Report

Organization: **Microsoft**

http://www.microsoft.com/

Publication date: Apr 22, 2025, 09:01 AM PDT

Reporting period: 2025

Section 1 - Risk identification and evaluation

a. How does your organization define and/or classify different types of risks related to AI, such as unreasonable risks?

As context for this question and our broader approach to contributing to the efforts of the G7 and OECD on the voluntary reporting framework:

- Microsoft is committed to helping advance shared expectations for transparency among developers and deployers of advanced AI systems, including by fulfilling our voluntary commitments to submit this Report. Through our experience with various frameworks, we have identified consistent themes and significant overlap in core expectations, such as risk assessment, risk mitigation, incident reporting, and governance. This Report helpfully enables us to provide an overview of our practices organized by key focus areas.
- Microsoft is committed to helping advance shared expectations for transparency among developers and deployers of advanced AI systems, including by fulfilling our voluntary commitments to submit this Report. Through our experience with various frameworks, we have identified consistent themes and significant overlap in core expectations, such as risk assessment, risk mitigation, incident reporting, and governance. This Report helpfully enables us to provide an overview of our practices organized by key focus areas.
- We're at an inflection point for the adoption of AI. As AI contributes to opportunities for economic growth and scientific advancement around the world, we're seeing new regulatory efforts and laws emerge in parallel. Making the most of AI's potential will require broad adoption—enabled by putting in place the infrastructure and skilling programs necessary for workers and companies to flourish as well as advancing the trust that underpins people and organizations actually using new technology to enhance their lives and serve their goals. That's where AI governance has a critical role.
- We recognize that good AI governance will result from an iterative process, and as AI technology continues to develop and evolve rapidly, we remain committed to building tools, processes, and practices that allow us to adapt AI governance at the speed of AI innovation. We also remain committed to advancing our understanding of AI risks and effective mitigations. We invite feedback from the AI ecosystem and policymakers to help inform our future efforts. We look forward to continuing to engage in dialogue related to advancing trustworthy AI, and to continuing to share our learnings with stakeholders and the broader public as our program grows and evolves.

Steps to define and/or classify risks are part of the comprehensive AI governance program Microsoft has put in place to map, measure, and manage risks. We have developed and regularly evolve a risk taxonomy to apply as applicable across technology scenarios and leverage for governance.

Under our comprehensive program, AI models and systems are subject to relevant evaluation, with mitigations then applied to bring overall risk to an appropriate level. Microsoft's Responsible AI Standard and integrated set of policies and tools detail requirements for that process across technology scenarios. In some cases, we have also made available further details regarding policies for particular scenarios.

In the context of the development and deployment of highly capable AI models, within our Frontier Governance Framework, we have defined risks that warrant additional governance steps. In particular, we have defined "tracked high-risk capabilities," which include:

- Chemical, biological, radiological, and nuclear (CBRN) weapons. A model's ability to provide significant capability uplift to an actor seeking to develop and deploy a chemical, biological, radiological, or nuclear weapon.
- Offensive cyberoperations. A model's ability to provide significant capability uplift to an actor seeking to carry out highly disruptive or destructive cyberattacks, including on critical infrastructure.
- Advanced autonomy. A model's ability to complete expert-level tasks autonomously, including AI research and development.

As the Frontier Governance Framework further details, models assessed as posing low or medium risk in relation to tracked high-risk capabilities may be deployed with appropriate safeguards. Models assessed as having high or critical risk are subject to further review and safety and security mitigations prior to deployment. If, during the implementation of our Frontier Governance Framework, we identify a risk that we cannot sufficiently mitigate, then we will pause development and deployment until the point at which mitigation practices evolve to meet the risk.

In the context of the development and deployment of advanced AI systems, we have defined areas of policy and established governance tools to address various risk scenarios, including: "restricted uses," "sensitive uses," and "unsupported uses."

- Restricted uses are subject to specific restrictions, typically on AI development or deployment. They are defined by our Office of Responsible AI and updated periodically.
- Sensitive uses involve scenarios that could have significant impacts on individuals or society, such as those affecting life opportunities, physical safety, or human rights. These uses require notification to our Office of Responsible AI and additional governance steps.
- Unsupported uses refer to reasonably foreseeable uses for which the AI system was not designed or evaluated or that we recommend customers avoid. As part of our Microsoft Responsible AI Impact Assessment process, we provide guidance to teams to think through unsupported uses.

This layered approach ensures that highly capable AI models and AI systems are developed and deployed responsibly, with ongoing efforts to map, measure, manage, and govern risks. By categorising AI technologies and use scenarios and implementing robust governance processes, Microsoft aims to safeguard against risks while promoting the effective use of AI technologies.

b. What practices does your organization use to identify and evaluate risks such as vulnerabilities, incidents, emerging risks and misuse, throughout the AI lifecycle?

At Microsoft, tactics we use to identify and prioritize AI risks include threat modeling, responsible AI impact assessments, customer feedback, incident response and learning programs, external research, and AI red teaming. These exercises inform decisions about planning, mitigations, and the appropriateness of deploying an AI model or application for a given context. Equally important is our ability to remain flexible and responsive to new or previously unforeseen risks that arise at any stage of development or deployment, including post-deployment.

Red teaming in particular has become an industry best practice to identify potential risks by simulating adversarial user behavior. For pre-deployment red teaming of our highest risk AI systems and models, we leverage the expertise of Microsoft's AI Red Team (AIRT), a centralized team of professional red teamers that operates independently of product teams. Guided by tools and resources developed by expert red teamers, product teams across Microsoft also perform pre-deployment red teaming of their AI systems and models. Risks that are identified during red teaming inform how we prioritize and design measurement and mitigation tasks.

A hallmark feature of Microsoft's AI governance program is the intentional collaborations we nurture between engineering, research, and policy. This collaboration is particularly important to quickly advance the science of AI measurement and evaluation. Our ability to develop effective and valid risk measurement capabilities that move at the speed of innovation became increasingly more evident in 2024 as AI capabilities and the creative ways they are used continued to grow more complex.

AI risk measurement helps us to prioritize mitigations and assess their efficacy. For example, we seek to measure our AI applications' abilities to generate certain types of content and the efficacy of our mitigations in preventing that behavior. In addition to regularly updating our measurement methods, we also share resources and tools that support the measurement of risks and risk mitigations with our customers.

We continue to leverage the power of generative AI models to scale our measurement practices. Our automated measurement pipelines involve three main components. The first component is the AI system or model that is being evaluated. The second component is an AI model, usually an LLM, or in some cases a multimodal model, that is instructed to interact with the first component by simulating adversarial user behavior. The interaction between these first two components generates simulated interactions that make up test sets. The third component is also an AI model that serves as a judge, assessing and annotating each simulated interaction in the test sets based on instructions developed by human experts. The accuracy of the AI annotations is compared against human annotations, which informs how the instructions provided to the AI model need to be adjusted. Finally, the annotated test sets are used to calculate metrics about the risks, which inform downstream mitigation tasks.

In 2024, we made significant improvements to our measurement pipelines with the primary goal of expanding risk coverage across different modalities and risk types, enhancing the reliability of metrics generated, and leveraging new approaches to expose safety vulnerabilities. We expanded our measurement pipeline to cover two new risk categories: the generation of election-critical information and reproduction of protected materials. Our broader approach to mapping, measuring, and managing AI-related risks for 2024 elections is covered in Section 5D.

Our testing coverage for protected materials included content such as song lyrics, news, recipes, and code from public, licensed GitHub repositories. We also expanded our ability to measure an AI system's ability to generate sexual, violent, and self-harm content and content related to hate and unfairness across both image generation and image understanding modalities. Furthermore, with increased support for audio modalities in the latest releases of generative AI models, we expanded measurement support for audio interactions by adding a transcription layer and running the text output through our measurement pipelines.

To improve the reliability of our metrics, we leveraged several prompt engineering techniques to optimize the performance of the annotation component of our measurement pipeline.To better measure safety vulnerabilities, we applied adversarial fine-tuning to components of our measurement pipeline to generate prompts that are more effective at revealing potential safety vulnerabilities in the system, which in turn guides risk management.

Looking ahead, we are integrating more advanced adversarial techniques and attack strategies to systematically measure vulnerabilities that could be exploited by malicious actors. We also plan to improve our evaluators for accuracy and support granular metrics, which in turn will improve their interpretability and provide transparency through safety scorecards.

Further, we will continue to expand our testing risk coverage while continuing to refine our existing evaluations across various settings, newer models, modalities, and tools.We will also continue fostering collaborations with Microsoft Research to incorporate the latest advances in the science of AI risk evaluation into our tools and practices. This includes building measurement frameworks to better understand, interrogate, and compare measurements comprehensively through multiple lenses.

c. Describe how your organization conducts testing (e.g., red-teaming) to evaluate the model's/system's fitness for moving beyond the development stage?

Microsoft leverages red teaming, quantitative evaluations, and external expertise to implement and improve our approach to testing. We consider red teaming as more focused on risk identification and systematic measurement as more enabling of risk evaluation. As an overarching framework for risk identification and evaluation (or "mapping and measurement") across the AI lifecycle, including during development, we focus on: 1) manual red teaming (or "adversarial testing"), during which, depending on risk level, the product team or an independent team of human experts manually probes AI technologies and products; 2) automated red teaming, during which we use tools to build upon human-generated adversarial probes to test variations at scale; and 3) systematic automated measurement, during which we use AI tools to test for risks at scale (enabling quantitative as well as qualitative evaluation).

As one approach to systematic automated measurement (detailed in *A Framework for Automated Measurement of Responsible AI Harms in Generative AI Applications*), we have leveraged a framework comprised of two key components: a) a data generation component, through which real-world AI content generation is simulated through the use of templates and parameters; and b) an evaluation component, through which AI-generated content is evaluated, providing both quantitative and qualitative outputs (e.g., numerical annotations of harm severity and written snippets about annotation reasoning)). More detail on our AI risk measurement practices is also included in Section 1B.

