess safeguards, and help mitigate national security and large-scale public safety risks. Our collaboration with the Australian Department of Industry, Science and Resources, coordinating through the Australian AI Safety Institute, focuses on research on human–AI interaction risks, including overreliance, emotional dependency, and user behavior in AI companion chatbots and conversational AI systems, especially for young people and vulnerable groups.

With Singapore's Infocomm Media Development Authority (IMDA), working closely with the Singapore AI Safety Institute, we collaborate on the co-development of a policy framework on trusted access to frontier capabilities for cybersecurity operations and safety and security testing, as well as multilingual safety evaluation across ASEAN languages. Through these bilateral partnerships, Microsoft seeks to support more robust and reproducible evaluation methods, greater technical exchange, and stronger shared approaches to assessing emerging frontier risks.

Public transparency reporting. Microsoft publishes an annual [Responsible AI Transparency Report](#), now in its third edition, describing how Microsoft governs AI, what it has learned and where challenges remain. The 2026 edition includes case studies tracing how individual models, platform services, and applications were assessed before release, an account of how Microsoft governs its own use of AI as a deployer, and a breakdown by impact category of AI safety and security cases reported to the Microsoft Security Response Center. Microsoft first published its [Frontier Governance Framework](#) in February 2025 and updated it in February 2026, adding disclosures on its red teaming and risk modelling practices and providing greater transparency into how Microsoft evaluates and monitors frontier model risks and engages third-party evaluators. Microsoft also publishes data summaries, model cards, and application and platform cards for in-scope technologies.

25. Does your organization use international technical standards or best practices for AI risk management and governance policies?

Yes

If yes, please indicate which standards or frameworks apply:

- ISO/IEC standards (e.g., ISO/IEC 27001, ISO/IEC 42001)
- NIST AI Risk Management Framework (AI RMF)
- OECD AI Principles
- Industry-led best practices or codes of conduct

- Other international or regional standards (please specify): EU General-Purpose AI Code of Practice; C2PA Content Credentials; OpenTelemetry; Agent Control Specification, EU Code of Practice on Transparency of AI-generated Content (covering the Article 50(2) machine-readable marking requirements)

If yes, describe which international technical standards or best practices your organization uses.

ISO/IEC. Microsoft was a key contributor to the development of ISO/IEC 42001:2023, the first global management system standard for AI, and has since been independently certified against it across a broad set of AI systems: Microsoft 365 Copilot, Microsoft 365 Copilot Chat, Microsoft Foundry, Microsoft Security Copilot, Microsoft Copilot Studio, GitHub Copilot, Dragon Copilot, Dragon Copilot (Radiologist), and Copilot Health. These span productivity, developer tools, healthcare, security, and platform services under a single converged audit framework operationalized over the last year. Microsoft online services are also certified to ISO/IEC 27001 for information security management through regular independent third-party audit. Microsoft has also been instrumental in the development of ISO/IEC 23894 on AI risk management and ISO/IEC 42005 on AI system impact assessments, building on Microsoft's internal AI system impact assessment requirements. Microsoft has also contributed to work on conformity assessment of AI systems, including ISO/IEC 42007. Microsoft participates in the standards development organizations responsible for this work, including ISO/IEC JTC1 SC42 Artificial Intelligence and CEN/CENELEC JTC21 Artificial Intelligence.

Technical standards initiatives. For agentic systems, Microsoft has contributed an agentic extension of the OpenTelemetry observability standard, and announced the Agent Control Specification as an open, vendor-neutral specification for runtime governance. Microsoft also supports the Agentic AI Foundation.

NIST. Microsoft actively participates in the standardization efforts led by the U.S. National Institute of Standards and Technology. Microsoft's responsible AI program aligns with the AI Risk Management Framework and its govern, map, measure, and manage functions.

OECD. Microsoft aligns with the OECD AI Principles and uses the OECD.AI Catalogue of Tools and Metrics for Trustworthy AI as a reference point. Microsoft contributed a voluntary report under version 1.0 of the Hiroshima AI Process reporting framework in 2025 and contributed to the development of version 2.0.