We find value in leveraging both internal and external expertise in risk identification and evaluation. As elaborated in Section 1F, external vendors and other experts contribute to our risk identification and evaluation at different stages along the AI lifecycle and for various reasons, including where our processes and practices may benefit from additional scale or specific expertise.

Microsoft internally manually red teams a wide range of AI technologies and products, including AI models, platform services, and applications, for AI security and responsible AI risks. Manual red teaming of AI models and products not designated as high risk is generally carried out by the teams developing them and reviewed as part of our launch readiness assessment. Manual red teaming of AI models and products designated as high risk is carried out by our internal AI Red Team, which is comprised of multi-disciplinary experts and independent from any products teams. Microsoft established our AI Red Team in 2018 to identify AI security vulnerabilities, and it has since evolved to evaluate other risks associated with AI.

We've also developed automated red teaming tools—used by both product-based red teams and our expert AI Red Team—and made some available open access. In February 2024, we released a red teaming accelerator, Python Risk Identification Tool for generative AI (PyRIT), enabling developers to proactively identify risks in their generative AI applications. PyRIT accelerates an evaluator's work by expanding on initial red teaming prompts and automatically scoring outputs using content filters.

PyRIT has received over 2,000 starts on GitHub and been copied more than 200 times by developers for use in their own repositories, where it can be modified to fit their use cases. In April 2025, Microsoft announced PyRIT's integration with Azure AI Foundry. Customers using this new capability in Azure AI Foundry can simulate adversarial attack techniques and generate red teaming reports that help track risk mitigation improvements throughout the AI development lifecycle.

d. Does your organization use incident reports, including reports shared by other organizations, to help identify risks?

Yes

e. Are quantitative and/or qualitative risk evaluation metrics used and if yes, with what caveats? Does your organization make vulnerability and incident reporting mechanisms accessible to a diverse set of stakeholders? Does your organization have incentive programs for the responsible disclosure of risks, incidents and vulnerabilities?

As the science of AI evaluations is still developing, there are numerous limitations to current approaches to evaluations, impacting Microsoft's internal quantitative tests and external benchmarks. Scientifically valid approaches to generative AI evaluations are nascent and rapidly evolving; there are gaps in how to systematize, operationalize, and measure relevant theoretical constructs needed to perform reliable testing (for a link to a published paper in which this is discussed, see Measurement and Fairness (arxiv.org)).

Similarly, many existing widely used benchmarks are not robust to distribution shifts and other measures of validity over time and fail to translate well to real-world settings (for a link to a published paper in which this is discussed, see: 2404.09932 (arxiv.org)). Benchmarks are also not advancing as rapidly as AI functionality; for example, the breadth of language capabilities goes beyond the breadth of languages benchmarked.

While evaluation is critical to risk management, there are also additional steps, such as leveraging vulnerability and incident reporting, that can be leveraged across the AI lifecycle as part of a comprehensive governance approach. For the past two decades, the Secure Development Lifecycle (SDL) that Microsoft has implemented across the company has included a response phase that focuses on handling unforeseen issues and applying these learnings to future releases. We apply the infrastructure, processes, and best practices developed over the years through SDL to our AI systems.

Across Microsoft, product teams are required to have repeatable processes to collect user feedback and to triage and address issues that arise after the release of an AI system. Teams are also required to build feedback collection mechanisms within their products so users can more easily report concerns. When possible, teams employ automation to enable quick action on well-understood problems.

Initial concerns reported via the Microsoft Security Response Center's (MSRC) researcher portal are triaged and assessed by expert teams. Microsoft employees are also provided with multiple avenues to raise concerns, including an anonymous reporting channel. If these concerns are assessed to warrant an incident response, appropriate teams are assembled and coordinated by response specialists, who manage root-cause analysis, mitigation, and communication. After the incident is mitigated, a postmortem analysis is typically conducted to distill 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 productsAI products, and efforts to build community and offer public thanks through conferences like BlueHat and the Microsoft Researcher Recognition Program.

f. Is external independent expertise leveraged for the identification, assessment, and evaluation of risks and if yes, how? Does your organization have mechanisms to receive reports of risks, incidents or vulnerabilities by third parties?

Microsoft leverages independent external expertise in conducting tests in multiple ways. In advance of the deployment of highly capable models subject to our Frontier Governance Framework, evaluations for a set of tracked high-risk capabilities (see Section 1A) involve qualified and expert external actors that meet relevant security standards, including those with domain-specific expertise, as appropriate. As our Frontier Governance Framework acknowledges, we also benefited from the advice of external experts in formulating our list of tracked high-risk capabilities for which we apply these governance steps.

More broadly, external experts contribute to our evaluation practices in important ways. We identify third-party evaluators who have special expertise in evaluating specific risks, and we engage them to help us with tailored evaluations. These engagements help us better understand how to conceptualize certain risks and may surface opportunities for us to improve our risk mitigations.

Specifically, external experts have contributed to our development of testing sets, including by writing and seeding prompts, and to our paradigms for red teaming and systematic measurement, with consultations with external subject matter experts helping refine our risk conceptualization. As highlighted in *A Framework for Automated Measurement of Responsible AI Harms in Generative AI Applications*, creating harm- or risk-specific measurement resources requires domain-specific expertise. They have also supported translation or localization guidance for AI evaluations.

We also have in place mechanisms to receive reports from third parties of risks and potential vulnerabilities and incidents, including via reported via the Microsoft Security Response Center's (MSRC) researcher portal, as described in Section 1E.

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

We see global efforts to build consensus-based frameworks, best practices, and standards as critical to promoting the trust and coherence that underpin broad adoption of a global technology. The development of AI standards for risk assessment and evaluation is especially important to advance rapidly; specifying the types of evaluations needed, ways to substantiate reliability, and expectations for evidence will be crucial in enhancing AI assessment and evaluation science.

We have made significant contributions to and use various best practices, including international or other formal standards or industry technical specifications. For example, in the context of industry standards, best practices, and technical specification efforts, we have been developing specifications for content provenance and authentication through the Coalition for Content Provenance and Authenticity (C2PA), of which Microsoft is a founding member. We have implemented C2PA in multiple services, such as LinkedIn, where content carrying the technology is automatically labelled, and Bing Image Creator, where all images created include 'Content Credentials.' Additionally, we are actively involved in developing tools and defining practices for AI evaluations, such as the MLCommons platform for AI risk and reliability tests and Frontier Model Forum (FMF) Issue Briefs.

Microsoft uses international standards for cybersecurity, including ISO/IEC 27001, ISO/IEC 27017, ISO/IEC 29147, and ISO/IEC 30111 for information security management systems, cloud security, and coordinated vulnerability disclosure and management. We have also led on standardization for securing AI systems with the intent of providing guidance and awareness of security risks to help organizations better protect AI systems against an evolving threat landscape.

Microsoft was a key contributor to the conception and development of ISO/IEC 42001:2023. In early 2025, we successfully achieved the ISO/IEC 42001:2023 certification for Microsoft 365 Copilot and Microsoft 365 Copilot Chat. This certification confirms that an independent third party has validated Microsoft's application of the necessary framework and capabilities to effectively manage risks and opportunities associated with the continuous development, deployment, and operation of M365 Copilot and M365 Copilot Chat.

Microsoft also actively participates in Standards Development Organizations (SDOs), including: 1) ISO/IEC JTC1 SC42 Artificial Intelligence, and 2) CEN/CENELEC JTC21 Artificial Intelligence. Our involvement in these committees is facilitated through the National Standards Body members, where we have taken on a number of leadership roles. In SC42, we have been leading efforts, sharing our experiences and knowledge in developing and implementing RAI practices, on key standards deliverables related to terminology and concepts, governance, responsible AI, risk management, data quality management, and conformity assessment.

In JTC21, we have focused on harmonized standards requested by the European Commission to support the EU AI Act. Microsoft also actively participates in the standardization efforts led by the U.S. National Institute of Standards and Technology (NIST) and contributed to the development of the NIST AI Risk Management Framework (RMF) by providing feedback on drafts.

We remain committed to advancing internationally recognized standards that help to establish consistent practices, enhance accountability, and foster trust in AI technologies.

h. How does your organization collaborate with relevant stakeholders across sectors to assess and adopt risk mitigation measures to address risks, in particular systemic risks?

Microsoft collaborates with industry peers, academia, civil society, and governments to develop, share, and adopt risk mitigation measures. These collaborations take many forms, including research, standards development, and open-source projects.

Through the Frontier Model Forum, for example, we have collaborated with others in industry to develop resources on safety evaluations (see: Issue Brief: Early Best Practices for Frontier AI Safety Evaluations - Frontier Model Forum) and security best practices (see: Issue Brief: Foundational Security Practices - Frontier Model Forum). Through Partnership on AI, we have also contributed to the development of guidance for safe foundation model deployment.

Section 2 - Risk management and information security

a. What steps does your organization take to address risks and vulnerabilities across the AI lifecycle?

Our iterative process for mapping, measuring, and managing risks during the design, development, deployment, and operation of AI models and systems is at the foundation of our approach to holistic risk management across the AI lifecyle. We incorporate risk and vulnerability detection tools and mitigations into our engineering pipelines and platforms, building responsible- and secure-by-design approaches into our product development. We use a combination of automated and manual testing to map and measure risks and to evaluate the efficacy of mitigations, and we use insights from monitoring and incidents to inform future risk and vulnerability assessments and mitigations (for more information, please see Section 1B). This process is guided by our internal policies, including our Responsible AI Standard and associated resources, such as our Impact Assessment Template, and our Security Development Lifecycle (SDL) and associated resources (for more information, see Section 2G).