Industry-led best practices and codes of conduct. Microsoft participated in multistakeholder efforts to develop the EU General-Purpose AI Code of Practice and has published a data summary and model card for MAI-Image-1 as required under the Transparency Chapter. Microsoft is also a signatory to the EU Code of Practice on Transparency of AI-generated Content, covering the Article 50(2) machine-readable marking requirements. Microsoft applies Frontier Model Forum consensus best practices (see 24.A), and implements C2PA Content Credentials for content provenance.

Any further comments and implementation documentation.

No answer provided

Section 5 – Content Authentication & Provenance Mechanisms

26. What mechanisms, if any, does your organization put in place to allow end-users, where possible and appropriate, to know when they are interacting with an advanced AI system developed by your organization?

- User disclosure (e.g., text notice, verbal statement, or system label)
- Persistent visual or audio indicators (e.g., watermark, icon, chime, banner) during AI interaction
- Distinct branding or labeling for AI-generated content (e.g., “AI-generated” tag, metadata label)
- Documentation or FAQs explaining when and how AI is used in the system
- Access to provenance information (e.g., model version, content authenticity signals)
- Disclosure to end-users when an AI system is acting autonomously or executing actions on their behalf (e.g., making purchases, sending communications, modifying files)

Provide any additional details or context on how or when users are informed they are interacting with advanced AI systems (e.g. if such mechanisms are planned, in testing, or applied only to certain products).

Microsoft employs several complementary mechanisms to ensure that end-users understand when they are interacting with an advanced AI system. These mechanisms are tailored to different products and scenarios.

User disclosures within the product interface. For applicable products, Microsoft's Responsible AI Standard requires product teams to disclose the role of AI in user interactions and to inform users that they are engaging with an AI system. In practice, these disclosures can include on-screen notices, contextual warnings, and consent flows embedded in the user experience. AI-generated responses may also include citations to source material so that users can verify information.

For example, across Copilot experiences, in-product disclosures inform users that they are interacting with an AI application. These disclosures may include notices or disclaimers about the potential for AI to make errors or produce unexpected content.

In certain risk scenarios, additional disclosures may apply. Microsoft's Sensitive Interactions Policy, which addresses open-ended conversation, memory, and personalization across sessions, requires disclaimers and factual responses, including encouragement to consult qualified professionals, when interactions relate to significant decisions such as health or finance. For AI systems acting on a user's behalf, Microsoft also surfaces actions directly in the user experience.

For example, in Copilot Cowork, actions are visible in the conversation, actions that send or modify content require explicit user approval before execution unless the user has already provided permission in their initial instructions, and outgoing messages include an attribution indicating AI assistance, so that recipients are informed when interacting with AI-produced content. Similarly, Browse with Copilot is an opt-in experience with an educational flow, preview labeling, and in-product warnings that highlight AI involvement; users can see actions in real time and interrupt or take control.

Distinct labeling of AI-generated content. The Responsible AI Standard requires that AI-generated or AI-modified content be clearly and conspicuously disclosed where appropriate. Microsoft is scaling engineering capabilities to mark AI-generated content and attach provenance metadata across relevant product surfaces. Provenance metadata is implemented to meet C2PA's requirements for cryptographically hardened marking. In July 2026, Microsoft signed the EU Code of Practice on

Transparency of AI-generated Content covering the Article 50(2) machine-readable marking requirements (see 27.D).

Documentation explaining when and how AI is used. Microsoft publishes transparency documentation to explain how AI is integrated into its products and services. This includes data summaries and model cards for models such as MAI-Image-1, MAI-Image-2, and Phi-4-Reasoning, as well as application and platform cards. Microsoft launched new templates for AI model, platform, and application cards and began implementing them in 2026. Since January 2026, Microsoft has published 20 application and platform cards describing technology behavior, capabilities, limitations, safety components and governance practices, and clarifying the respective roles of Microsoft and customers in responsible deployment, including an application card for Copilot Cowork describing its capabilities, intended uses, limitations, and best practices.

Access to provenance information. LinkedIn provides access to provenance signals about AI-generated content, enabling downstream users to inspect whether content was AI-generated using tools that support the C2PA standard (i.e., LinkedIn displays a Content Credentials indicator on signed images, see also 27.D). Under the Deployer Chapter of the Responsible AI Standard, teams deploying third-party AI applications assess how applications that Microsoft is procuring mark or apply content provenance to AI-generated or modified content.

27. Does your organization use content provenance detection, labeling or watermarking mechanisms that enable users to identify content generated by advanced AI systems?

Yes

Content Credentials

27.A. If yes, how?

- Latent disclosure (e.g., hidden provenance information)
- Manifest disclosure (e.g., visible label)
- Detection tool(s) for hidden provenance signals (e.g., watermark or metadata checks)
- Use of international or industry standards (e.g., C2PA)

27.B. If yes, does your organization use international technical standards or best practices when developing or implementing content provenance?

Yes

27.C. Does your organization collaborate on or invest in research to advance content authentication and provenance?

Yes

Describe how provenance mechanisms are implemented and which standards or tools are used, and any collaborations or investments to improve content authentication (e.g., C2PA credentials, joint R&D projects, open-source initiatives, standard-setting groups). You may also indicate whether these mechanisms are tested for robustness against removal, tampering, or circumvention.

Implementation. Microsoft co-founded the Coalition for Content Provenance and Authenticity in 2021 and has held leadership roles since its founding. AI-generated images from consumer tools such as

Microsoft Designer include C2PA Content Credentials, and LinkedIn displays a Content Credentials indicator on signed images. Microsoft's layered strategy pairs cryptographically sealed C2PA provenance metadata with watermarking as complementary mechanisms; invisible watermarks are identifiable by specialized tools, resilient to common modifications such as resizing or cropping, and can help recover provenance information where a C2PA manifest has been removed. Over the past year, Microsoft scaled its readiness to apply both mechanisms across relevant product surfaces, with central solutions to support latency stabilization, production hardening, and monitoring at scale. Content Credentials have also proved useful in enforcement: in a Microsoft Digital Crimes Unit investigation into misuse of AI image-generation systems, provenance metadata helped investigators establish the origin of many of the AI-generated images discovered.

Standards and collaboration. Over the past year, C2PA released Content Credentials 2.3 and committed to supporting adoption across additional languages; with C2PA partners, Microsoft is working to extend provenance standards beyond an English-ready baseline. Microsoft signed the Tech Accord to Combat Deceptive Use of AI in 2024 Elections and piloted Content Integrity Tools, including a capture app developed with TruePic, enabling campaigns, election authorities, and media organizations to attach Content Credentials to authentic content.

Regulatory alignment. These mechanisms underpin Microsoft's compliance with emerging transparency requirements, including California's AI Transparency Act and other U.S. state laws as well as Article 50(2) of the EU AI Act and the EU Code of Practice on Transparency of AI-generated Content. The Code identifies digitally signed metadata and imperceptible watermarking as its central measures for providers, consistent with Microsoft's layered approach.

Robustness testing. Provenance information is only useful if it cannot easily be stripped off a file or falsified. As co-founder and co-chair of C2PA, Microsoft helps develop the Content Credentials specification. Under the specification, provenance information is digitally signed together with the file, so that any later alteration to the file or to the information attached to it breaks the signature and can be detected.

For content that carries no credential, Microsoft's AI for Good Lab, with Northwestern University and the human rights organization WITNESS, published a free test set of more than 50,000 real, AI-generated, and manipulated images, videos, and audio clips, including examples journalists and human rights defenders have encountered and examples built to fool detectors. It allows anyone to measure how a deepfake detector performs on realistic material rather than on easy test cases.