Emerging risks may be identified by subject matter experts, either from their own discoveries or from reporting by external sources. These subject matter experts come from diverse disciplines, including research, red teams, policy, and engineering. Microsoft may publish findings on emerging risks discovered by internal subject matter experts, as we did with Crescendo and Skeleton Key attacks (for an example publication based on the research of a Microsoft subject matter expert, see: https://www.microsoft.com/en-us/security/blog/2024/06/26/mitigating-skeleton-key-a-new-type-of-generative-ai-jailbreak-technique/).

Misuses may be identified through internal and product safety monitoring tools and manual reporting from internal and external sources. Examples of abuse monitoring and content safety tools include our built-in abuse monitoring of the Azure OpenAI Service, based on our Azure OpenAI Service Code of Conduct, and Azure Content Safety. We also welcome external reports of abuse (for more information, see https://msrc.microsoft.com/report/).

From a product perspective, Microsoft requires evaluations pre-deployment and ongoing monitoring post deployment. Product teams may engage in manual or automated testing as part of their ongoing system improvement and monitoring processes. In addition, our AI Red Team, our central team of multi-disciplinary experts most often leveraged in the context of high-risk AI technologies and products, leverages manual and automated probing post deployment, focusing on situational and time-bound events (e.g., elections), novel techniques, and hardening systems against adversarial trends. This post-deployment red teaming complements other aspects of our risk management program, such as incident response, feedback collection and processing, and ongoing monitoring.

b. How do testing measures inform actions to address identified risks?

As detailed in Section 1B, AI risk measurement helps us to prioritize mitigations and assess their efficacy. For example, we seek to measure our AI applications' abilities to generate certain types of content and the efficacy of our mitigations in preventing that behavior. In addition to regularly updating our measurement methods, we also share resources and tools that support the measurement of risks and risk mitigations with our customers.

As another example, with increased support for audio modalities in the latest releases of generative AI models, we expanded measurement support for audio interactions by adding a transcription layer and running the text output through our measurement pipelines. To improve the reliability of our metrics, we leveraged several prompt engineering techniques to optimize the performance of the annotation component of our measurement pipeline.To better measure safety vulnerabilities, we applied adversarial fine-tuning to components of our measurement pipeline to generate prompts that are more effective at revealing potential safety vulnerabilities in the system, which in turn guides risk management.

c. When does testing take place in secure environments, if at all, and if it does, how?

For internal evaluations, Microsoft product teams leverage secure testing platforms, and our AI Red Team a) creates a secure virtual machine for testing whenever possible and b) manages reporting and storage of testing results in an environment that requires custom access requests for every operation. Credentials are also required for external evaluators to send API requests for testing access.

d. How does your organization promote data quality and mitigate risks of harmful bias, including in training and data collection processes?

Microsoft is committed to implementing and supporting responsible data policies and practices for inputs and outputs of AI models and applications. Our Responsible AI Standard and accompanying privacy, security, and accessibility standards, which apply to all AI systems that we develop and deploy, establish numerous data requirements consistent with our Responsible AI principles.

As a result, depending on their product scenario, product teams may be required to assess the quantity and suitability of data sets, inclusiveness of data sets, representation of intended uses in training and test data, limitations to generalizability of models given training and testing data, and how they meet data collection and processing requirements among other mandates. Our Impact Assessment and other Responsible AI tools help teams conduct these assessments and provide documentation for review.

Through transparency mechanisms, Microsoft provides context to customers and other stakeholders on data processed by AI systems like the Azure OpenAI Service, including user prompts and generated content, augmented data included with prompts (i.e., for grounding), and user-provided training and validation data. Customer data is also processed to analyze prompts, completions, and images for harmful content or patterns of use that may violate our Code of Conduct or other applicable product terms. Microsoft Purview Data Quality also enables business domain and data owners to assess and oversee the quality of their data ecosystem, facilitating targeted actions for improvement.

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

Microsoft leverages multiples measures and safeguards to mitigate the risk of our AI tools being misused for copyright infringement. These measures include meta-prompts, classifiers, and controls that add instructions to a user prompt to limit potentially harmful or infringing outputs. For example, Copilot will decline to provide song lyrics or provide extracts from books that are available online. The operation of meta-prompts and classifiers are further explained in Microsoft's white paper, Governing AI: A Blueprint for the Future.

Microsoft continues to improve current mitigations and implement new ones in response to our learnings and encourages rightsholders to help us think through effective practices. GitHub's reference feature was developed with engagement and feedback from the developer community. It enables developers to choose whether to block code that matches code in public repositories or allow the code suggestions with information about the matching public code on GitHub.

Microsoft has also committed to indemnify and defend customers of our commercial Copilot offerings and Azure OpenAI against claims of copyright infringement as a result of the output content generated by these tools, provided that the customer has used the guardrails built into the products. This Copilot Copyright Commitment, first launched in 2023, reflects Microsoft's commitment to building responsible AI-powered products and tools that limit the risk of infringing outputs, allowing users to focus on delivering value without fear of unreasonable litigation costs related to potential qualifying claims. It also provides a strong incentive for Microsoft customers to adopt responsible practices to mitigate these risks. This program helps Microsoft educate users on appropriate uses of AI technology and reinforce how users can respect intellectual property rights. In 2024, after listening closely to questions and feedback from our partners, we extended our Customer Copyright Commitments to include our reseller partners. This means that our resellers can assure their customers that they will receive the same protections as customers who purchase qualifying offerings directly from Microsoft.

We also recognize that some rightsholders prefer that their work is not used in AI training. While there is no basis in copyright law for a rightsholder to prevent the use of a publicly available or otherwise legally accessed work for AI training, organizations may still take steps to identify and respect a rightsholder's interest in preventing their works from being used for AI training, through voluntary measures. These norms already exist in search; for decades, web publishers have relied on a series of machine-readable signals associated with robots.txt to declare whether the site welcomes a crawler to index its pages and present information within search results. Microsoft abides by these well-established norms and keeps applying learnings to give content owners a clear, transparent way to signal their preferences regarding the use of their data in AI training.

These protocols even include preferences that relate to how much of a work to display within a search result (such as maximum snippet length (See Robots Meta Tags (bing.com)). Similar norms are being deployed with AI training on a voluntary basis in which an AI crawler or Bot will avoid a publisher's site, a specific page, or other content for the purpose of AI training based on information found in the Robots.txt file. Examples can be found here: https://platform.openai.com/docs/bots and Robots Meta Tags (bing.com).

f. How does your organization protect privacy? How does your organization guard against systems divulging confidential or sensitive data?

Microsoft policies dictate that training data and fine-tuned models are available exclusively for use by the customer, can be double encrypted at rest (by default with Microsoft's AES-256 encryption and optionally with a customer-managed key), and can be deleted by the customer at any time. Microsoft will not store or process customer data outside a customer-specified geography without customer authorization. For our platform for AI enterprise operations Azure AI Foundry, privacy measures are effectively in place to ensure that customer data is handled securely and in accordance with relevant data protection laws. Azure AI Foundry also adheres to strict data encryption standards and allows customers to specify the geography for data storage and processing, ensuring that their data remains within the chosen geography.

We provide deeper information about how we protect individuals' privacy in Microsoft Copilot and our other AI products in our transparency materials, such as M365 Copilot FAQs and The New Bing: Our Approach to Responsible AI. Our users and the public can also review the Microsoft Privacy Statement, which provides information about our privacy practices and controls for all of Microsoft's consumer products. Microsoft provides customers with the ability to control their interactions with Microsoft products and services and honors their privacy choices.

We also provide the following resources for customers and policymakers:

- Through the Microsoft Privacy Dashboard, our account holders can access, manage, and delete their personal data and stored conversation history.
- Microsoft Purview provides a variety of capacities that offer additional data security and compliance controls: Microsoft Purview data security and compliance protections for Microsoft Copilot and other generative AI apps | Microsoft Learn.
- To understand our commitment to protecting our enterprise customer's data see: FAQ: Protecting the Data of our Commercial and Public Sector Customers in the AI Era - Microsoft Community Hub.
- To better understand the use of data by Azure see: Data, privacy, and security for Azure OpenAI Service - Azure AI services | Microsoft Learn.
- To better understand data use for Dynamics 365 and Power Platform see: FAQ for Copilot datasecurity and privacy for Dynamics 365 and Power Platform - Power Platform | Microsoft Learn.

g. How does your organization implement AI-specific information security practices pertaining to operational and cyber/physical security?

- **i. How does your organization assess cybersecurity risks and implement policies to enhance the cybersecurity of advanced AI systems?**
- **ii. How does your organization protect against security risks the most valuable IP and trade secrets, for example by limiting access to proprietary and unreleased model weights? What measures are in place to ensure the storage of and work with model weights, algorithms, servers, datasets, or other relevant elements are managed in an appropriately secure environment, with limited access controls in place?**
- **iii. 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?**
- **iv. How often are security measures reviewed?**
- **v. Does your organization have an insider threat detection program?**

How does your organization assess cybersecurity risks and implement policies to enhance the cybersecurity of advanced AI systems?

Microsoft's foundational security policy, our Security Development Lifecycle (SDL), is regularly updated in response to emerging risks, techniques, and technologies. This includes adding AI-specific information security requirements and adding AI-specific considerations to existing information security requirements. For example, our threat modeling requirement has been updated with specific guidance on creating threat models for AI systems.

Our product and operational security measures are implemented by security subject matter experts from both product teams and centralized functions, including security-focused experts from our AI Red Team. Our Microsoft Threat Intelligence community is made up of more than 10,000 security researchers, analysts, and threat hunters with a variety of backgrounds. Our Microsoft Threat Intelligence Center carries out proactive monitoring of our systems, tracking over 300 unique threat actors in 2023, including 160 nation-state actors and 50 ransomware groups.