Microsoft Research's February 2026 report, [Media Integrity and Authentication: Status, Directions, and Futures](#), tested signed metadata, watermarks, and fingerprinting across the full life of a file, from capture and editing to sharing and verification, and against attempts to defeat them, including attacks designed to make authentic content look fake. It concluded that only signed metadata can reliably show whether a file is unchanged and who signed it, and that watermarks are best used as a backup for recovering that information when the metadata has been stripped, a known limit of any metadata-based approach: provenance data can be removed through re-encoding, screenshotting, or platform processing. Building on those findings, Microsoft tests in practice whether watermarks survive common edits such as resizing and cropping, whether provenance metadata can still be verified after a file has been shared, how often detectors are wrong in either direction and how they hold up against attacks, and how people understand the signals they see. In one study, an attacker's success rate against a

Microsoft AI detector fell from 41 percent to 2 percent once the detector had been trained on adversarial examples.

Any further comments and implementation documentation.

No answer provided

Section 6 – Research & Investment to Advance AI Safety & Mitigate Societal Risks

28. Does your organization advance research and investment in the following areas?

- Security
- Safety
- Data quality and representativeness
- Information integrity
- System performance and outcomes (e.g. accuracy, consistency)
- Explainability or interpretability
- Transparency
- Robustness or reliability
- Other (please specify): psychosocial risk and human-AI interaction, including AI companionship and appropriate reliance

If yes, please describe key research or investment initiatives in the areas selected above (e.g., dedicated safety teams, funding of research programs, contributions to open-source safety tools, publishing research, partnerships and collaborative research initiatives). You may also indicate whether related outputs are publicly shared, shared with partners or within consortia or not publicly disclosed.

Dedicated teams and programs. Multiple organizations contribute to Microsoft's investments in applied and basic research, including the Office of Responsible AI and its Sensitive Uses and Emerging Technologies program; the AI Red Team; Microsoft Research, including the Sociotechnical Alignment Center; the AI for Good Lab; and the AI Economy Institute. The Emerging Technologies program, launched in 2025, proactively examines frontier capabilities such as memory and computer use agents and identifies emerging risks not yet addressed by formal policy, with applied research that ultimately informs core or scenario requirements in the Responsible AI Standard.

Security and safety. Microsoft continued to invest in dedicated security, research, safety, and evaluation expertise for emerging AI risks. Microsoft's AI Red Team conducts human-led and tool-based adversarial testing for novel, high-impact technologies and risk scenarios; in 2025, its testing grew 1.5-fold, executing more than 100 operations, while developing expertise on emerging attack vectors including adversarial memory poisoning and backdoors in open-weight models. Microsoft also expanded centralized evaluation capabilities, including greater coverage of agentic systems, and invested in specialized expertise across frontier risk domains such as advanced cyber capabilities, biological and chemical risks, advanced autonomy and loss-of-control scenarios, and harmful manipulation.

Research, evaluation, and governance. In September 2026, Microsoft AI published the first draft of its [Humanist AI Code of Conduct](#) for public consultation. The Code is intended to become the primary governing document for MAI models, translating Microsoft AI's broader commitment to human-centered AI into concrete expectations for model behaviour, values, and boundaries and, over time, informing how MAI models are trained, evaluated, operated, and governed. It sets out a forward-looking framework focused on enabling useful and capable models while preserving meaningful human control, establishing clear lines of authorization and accountability, and defining firm safety boundaries for frontier risks. The consultation invites external scrutiny of how these commitments can be made sufficiently precise, operationalized, and evaluated, including in multi-agent settings. Microsoft AI will publish a summary of the feedback received and resulting changes alongside a revised Code, with further consultations planned as the framework evolves.

Robustness, reliability, and system performance. Microsoft has invested in the science of evaluation. An internal evaluation framework translates risk definitions into measurement techniques, automating risk systematization, targeted query generation, annotation, and metrics calculation with policy and domain experts in the loop. This work is grounded in the Sociotechnical Alignment Center's social science measurement framework, which defines how an issue is measured and whether those measurements are valid.