In November 2023, we launched the Secure Future Initiative (SFI), a multiyear commitment that advances the way we design, build, test and operate our Microsoft technology to ensure that our solutions meet the highest standards for security. There are three principles and six pillars to SFI. The three principles are: secure by design; secure by default; and secure operations. The six pillars are protect identities and secrets; protect tenants and isolate systems; protect networks; protect engineering systems; monitor and detect cyberthreats; and accelerate response and remediation. Through SFI, we have increased our investment in tools that scale continuous evaluation, building on our own research as well as learnings from external reports.

Microsoft products undergo rigorous security assessments by third parties, as reflected by our compliance offerings and associated documentation for ISO/IEC 27001, ISO/IEC 27017, SOC 2, and other programs (see, e.g.: Azure compliance documentation | Microsoft Learn).

Finally, we have annual security training requirements for employees to ensure skills and knowledge remain current.

How does your organization protect against security risks the most valuable IP and trade secrets, for example by limiting access to proprietary and unreleased model weights? What measures are in place to ensure the storage of and working with model weights, algorithms, servers, datasets, or other relevant elements are managed in an appropriately secure environment, with limited access controls in place?

SDL protection, detection, and response requirements and SFI commitments apply to AI technology. For instance, Microsoft employs strong identity and access control, holistic security monitoring (for both external and internal threats) with rapid incident response, and continuous security validation (such as simulated attack path analysis) for our AI environments.Microsoft separates development and production environments, follows least privilege principles, and uses just-in-time access control.

Model weights, and other relevant elements, are encrypted-at-rest and encrypted-in-transit to mitigate the potential risk of model theft, and more stringent security controls are applied based on risk, such as for protecting highly capable models. Robust physical, operational, and network security measures, including for supplier management, identity and access management, and insider threat monitoring, also protect cloud infrastructure, including AI datacenters that enabling training and hosting of models.Supplier security and privacy are governed by our Supplier Security and Privacy Assurance program.

Access to physical datacenter facilities is tightly controlled, with outer and inner perimeters and increasing security at each level, and subject to a least privileged access policy, whereby personnel with an approved business need are granted time-limited access.We log and retain access requests and analyze data to detect anomalies and prevent and detect unnecessary or unauthorized access.We also employ multiple strategies for securing network boundaries.

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?

Microsoft's vulnerability management process helps ensure timely response, including investigation and remediation as appropriate, to potential vulnerabilities, whether discovered internally or via external sources. We leverage international standards for vulnerability disclosure and management, including ISO/IEC 29147 and ISO/IEC 30111, consistent with our commitment to Coordinated Vulnerability Disclosure (CVD) (see more information in our responses to Sections 1E and 1G).

Incident detection and response are handled through robust monitoring and incident response processes based on SDL and SFI, which includes handling unforeseen issues and learning from them for future releases. AI product teams are required to have repeatable processes to collect user feedback and triage and address issues that are found or reported.

We also collaborate with other stakeholders to address identified risks and help support appropriate action in response to remediated vulnerabilities. Through the Frontier Model Forum, we have recently joined with other members in signing a first-of-its-kind agreement to facilitate information sharing about vulnerabilities, threats, and capabilities of concern unique to frontier AI. This builds upon our longstanding commitment to work closely with others, including security researchers through CVD and industry through our Microsoft Active Protections Program (MAPP) and other initiatives, to help improve ecosystem security. Members of MAPP receive security vulnerability information from the Microsoft Security Response Center in advance of Microsoft’s monthly security update, enabling them to more quickly provide protections through their security software or devices.

How often are security measures reviewed?

Microsoft has long had a rigorous approach to continuous security improvement, and our Secure Future Initiative has helped us double down on a security-first culture and security governance that help us integrate feedback and learnings from issues, vulnerabilities, and incidents on an ongoing basis.

Security standards are reviewed at least annually, though individual requirements may be added or updated between reviews. Security reviews of developed systems are typically conducted during initial design, before release phases or major updates, and at least annually.

We also leverage practices and tools help regularly measure and strengthen system security, including vulnerability scanners, scheduled security updates, asset discovery, continuous monitoring, patch management software, and compliance validation. Security Configuration Management (SCM) software helps track changes and ensure compliance with security policies.

Does your organization have an insider threat detection program?

Microsoft has a dedicated team, the Digital Security and Resilience Insider Threat Team (DSR ITT), which focuses on insider threat detection and prevention. Microsoft also assumes insider threats in the threat modeling covered by the Security Development Lifecycle (SDL) mentioned above. All red team operations assume insider access, resulting in the privileged access requirements described above

h. How does your organization address vulnerabilities, incidents, emerging risks?

Microsoft takes action whether risks or vulnerabilities are identified by internal or external sources, though the paths through which action is guided varies. Externally reported vulnerabilities that meet our triage bar are assigned to the appropriate product development team to mitigate, and the mitigation is verified by the Microsoft Security Response Center and, when feasible, with the reporter of the vulnerability. Internally reported vulnerabilities are reported directly to the appropriate product development team that will then triage and, as the triage bar is met, mitigate the vulnerability.

For Microsoft products, more systematic steps may also be taken as warranted. Identified risks and vulnerabilities may also be addressed via updates to our Responsible AI Standard, Privacy Standard, or Security Development Lifecycle (SDL) as well as accompanying guidance documentation, training materials, or communication to Responsible AI leaders in relevant product development teams. For some identified risks and vulnerabilities, we may also implement automated detection and/or mitigation of the risk in our platform services or development environments or in standalone tools.

When identified risks and vulnerabilities affect the broader ecosystem of AI technologies and products, we also take action to support or collaborate with external stakeholders. For example, as we have with the Crescendo and Skeleton Key attacks, we may share information publicly about emergent risks and mitigations. When products of other organizations are affected by risks or vulnerabilities that Microsoft identifies, we also share information directly with impacted organizations, consistent with the principles and guidelines of Coordinated Vulnerability Disclosure, and as highlighted above with the FMF agreement and MAPP.

Section 3 - Transparency reporting on advanced AI systems

a. Does your organization publish clear and understandable reports and/or technical documentation related to the capabilities, limitations, and domains of appropriate and inappropriate use of advanced AI systems?

- **i. How often are such reports usually updated?**
- **ii. How are new significant releases reflected in such reports?**
- **iii. Which of the following information is included in your organization's publicly available documentation: details and results of the evaluations conducted for potential safety, security, and societal risks including risks to the enjoyment of human rights; assessments of the model's or system's effects and risks to safety and society (such as those related to harmful bias, discrimination, threats to protection of privacy or personal data, fairness); results of red-teaming or other testing conducted to evaluate the model's/system's fitness for moving beyond the development stage; capacities of a model/system and significant limitations in performance with implications for appropriate use domains; other technical documentation and instructions for use if relevant.**

A key goal of our responsible AI program is transparency. As a provider of AI tools, services, and components, we understand our role in equipping customers with the information they need to innovate responsibly. Our transparency mechanisms cover product-level information as well as broader context on our AI governance program and practices.

We provide product-level information for models, platform services, and AI systems or applications. For models that we train and deploy, we provide and update model cards and, as relevant, provide additional documentation, such as technical reports, when new models are released or when significant updates are made. For example, we have provided our Phi-4 Model Card and Technical Report. We also published a detailed report about Phi-3 safety post-training and our "Break-Fix" Cycle: [2407.13833] Phi-3 Safety Post-Training: Aligning Language Models with a "Break-Fix" Cycle.

For platform services, we leverage Transparency Notes, which convey important information about the responsible use of these services (see, e.g., for Azure OpenAI Service: Transparency Note for Azure OpenAI - Azure AI services | Microsoft Learn). Since 2019, we've published 40 Transparency Notes, all of which follow a specific template that includes the capabilities, limitations, and intended uses of these platform services.

In 2023, we expanded our transparency documentation policies to require AI systems or applications, such as our Copilots, to publish Responsible AI Frequently Asked Questions (FAQs) and other important disclosures. This includes, for example, in-product disclosure in products like Microsoft Copilot and M365 Copilot to inform users they're interacting with an AI application, as well as citations to source material to help users verify information in the responses and learn more. Other important notices may include disclaimers about the potential for AI to make errors or produce unexpected content. These user-friendly transparency documents and product integrated notices are especially important in our Copilot experiences, where users are less likely to be developers or to seek out documentation.

Transparency Notes and FAQs are revised to reflect meaningful updates to capabilities, features, or functionality of the underlying service. For example, our Transparency Note for Azure OpenAI Service, originally published in 2022, was updated in May 2024 to address the addition of GPT-4o.

In May 2024, Microsoft released our inaugural Responsible AI Transparency Report. This report provides extensive insight into Microsoft's overall responsible AI program and will be updated annually. It describes how we build applications that use generative AI, make decisions and oversee the deployment of those applications, support our customers as they build their own generative applications, and learn, evolve, and grow as a responsible AI community. We plan to publish our second annual Responsible AI Transparency Report next month.

b. 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?

Our research and information-sharing practices, which are integral to supporting research efforts, strengthening the outcomes of evaluations, and enhancing our understanding of the impacts of AI systems. The transparency mechanisms (e.g., technical reports on models, Transparency Notes for platform services, and FAQs for applications) described in Section 3A are the primary way we publicly share information with diverse stakeholders regarding risk evaluations and impacts of advanced AI systems

By sharing our findings and collaborating with stakeholders, we aim to contribute to the broader knowledge base and ensure that our practices are aligned with the highest standards of transparency and accountability. We provide additional details on our investments in research and the advancement of AI safety in Section 6C.

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

Transparency is a core governance objective in integrating AI into Microsoft products and services in a way that promotes user control and privacy and builds trust. That's why we are committed to building transparency into people's interactions with our AI systems.