Interpretability-related safety research. Microsoft Research has developed techniques that inspect model internals to detect unsafe behavior, including TaskTracker, which detects indirect prompt injection by examining an LLM's internal activation states. Microsoft researchers have also explored techniques to detect backdoored language models and developed a scanner to identify potential triggers at scale.

Data quality and representativeness. Through [Project Gecko](#), Microsoft co-designs AI technologies with local communities in East Africa and South Asia, including the Paza family of automatic speech recognition models operating on mobile devices across six Kenyan languages and a Multimodal Critical Thinking Agent that reasons over community-generated video, voice, and text. Microsoft 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](#). Samiksha, developed with Karya and The Collective Intelligence Project in India, advances community-centered evaluation of AI behavior in real-world contexts.

Information integrity. Microsoft's AI for Good Lab, with Northwestern University and WITNESS, published a benchmark of more than 50,000 real, AI-generated, and manipulated media samples for evaluating deepfake detection under real-world conditions, and Microsoft Research's February 2026 report on media integrity and authentication evaluated provenance, watermarking, and fingerprinting against technical and sociotechnical attacks (see 27.D).

Transparency. Microsoft refreshed its data summary and model card templates to incorporate learnings and address regulatory requirements, and scaled documentation across the technology stack with new templates for AI applications and platform services, publishing 20 application and platform cards since January 2026. These templates were shaped by partner research, including work Microsoft supported at UC Berkeley's Artificial Intelligence Research Lab, which mapped existing transparency tools and surveyed how AI builders, end users, and policymakers experience transparency in practice.

Psychosocial risk and human-AI interaction. Microsoft pairs product-level interventions with research and direct learning from youth. Microsoft Research is examining what design choices promote human agency and appropriate reliance, including research on [the role of explanations, sources, and](#)

[inconsistencies in shaping users' reliance on LLM outputs](#). Microsoft is also learning directly from teens through its AI Futures Youth Council and workshops with Cyberlite to co-design guidance on AI companions. Microsoft participates in the MLCommons AILuminate Psychosocial Task Force, which is examining how psychosocial risks are conceptualized and how the impact of mitigations can be measured, with the aim of developing shared benchmarks across areas including crisis prevention and unqualified AI advice.

Funding of external research. Microsoft supports independent research through the [Agentic AI Research and Innovation program](#), supporting twelve projects on topics including reliability, uncertainty handling, and cultural misalignment in AI agents; the [Computing Research Association's Trustworthy AI Research Fellowship](#) for early-career scholars; and the [US National AI Research Resource pilot](#), including a Deep Partnership track giving researchers access to Microsoft Discovery. In July 2026, Microsoft's AI Red Team extended its capability through the [External Red Team Alliance \(EXTRA\)](#), a formalized global extension of the AI Red Team that funds independent AI safety and security research and builds a distributed network of external specialists for red teaming in specialized domains.

Disclosure of outputs. Research findings are generally released publicly, through peer-reviewed publication, Microsoft Research outputs, open-source repositories, and public datasets. A narrower set is shared within consortia, including under the Frontier Model Forum information sharing agreement, or bilaterally with government partners. The channels used are described at Question 24.

29. Does your organization pursue research or investment related to socio-economic or environmental considerations in the development and deployment of AI?

- Socio-economic considerations (e.g., AI adoption and productivity, workforce transitions, skills development, access to AI tools)
- Environmental considerations (e.g., energy use, computational efficiency, resource use)

If yes, please provide examples of relevant research or investment initiatives for the areas selected above.

Socio-economic considerations.

Research and labour markets. In 2025, Microsoft launched the [AI Economy Institute](#) to study how AI can diffuse responsibly across education systems and labor markets; in fall 2025 it launched a [research cohort](#) focused on education in the AI economy. Microsoft also publishes labour market research through the [LinkedIn Economic Graph Research Institute](#), tracking AI skills demand, hiring patterns, occupational exposure to AI and entry-level hiring across economies (see 20.A).