This approach starts with providing clarity to users when they are interacting with an AI system if there is risk that they will be confused. We also provide real-time information to help people better understand how AI features work. Microsoft Copilot uses a variety of transparency approaches that meet users where they are. Copilot provides clear information about how it collects and uses data, as well as its capabilities and its limitations. Our approach to transparency also helps people understand how they can best leverage the capabilities of Copilot as an everyday AI tool and provides opportunities to learn more and provide feedback.

- **Transparent choices and disclosures while users engage with Microsoft Copilot**: To help people understand the capabilities of these new AI tools, Copilot provides in-product information that clearly lets users know that they are interacting with AI and provides easy-to-understand choices in a conversational style. As people interact, these disclosures and choices help provide a better understanding of how to harness the benefits of AI and limit potential risks.
- **Grounding responses in evidence and sources:** Copilot also provides information about how its responses are centered, or "grounded," on relevant content. In our AI offerings in Bing, Copilot.microsoft.com, Microsoft Edge, and Windows, our Copilot responses include information about the content from the web that helped generate the response. In Copilot for Microsoft 365, responses can also include information about the user's business data included in a generated response, such as emails or documents that you already have permission to access. By sharing links to input sources and source materials, people have greater control of their AI experience, can better evaluate the credibility and relevance of Microsoft Copilot outputs, and can access more information as needed.
- **Data protection user controls:** Microsoft provides tools that put people in control of their data. We believe all organizations offering AI technology should ensure consumers can meaningfully exercise their data subject rights. Microsoft provides the ability to control your interactions with Microsoft products and services and honors your privacy choices. Through the Microsoft Privacy Dashboard, our account holders can access, manage, and delete their personal data and stored conversation history. In Microsoft Copilot, we honor additional privacy choices that our users have made in our cookie banners and other controls, including choices about data collection and use.
- **Additional transparency about our privacy practices:** Microsoft provides deeper information about how we protect individuals' privacy in Microsoft Copilot and our other AI products in our transparency materials such as M365 Copilot FAQs and The New Bing: Our Approach to Responsible AI, which are publicly available online. These transparency materials describe in greater detail how our AI products are designed, tested, and deployed – and how our AI products address ethical and social issues, such as fairness, privacy, security, and accountability. Our users and the public can also review the Microsoft Privacy Statement which provides information about our privacy practices and controls for all of Microsoft's consumer products.

AI systems are new and complex, and we are still learning how we can best inform our users about our ground-breaking new AI tools in a meaningful way. We continue to listen and incorporate feedback to ensure we provide clear information about how Microsoft Copilot works.

d. 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?

Microsoft provides transparency with regard to AI models through model cards and technical reports, which typically include high level information regarding data sources for training. Through documentation, Microsoft also provides context to customers and other stakeholders on data processed by AI systems like the Azure OpenAI Service, including user prompts and generated content, augmented data included with prompts (i.e., for grounding), and user-provided training and validation data. Customer data is also processed to analyze prompts, completions, and images for harmful content or patterns of use that may violate our Code of Conduct or other applicable product terms.

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

Beyond the examples provided in Section 3A-D, we may provide further documentation to empower AI developers with the information they need to innovate responsibly.

For example, with Phi models:

Phi-3 – Model Cards. https://huggingface.co/collections/microsoft/phi-3-6626e15e9585a200d2d761e3

Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. https://arxiv.org/pdf/2404.14219

Phi-3 Safety Post-Training: Aligning Language Models with a "Break-Fix" Cycle. https://arxiv.org/pdf/2407.13833

Phi-4 - A Microsoft Collection. https://huggingface.co/collections/microsoft/phi-4-677e9380e514feb5577a40e4

Phi-4 Technical Report. https://arxiv.org/abs/2412.08905

Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. https://arxiv.org/abs/2503.01743

For example, with Azure OpenAI Service:

Data, privacy, and security for Azure OpenAI Service - Azure AI services | Microsoft Learn

Azure OpenAI Service frequently asked questions - Azure AI services | Microsoft Learn

For example for Azure AI Foundry:

Run evaluations online in Azure AI Foundry - Azure AI Foundry | Microsoft Learn

How to manually evaluate prompts in Azure AI Foundry portal playground - Azure AI Foundry | Microsoft Learn

How to view evaluation results in Azure AI Foundry portal - Azure AI Foundry | Microsoft Learn

Evaluation and monitoring metrics for generative AI - Azure AI Foundry | Microsoft Learn

How to generate synthetic and simulated data for evaluation - Azure AI Foundry | Microsoft Learn

Responsible AI for Azure AI Foundry - Azure AI Foundry | Microsoft Learn

For example, expectations for customers:

Code of Conduct for Microsoft AI Services | Microsoft Learn

We recognize the need to support informed analysis and decisions by AI researchers, policymakers, and value chain actors as well as the public, and we encourage consideration of what information is needed, for whom, through what mechanism, at what altitude, and with what trade-offs—all in the context of a risk management outcome being advanced. Holistic and rigorous analysis, grounded in context from different communities about how information will be used in their respective processes to produce an intended effect, will help align prioritization of resources across actors in the value chain in pursuit of common goals—and minimize waste caused by mismatched goals and expectations.

Section 4 - Organizational governance, incident management and transparency

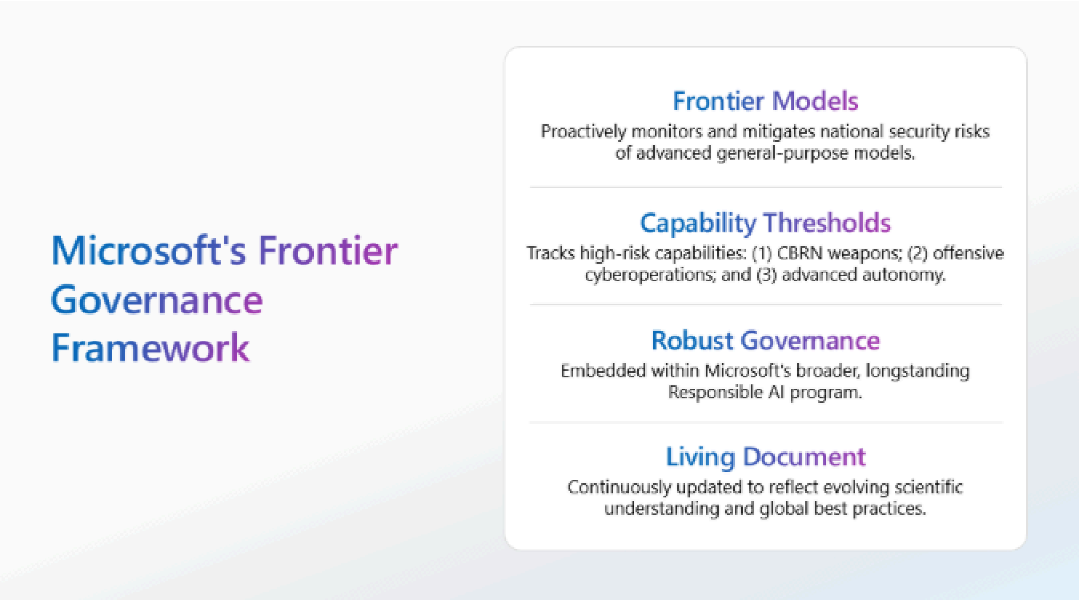
a. How has AI risk management been embedded in your organization governance framework? When and under what circumstances are policies updated?

At Microsoft, we are clear-eyed about the role we play in shaping this technology and in our understanding that people do not use technology they cannot trust. For us, the daily work of earning trust in the age of AI requires keeping humans at the center of how we design, develop, deploy AI – a practice that started in 2016 with the first draft of our AI principles. Formally adopted in 2018, our AI principles of fairness, transparency, accountability, reliability and safety, privacy and security and inclusiveness continue to serve as our enduring North Star.

The bedrock of our governance work, and our AI work at large, is our Responsible AI Standard. The Standard, which we first developed in 2019 and later revamped and released publicly in 2022, serves as our internal playbook for building AI systems in alignment with our AI principles.

In 2023, we formalized a set of specific internal requirements for generative AI systems to help us navigate the novel risks they presented. In 2024, we continued to update and improve these requirements, including establishing new policies for model development and deployment as part of our proactive, layered approach to compliance with new regulatory requirements, including the European Union’s AI Act. Cross-functional working groups worked together to identify key requirements to help our Microsoft teams get ready for enforcement deadlines and to support our customers with their own compliance efforts.

Our new policies for model development and deployment include our Frontier Governance Framework, which we shared with the public in February 2025. This framework originated from the voluntary Frontier AI Safety Commitments that Microsoft and fifteen other AI organizations made in May 2024 with the support of governments from around the world. The framework serves as a monitoring function, tracking the emergence of new and advanced AI model capabilities that could be misused to threaten national security or pose at-scale public safety risks. It also sets out a process for assessing and mitigating these risks so that frontier AI models can be deployed in a secure and trustworthy way.



Microsoft's Frontier Governance Framework

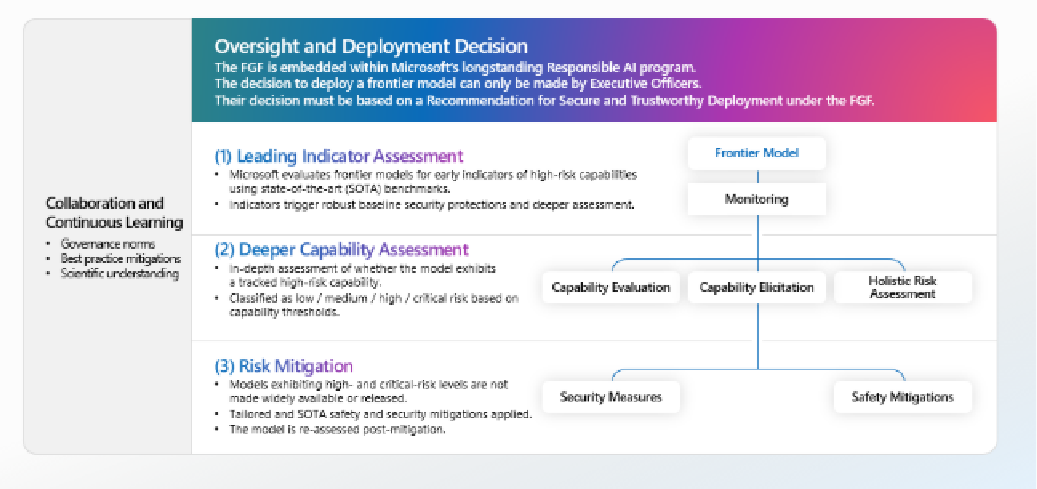
Frontier Models
Proactively monitors and mitigates national security risks of advanced general-purpose models.

Capability Thresholds
Tracks high-risk capabilities: (1) CBRN weapons; (2) offensive cyberoperations; and (3) advanced autonomy.

Robust Governance
Embedded within Microsoft's broader, longstanding Responsible AI program.

Living Document
Continuously updated to reflect evolving scientific understanding and global best practices.

Our Frontier Governance Framework integrates with our broader AI governance program by drawing on best practices for risk assessment, testing, and safety and security mitigations. We expect this framework to be updated over time as our understanding of AI risk and risk mitigation techniques improves and we look forward to working with others to continuously improve it.



Oversight and Deployment Decision
The FGF is embedded within Microsoft's longstanding Responsible AI program.
The decision to deploy a frontier model can only be made by Executive Officers.
Their decision must be based on a Recommendation for Secure and Trustworthy Deployment under the FGF.

Collaboration and Continuous Learning

- Governance norms
- Best practice mitigations
- Scientific understanding

(1) Leading Indicator Assessment

- Microsoft evaluates frontier models for early indicators of high-risk capabilities using state-of-the-art (SOTA) benchmarks.
- Indicators trigger robust baseline security protections and deeper assessment.

(2) Deeper Capability Assessment

- In-depth assessment of whether the model exhibits a tracked high-risk capability.
- Classified as low / medium / high / critical risk based on capability thresholds.

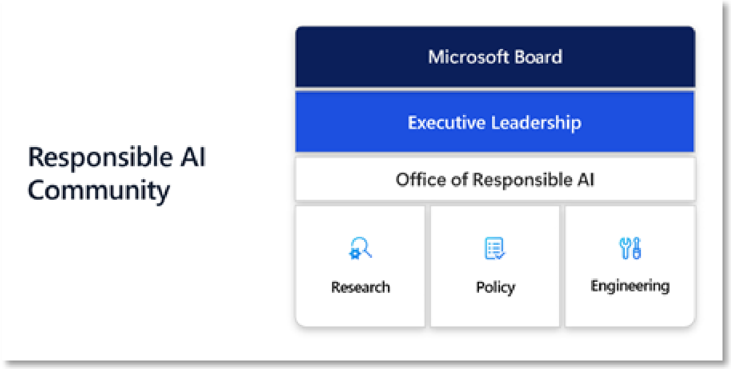
(3) Risk Mitigation

- Models exhibiting high- and critical-risk levels are not made widely available or released.
- Tailored and SOTA safety and security mitigations applied.
- The model is re-assessed post-mitigation.

b. Are relevant staff trained on your organization's governance policies and risk management practices? If so, how?

While establishing principles and policies are a critical building block to our responsible AI program, it takes a broad, cross-company effort to bring Microsoft's governance framework to life.

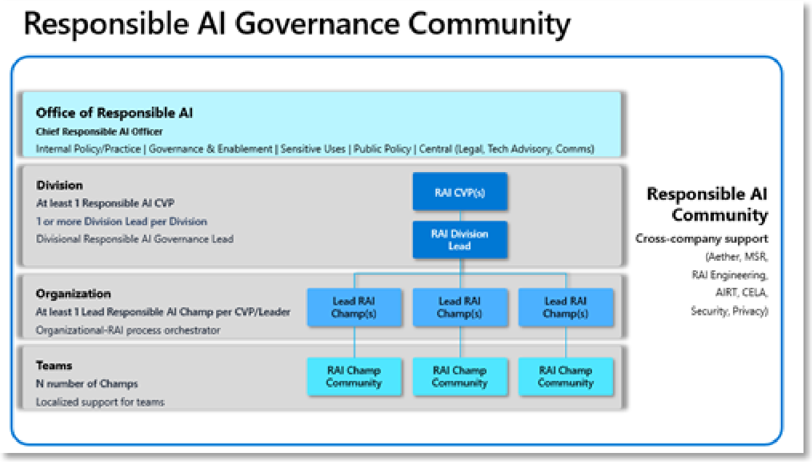
At the executive level, our adherence to our Responsible AI principles starts with our CEO Satya Nadella and our Responsible AI Council, led by Vice Chair and President Brad Smith and Chief Technology Officer Kevin Scott. Quarterly Responsible AI Council meetings provide executive-level oversight into the company-wide progress we are making on our commitments. Our progress is also regularly reported to, and guidance is solicited from, the Microsoft Board of Directors through the Board's Environmental, Social, and Public Policy (ESPP) committee.



Orchestrating these activities requires the muscle of a broader Responsible AI governance community. At Microsoft, this includes the Office of Responsible AI and a dedicated network of Responsible AI leaders and champions embedded throughout divisions across the company. Our Office of Responsible AI advises teams across the company on legal and regulatory requirements and manages the Responsible AI Governance framework and community, defining roles and responsibilities, creating documented processes, and leading oversight processes such as the Sensitive Uses and Emerging Technologies program.

Over the years, our Responsible AI Governance Community has matured, creating more specialized roles within each division at Microsoft. Since the launch of our Responsible AI Champions program in 2020, our Responsible AI Governance Community has grown to include Responsible AI Corporate Vice-Presidents (CVPs), Division Leads, and Lead Responsible AI Champs.

Responsible AI CVPs are accountable executives who provide oversight of their group's implementation of and adherence to RAI policies and serve on the Responsible AI Council. They are kept informed of progress by a Division Lead who drives operations and implementation of our processes to uphold customer trust in Microsoft's AI-powered products and services. Division Leads partner with Lead Responsible AI Champs to keep their teams informed of updates, implement procedures, and ensure adherence to our policies.



Responsible AI Governance Community

Office of Responsible AI
Chief Responsible AI Officer
Internal Policy/Practice | Governance & Enablement | Sensitive Uses | Public Policy | Central (Legal, Tech Advisory, Comms)

Division
At least 1 Responsible AI CVP
1 or more Division Lead per Division
Divisional Responsible AI Governance Lead

RAI CVP(s)
RAI Division Lead

Organization
At least 1 Lead Responsible AI Champ per CVP/Leader
Organizational-RAI process orchestrator

Lead RAI Champ(s)
Lead RAI Champ(s)
Lead RAI Champ(s)

Teams
N number of Champs
Localized support for teams

RAI Champ Community
RAI Champ Community
RAI Champ Community

Responsible AI Community
Cross-company support
(Aether, MSR, RAI Engineering, AIRT, CELA, Security, Privacy)

We continue to invest in growing, training, and cultivating a thriving and empowered Responsible AI community, which includes the Responsible AI governance community and other teams like RAI engineering, Aether, Microsoft Research, and the AI Red Team (AIRT) that carry out critical functions in the mapping, measurement, and management of AI risks.

The Responsible AI Community is provided with in-depth, ongoing training to best equip them to implement responsible AI practices within their teams and divisions. In 2024, Responsible AI community members participated in a variety of trainings that cover responsible AI policies, procedures, and tools. Training included updates to RAI policies as we prepared to implement new regulatory requirements, new guidance and tooling to support teams in measuring and mitigating risks, and specialized topics at the intersection of AI and security.

Throughout 2024, we continued to offer all Microsoft employees training on responsible AI and AI more broadly that caters to different technical knowledge levels and the contexts in which they develop or use AI. This includes both live training and self-paced training. At the broadest level, all Microsoft employees are required to complete the Trust Code (Standards of Business Conduct), our companywide ethics course, which includes training on Responsible AI. As of January 22, 2025, 99 percent of all employees had completed this course.

In addition to this training, employees also have the option to participate in hackathons and learning series focused on RAI. Throughout 2024, there were seven learning sessions focused on RAI hosted as part of the AI/ML Learning Series, which features insights from research and practice on AI. Cumulatively, these RAI-focused sessions had 6,798 attendees. In late 2024, an RAI-focused hybrid event featured 22 deep-dive sessions on insights and best practices related to RAI. This event had

1,020 unique attendees at the live session and the content was posted online for employees across the company to view at their own pace.

For more hands-on learning experience, Microsoft's annual Hackathon offers employees an opportunity to step out of their day-to-day work and team up with colleagues from across the company to build something innovative. The 2024 Hackathon had 738 hacks on AI that included a focus on responsible AI.

Since the publication of the 2024 transparency report, Microsoft’s Responsible AI Community has grown by 33% to over 540 employees, with over half dedicated to responsible AI full-time. During this time, full-time personnel focused on responsible AI has grown by 36.7%, with increases in investment across engineering, research, policy, and customer engagement organizations.

Within the Office of Responsible AI (ORA), our Sensitive Uses and Emerging Technologies program continues to serve as Microsoft’s review and oversight process of high-impact and higher-risk uses of AI. Through this process, we provide guidance for all types of AI systems, generative or otherwise, developed or deployed by Microsoft, whose foreseeable use or misuse meets one of three reporting criteria:

- Systems that could have a consequential impact on an individual’s legal status or life opportunities, e.g., court case prioritization systems or certain employment-related use cases.
- Systems that could present the risk of significant physical or psychological injury, e.g., systems designed to detect health conditions or workplace safety alert systems in industrial environments.
- Systems that could restrict, infringe upon, or undermine the ability to realize an individual’s human rights, e.g., computer vision-based systems deployed in public spaces that could restrict rights to assembly or synthetic media systems that could generate disinformation without guardrails.