Skills development and worker voice. In July 2025, Microsoft launched [Microsoft Elevate](#), which aims to reach 20 million people by 2030 with upskilling opportunities ranging from foundational AI fluency to advanced technical training (see Question 30.A). Microsoft also works with organised labour to ensure worker perspectives inform AI adoption. This includes a [partnership with the AFL-CIO](#) in the United States and, since January 2026, a [framework agreement with the Australian Council of Trade Unions](#) (ACTU) focused on information-sharing, worker input into technology development, and public policy skills. Microsoft and the ACTU also convened a [Workers' Summit](#) in Sydney in April 2026.

Access across languages. Microsoft is investing 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. Microsoft's AI for Good Lab also

developed a [reproducible pipeline](#) for adapting open-weight large language models to low-resource languages, demonstrating measurable gains for languages including Chichewa, Inuktitut, and Māori.

Environmental considerations.

Community-first AI infrastructure. In January 2026, [Microsoft announced a community-first approach to building and operating AI datacentres](#) in the United States. This includes commitments to pay the full cost of electricity and associated grid upgrades so datacentre demand is not passed on to residents, minimise water use, replenish more water than is consumed within the same watershed, pay full local property taxes without abatements, invest in workforce development, and expand community investment in host regions.

Water efficiency. In Quincy, Washington, Microsoft partnered with the city on a first-of-its-kind cooling-water reuse facility that achieved an average [97% reduction in potable water consumption](#).

Research supporting sustainability. Microsoft supports the US National AI Research Resource (NAIRR) Pilot by providing access to Azure infrastructure, AI models and technical expertise. Through a Deep Partnership track launched with NAIRR, Microsoft supports “grand challenge” scientific research projects spanning fields including materials science and sustainability.

Any further comments and implementation documentation.

No answer provided

Section 7 – Advancing Human and Global Interests

30. Does your organization support any digital literacy, education or training initiatives to improve user awareness and/or help people understand the nature, capabilities and limitations of advanced AI systems?

Yes

If yes, describe any digital literacy or training initiatives your organization supports (e.g., user education, public AI-education programs, school or university partnerships, community workshops). You may also indicate whether these efforts include or prioritise underserved or underrepresented communities.

Microsoft invests in AI literacy across age groups and communities, with dedicated initiatives for young people, older adults, journalists, civil society, and underserved communities.

AI literacy and skills at scale. In July 2025, Microsoft launched Microsoft Elevate, working across Microsoft, LinkedIn, and GitHub and with partners, to reach 20 million people by 2030 with upskilling opportunities ranging from foundational AI fluency to advanced technical training. Public resources include the [Microsoft AI Skills Initiative](#) and the [Microsoft Learn AI skilling catalog](#), and Microsoft has worked with UNESCO's AI Business Council to build a repository of AI skilling resources. Over the past year, Microsoft teams delivered responsible AI training and knowledge-sharing programs reaching more than 20,000 engineers, policymakers, and customers, including AI Red Teaming 101, a free, self-paced training series for security professionals and developers on evaluating generative AI systems and addressing security risks.

Education and young people. Through Minecraft Education, Microsoft offers Reed Smart: AI Detective, teaching how AI works and helping students develop critical-thinking skills for evaluating AI-generated content, and AI Adventurers for younger learners. Microsoft created an AI classroom toolkit for educators covering fabricated content, privacy, gaps in linguistic or cultural context, and mental well-being, and partnered with Childnet on materials addressing AI-enabled intimate image abuse. In 2026, Microsoft co-developed a "by teens, for teens" guide to AI companions with Cyberlite, working with 131 students aged 12 to 18 across India and Singapore, and launched its first AI Futures Youth Council bringing together teens aged 13 to 17 from the US and Europe.