In 2024, employees submitted 396 cases to our Sensitive Uses and Emerging Technologies team for consultation and responsible AI guidance, 77% of which were related to generative AI. As teams introduced new generative models with powerful capabilities into their products and services, they recognized these models came with an expanded risk surface area that needed to be managed; a trend that contributed to the increase in cases being submitted to the Sensitive Uses and Emerging Technologies team for review and guidance. While we saw more Sensitive Use cases in 2024 than in 2023, we realize that trend may not continue over time, nor do we view increasing case numbers as a sign of success in and of itself. As we continue to iterate on our Sensitive Uses process and as our Responsible AI work continues to mature, we expect that year-over-year case numbers may eventually remain stable or even decrease. Our end goal is to make AI models safer and more trustworthy, and our Sensitive Uses review process plays a key role in meeting this objective.

The Sensitive Uses review process is designed to provide responsible AI consultation tailored to specific products or use case scenarios. It culminates in requirements that apply and often go beyond the baseline, generalized requirements outlined in the RAI Standard and related policies. The Sensitive Uses and Emerging Technologies team consists of a multidisciplinary set of experts who provide hands-on counseling for high-impact and higher-risk uses of AI, as well as research and guidance for emerging issues or novel AI technologies. The team includes members with backgrounds in engineering, cybersecurity, product management, public policy, international relations, user experience research, data science, social sciences, and law. The team's expertise is further augmented by professionals from across our research, policy, and engineering organizations with expertise in human rights, social science, privacy, and security, who lend their expertise on complex sociotechnical issues as part of the Sensitive Uses Panel.

Case Study: Smart Impression

One example of an AI system that triggered our Sensitive Uses review process is Smart Impression, an AI-powered feature within a suite of productivity tools for radiologists called PowerScribe One. In radiology reports, the “impression” of the report refers to the section that summarizes all clinically significant findings and recommendations to inform downstream patient care. Smart Impression uses a transformer-based language model to compose draft impressions based on a radiologist's findings. By enhancing the speed and accuracy of composing these impressions, Smart Impression can improve efficiency and reduce the risk of burnout among medical professionals.

The use of generative AI in a healthcare decision support tool warrants additional oversight per Microsoft's Responsible AI policies and processes. In this case, the team went through the Sensitive Uses and Emerging Technologies review process, where they received hands-on consultation. Through the review process, the product team identified the possibility of generating impressions that don't align with findings included in the report—also known as ungrounded content—as a key risk to address. To measure and evaluate the risk of generating ungrounded content, the product team conducted both automated evaluations and human evaluations for each product release cycle, relying on practicing radiologists for their expertise.

After conducting red teaming, the team developed a method to measure how ungrounded content could appear in AI-generated impressions. Their measurement approach tracked various types of discrepancies, such as the tool mentioning a finding that wasn't in the reference impression or report, omitting a finding that was in the reference impression, and omitting the anatomic location or position of a finding present in the reference impression, among other variations.

The product team put in place several layers of mitigations to align with the requirements issued by the Sensitive Uses and Emerging Technologies team, including:

-

- **Off by default**: Smart Impression feature is off by default, meaning it cannot generate content without being invoked by the user and, if invoked, its final output can be discarded or edited. The radiologist has the option to write their own impression or invoke Smart Impression to generate draft impressions via a button push or voice command.
- **Human review**: If the Smart Impression feature is invoked by the user to generate draft impressions, a message is displayed to indicate that there is AI-generated content that needs to be reviewed, which also triggers an implicit workflow for review and approval. The draft impression cannot be submitted into a patient's medical record without review and signoff by the radiologist.
- **Feedback mechanism**: Radiologists can submit concerns or feedback about their experience with the tool through the user feedback feature, which is used to make ongoing improvements to the Smart Impression feature.
- **Transparency**: The team developed transparency documentation to equip radiologists with the information they need to use the Smart Impression feature responsibly. The Transparency documentation follows Microsoft's Transparency Note template, which includes information about the AI-powered feature's capabilities, limitations, intended uses, and best practices for use. The Transparency Note for Smart Impression is made available to Smart Impression customers and can be requested by prospective customers.

In line with Microsoft's staged release approach, Smart Impression was released for private preview to a small set of practicing radiologists to gather real-world insights ahead of broader release efforts. The private preview included 40,000 radiology reports generated by over 100 radiologists from community care hospitals, private practices, and university hospitals in the US. In two rounds of focused interviews as well as three site visits, radiologists relayed that they found the results to be acceptable for use in real radiology interpretation scenarios and found the draft impressions to be helpful both in terms of time savings as well as reducing cognitive load.

Before public preview, the product team used insights from interviews, feedback submitted by radiologists through the user feedback feature, and analysis of generated impressions that were accepted as-is or with minimal edits to make changes and improve performance of the model. By the end of public preview, half of the AI-generated impressions were accepted as-is, and almost three quarters were accepted with only minor edits by practicing radiologists. Smart Impression is now available to radiologists across the US, and the team continues to monitor and address issues to improve outcomes for healthcare professionals and their patients.

Case Study: Consumer Copilot with Voice

Another example of an AI system that triggered our Sensitive Uses and Emerging Technologies review process in the past year was Copilot Voice. As part of Microsoft's broader consumer Copilot service relaunch in the fall of 2024, the team building the system sought to leverage the audio capabilities of GPT-4o to provide a more natural conversational experience known as "Copilot Voice."

Deploying an AI system with new modalities such as voice and audio capabilities introduce novel risks that warrant additional oversight. The Copilot Voice product team went through Sensitive Uses review, where they received hands-on consultation from the Sensitive Uses team. Through the review process, the team identified, measured, and mitigated risks associated with audio generation, as well as user voice inputs, before deploying Copilot Voice.

The Sensitive Uses and Emerging Technologies team coordinated extensive red teaming conducted by both internal and external red teams. The red teams probed both the underlying GPT-4o model powering the Copilot Voice experience and the user-facing application with and without additional safeguards applied. The intention of these focused exercises was to produce harmful responses, surface potential avenues for misuse, and identify capabilities and limitations related to voice and audio scenarios. This process helped the product team better understand how the system could be used by a variety of users and helped improve mitigations.

Taking signals from red teaming, the product team conducted further evaluations through partially automated measurement pipelines. The Sensitive Uses and Emerging Technologies team worked closely with the product team to design broad coverage of risk areas with a focus on voice and audio-enabled risks. This included risks related to the system's ability to generate third-party content, reproduce or mimic a person's voice, succumb to voice-based jailbreaking techniques, and more.

The measurement pipelines included a set of conversations, or interactions, with Copilot Voice collected from human evaluators and synthetic conversations generated with LLMs prompted to test policies in an adversarial fashion. Each of the newly developed conversation sets were annotated by human labelers who read the text content or listened to the audio output to validate the LLM-based evaluations.

These measurement efforts informed the development of a range of mitigations, including post-training, system prompts, and both input and output classifiers. For example, to mitigate the risk of the system reproducing or mimicking a person's voice, the product team implemented a mechanism to assess and block voice outputs from the system that diverge significantly from the set of voices the user can choose from.

Finally, the Sensitive Uses and Emerging Technologies team helped develop and review transparency documentation designed to equip users with information they need to use the product responsibly. The transparency documentation follows Microsoft's Transparency Note template, which includes information about the AI-powered feature's capabilities, limitations, intended uses, and best practices for use.

By design, cases submitted to the Sensitive Uses and Emerging Technologies team tend to be applications of AI with complex risk profiles. This means that cases are often fact-intensive and require bespoke engagement and guidance. However, we have observed several common trends in Sensitive Use cases over the past 12 months:

1. **Use of generative AI has continued to accelerate, with emphasis on multimodal and agentic applications.** Almost 80% of Sensitive Use cases in the past year involved leveraging a generative model, deployments which increasingly have image- and audio-based capabilities. More recently, many of these cases have also involved agentic applications of AI, such as generative orchestration, which involves coordination of information flows between user input across one or more models and across additional tools and sources of data.
2. **Health & Life Science projects make up an increasing share of cases submitted.** While the Sensitive Uses and Emerging Technologies team reviews cases across almost every industry, we receive a higher volume of health & life science submissions compared to any other domain. These projects typically involve administrative and decision-support tools for healthcare professionals and educational tools for patients. These deployment scenarios often meet two of the Sensitive Use and Emerging Technologies team's reporting criteria: Systems that could have a consequential impact on an individual's legal status or life opportunities (e.g., access to healthcare services), and systems that could present the risk of significant physical or psychological injury (e.g., systems designed to detect health conditions). Depending on their use cases, reviews for these features often evaluate them for accuracy and develop strategies to reduce risks from overreliance.
3. **AI for scientific research is an emerging trend among cases submitted.** Microsoft's research and development efforts in the sciences are driven by experts in quantum physics, computational chemistry, molecular biology, software engineering, and other disciplines. Their work increasingly makes use of AI innovations, with teams like Microsoft's AI for Science research organization. Some of these research advances include projects such as MatterGen, which leverages AI to accelerate our ability to discover new molecules for drug discovery or new materials in broad domains including batteries, magnets, and fuel cells. Given the many potential downstream applications of these research advances, the Sensitive Use reviews typically focus on developing a detailed threat model, which is used to build customized research safeguards and deployment mitigations.