Underserved and underrepresented communities. In April 2026, Microsoft announced a partnership with Ampere to advance rural, remote, and Indigenous AI fluency in Canada, with objectives including embedding AI skilling into regional education systems, strengthening workforce readiness, and equipping more than 650 youth and 2,000 community members with long-term digital literacy skills. In Australia, Microsoft partnered with Deadly Coders, an Indigenous-owned nonprofit, to deliver culturally relevant AI and Minecraft Education sessions to 5,000 Aboriginal and Torres Strait Islander students in 2025. These initiatives form part of Microsoft's broader efforts to expand access to AI literacy among underserved and underrepresented communities.

Older adults. With AARP and Older Adults Technology Services (OATS), Microsoft supports OATS's AI Guide for Older Adults, which Microsoft amplified through in-kind advertising generating more than 1 billion impressions since May 2025; a scam and fraud hub with videos and interactive learning resources that reached more than 3,000 participants through virtual and in-person training in its first month; and AARP's fraud helpline awareness campaign, which generated more than 10 million impressions during the 2025 holiday season.

News and information literacy. Microsoft also works with organizations supporting journalists, creators, election officials, and civil society. This includes 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, as well as a pilot News Creator Training program for journalists and digital creators.

31. Does your organization pursue AI projects that aim to deliver societal or public benefits, including through partnerships with civil society and community groups?

Yes

If yes, please provide examples, including any collaborations with civil society, community groups, or efforts to address broader societal challenges.

AI for Good and societal challenges. The Microsoft AI for Good Lab focuses on areas where AI can make a tangible difference, including healthcare, education, and inequality. Collaborations include a partnership with Planet and the Institute for Health Metrics and Evaluation using satellite imagery and AI to produce high-resolution population maps for climate-event resilience; a machine learning study with Catholic Relief Services predicting food insecurity in Malawi; and work with SEEDS developing AI models for early warnings of extreme heat risks.

Community co-design and language inclusion. Over the past year, Microsoft has expanded community co-design of AI. Through Project Gecko, Microsoft co-designs AI technologies with local communities in East Africa and South Asia to support agriculture, including speech recognition models running on mobile devices across six Kenyan languages. Microsoft also invests in data and model

capability for African languages through LINGUA Africa, a collaboration involving the Masakhane African Languages Hub, Microsoft's AI for Good Lab, and the Gates Foundation.

Civil society partnerships on safety and wellbeing. Microsoft supports the Paris Peace Forum's International Research-driven Alliance for AI Serving Every Child (iRAISE), contributing to the publication [Adolescents & Anthropomorphic AI: Rethinking Design for Wellbeing](#). Microsoft also partnered with online safety education organization Cyberlite on [AI Companionship Youth Co-Design Sessions](#) involving 131 students aged 12 to 18 across India and Singapore, which informed a "by teens, for teens" set of resources on AI companionship, and launched its first [AI Futures Youth Council](#) bringing together teens aged 13 to 17 from the US and Europe to share their ideas and experiences and help Microsoft better understand young people's expectations of AI and responsible AI practices.

Fraud and scam prevention. Microsoft is a signatory to the [Industry Accord Against Online Scams and Fraud](#), a cross-industry effort to help protect people from scams. In May 2025, Microsoft's 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](#).

Responsible AI and global perspectives. 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. 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.

Public-interest commitments. Microsoft's public-interest work also reflects voluntary commitments made at the 2026 AI Impact Summit in New Delhi. Under Commitment 1, to publish statistical insights derived from anonymized, aggregated usage data to advance understanding of global AI adoption, Microsoft publishes the AI Diffusion Report and peer-reviewed analyses of real-world Copilot usage (see 20.A). Under Commitment 2, to evaluate multilingual capabilities and collaborate with local ecosystems for under-represented languages and cultural contexts, Microsoft co-chairs the MLCommons working group expanding the AILuminate benchmark across languages and cultures, contributed to the Frontier Model Forum's July 2026 issue brief on multilingual evaluations for CBRN and advanced cyber risks, and advances community-centered evaluation through Samiksha and Project Gecko (see 28.A).

Any further comments and implementation documentation.

No answer provided