In addition to the case consultation function, our Sensitive Uses and Emerging Technologies team develops early guidance for emerging AI technologies and risks. By identifying signals from Sensitive Use cases, partnering with researchers and product strategists across Microsoft, and scanning the horizon of AI innovations, the team creates early-steer guidance ahead of broader, more formal policy development.

This provides engineering teams with actionable recommendations when building with novel and emerging AI technologies. In 2024, this included drafting guidance for audio-based systems, image editing, and agentic AI systems. Emerging technologies guidance also serves as a foundational resource for developing more formal internal policies and helps inform our public policy efforts around new AI models and applications.

c. Does your organization communicate its risk management policies and practices with users and/or the public? If so, how?

Yes, we provide transparency into our risk management policies and practices by making available resources about our policies and commitments, such as our Responsible AI Standard, AI governance program information (via our Responsible AI Transparency Report), Security Development Lifecycle, and Secure Future Initiative.

d. Are steps taken to address reported incidents documented and maintained internally? If so, how?

Yes, steps are taken to address reported incidents; the incident detection and response work that happens after the release of an AI product deserves just as much attention and planning as the work that happens leading up to the release. At the conclusion of each incident a post incident review is conducted to understand the nature of the incident and track repair items that need to be addressed. This process divides the repair items by impact severity which then places a short, medium, or long-term time frame to complete the tasks. Short- and medium-term repair items are tracked by a dedicated team through their closure. Long term repair items are inserted into the normal engineering workflows for future releases. Post Incident Reviews are presented to key executive leaders including the relevant deputy Chief Information Security Officer (CISO) staff.

We also invest in crisis management to drive more consistency and efficiency in how we detect and respond to AI issues and incidents. When a crisis occurs, teams benefit from the expanded capacity of specialized roles like crisis managers, forensic investigators, and communications managers. As part of the service release cycle, we work with service teams to ensure they are connected to these central processes. Additionally, Microsoft has built a robust process to reflect on our handling of incidents, track repair items, and synthesize themes for use in future engineering releases. These experiences also inform our AI policies and practices, including the AI bug bar that we use to triage concerns raised to MSRC.

We also draw insights from externally reported issues. When an issue is confirmed in an AI service, we open companion cases for other services using similar models so that they may evaluate any impact. Our team also looks broadly across reported issues to synthesize themes, which are then used to educate product and service teams, update engineering processes to catch them earlier, and update policies as applicable.

Security research and incidents provide valuable insights into how we can engineer and operate our AI services better. In 2024, we formalized a process that brings those insights together to educate our engineering teams and update our engineering processes. We draw insights from every reported issue, each incident we experience, and by observing other’s experience in the threat landscape, including nation-state actor analyses reported by Microsoft’s Threat Analysis Center.

We continuously improve our insights to handle security incidents, as described in Section 4C. Below we highlight some key insights and learnings we gained through incident response efforts in 2024:

- No incident-level events in 2024 were a result of AI system malfunctions or issues arising during benign use. Every incident included patterns of malicious use where actors were actively trying to bypass security measures or misuse Microsoft AI products or services.

- Threat actors identified in 2024 exhibited varying levels of sophistication, ranging from actors who worked in isolation to individuals who coordinated across a network of actors working towards the same goal.
- Threat actors often exploit differences in safety systems across different AI products and services, making it increasingly important to share the latest and greatest innovations in AI safety across the tech industry and with customers.
- Threat actors are not only working to circumvent safety systems built around generative AI systems and models, but they are also using generative AI as a tool, as they would PowerShell or mimikatz. Defenders are best equipped to think of AI as another tool in attackers' toolbox.

e. How does your organization share relevant information about vulnerabilities, incidents, emerging risks, and misuse with others?

Cross-industry sharing of threat and vulnerability information will help make AI more secure and trustworthy, accelerating adoption across sectors. As highlighted in Section 2G, as a member of the Frontier Model Forum (FMF), Microsoft has recently agreed to facilitate information sharing about threats, vulnerabilities, and capability advances unique to AI, and we also share information about security vulnerabilities via our Microsoft Active Protections Program (MAPP). When required, Microsoft discloses incidents to appropriate government agencies. When appropriate, Microsoft shares with other AI model providers common threats and research.

f. Does your organization share information, as appropriate, with relevant other stakeholders regarding advanced AI system incidents? If so, how? Does your organization share and report incident-related information publicly?

Beyond our processes and programs described in other sections, including Section 4E and 2G, Microsoft shares information with other relevant stakeholders, including to take legal action against cybercriminals misusing AI.

For example, in January 2025, a complaint unsealed in the Eastern District of Virginia revealed that Microsoft's Digital Crimes Unit observed that an international threat–actor group had developed sophisticated software to exploit exposed customer credentials scraped from public websites. In February, Microsoft filed an amended complaint that named four of the primary developers of these malicious tools. These individuals, part of a global cybercrime network, exploited exposed customer credentials scraped from public sources to unlawfully access accounts with certain generative AI services. They then altered the capabilities of these services and resold access to other malicious actors, providing detailed instructions on how to generate harmful and illicit content.

Upon our discovery of this activity, Microsoft revoked cybercriminal access, put in place countermeasures, and enhanced our safeguards to further block such malicious activity in the future. The court order also enabled us to seize a website instrumental to the criminal operation so that we could gather crucial evidence about the individuals behind these operations, decipher how these services were being monetized, and disrupt additional technical infrastructure we found.

Seizing this infrastructure allowed us to effectively disrupt a cybercriminal network and create a powerful deterrent impact among its members. We take the misuse of AI very seriously and remain committed to protecting users by embedding robust AI guardrails and safeguarding our services from illegal and harmful content.

When identified risks and vulnerabilities affect the broader ecosystem of AI technologies and products, we also take action to support or collaborate with external stakeholders. For example, as we have with the Crescendo and Skeleton Key attacks, we share information publicly about emergent risks and mitigations. When products of other organizations are affected by risks or vulnerabilities that Microsoft identifies, we also share information directly with impacted organizations, consistent with the principles and guidelines of coordinated vulnerability disclosure.

g. How does your organization share research and best practices on addressing or managing risk?

Microsoft Research publishes extensively to share its findings with others, and subject matter experts from engineering teams also publish findings on emerging risks. We also use the following channels to convey and publish best practices:

1. Microsoft Security Blog: Microsoft regularly publishes articles on AI security risk management, sharing frameworks and best practices to help organizations secure their AI systems
2. Microsoft Purview Insider Risk Management: This tool helps organizations identify and respond to risky AI usage, integrating insider risk context into security operations
3. Microsoft Learn: Microsoft provides resources and frameworks for AI risk assessment, enabling organizations to conduct security risk assessments and improve their AI systems' security posture.

In 2024, the Sociotechnical Alignment Center (STAC), a team of researchers, applied scientists, and linguists in Microsoft Research, collaborated with researchers in the AI & Society Fellows Program, which is discussed in more detail later in this report, to publish a paper that contributes a measurement framework grounded in measurement practices that emerged in the social sciences. The researchers argue that, unlike more narrowly scoped measurement tasks involved in supervised machine learning systems, measurement tasks for generative AI systems, including risks and capabilities, often require measuring more complex, nuanced, and contested concepts.

STAC draws on measurement approaches in the social sciences that often grapple with similarly complex social concepts, to propose a measurement framework for generative AI. This framework emphasizes the critical step of systematizing, or clearly defining, complex and nuanced concepts, amounts, populations, and instances before jumping straight into operationalizing measurements. This provides clarity on what is being measured, enables stakeholders to better understand, interrogate, and compare measurements, and inform how effective risk mitigations are designed.

In 2024, Microsoft Research’s AI Frontiers lab worked on identifying and addressing some of the prevailing challenges plaguing current AI benchmark evaluation practices, including benchmark saturation and lack of transparency in evaluation methods. To address these challenges and meet the need for more rigorous and nuanced evaluation of large foundation models, AI Frontiers developed Eureka, a reusable and open evaluation framework that aims to create transparency and reproducibility while standardizing evaluations of large foundation models. We also released EUREKA-BENCH, a collection of benchmarks that state-of-the-art foundation models still find challenging to meet. These benchmarks represent fundamental but overlooked capabilities for completing tasks in both language and vision modalities.

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

Across jurisdictions, horizontal and issue-specific AI laws, norms, and standards are advancing at the same time. Horizontal approaches, such as the European Union's AI Act, address multiple layers of the tech stack, multiple sectors, and multiple issues, while narrower approaches focus on specific governance measures or topics, like synthetic media or frontier model safety.

Microsoft engages in global efforts to build consensus-based frameworks, promoting coherence across borders while instilling clear allocation of responsibilities across the AI value chain. We also use international technical standards and best practices for AI risk management and governance, including the NIST Artificial Intelligence Risk Management Framework (AI RMF), the NIST AI RMF Playbook, and ISO/IEC 42001, for which we recently achieved certification for Microsoft 365 Copilot and Microsoft 365 Copilot Chat. These frameworks help integrate AI risk management into broader risk management strategies, ensuring a cohesive approach to handling AI, cybersecurity, and privacy risks. Additionally, Microsoft's AI governance policies are designed to align AI activities with ethical standards, regulatory requirements, and business objectives. This includes using tools like Azure Policy and Defender for Cloud to manage AI models and detect risks.

Previous waves of technology have demonstrated that there are two components to the trust that underpins broad adoption and iterative innovation: first, trust in how technology itself performs; and second, confidence that people and organizations can deploy it successfully. In order to continue to advance trust and confidence in AI, we must work towards globally coherent governance frameworks that can help accelerate adoption and allow organizations of all kinds to innovate and use AI across borders. Microsoft will continue to share the lessons learned from its internal governance work so that others may build on it, focusing on three key areas:

- **Strengthening feedback loops between innovation and governance**. We know even greater AI capability is on the horizon, with the opportunity to unlock innovation in science, education, and countless other fields. AI deployment, experimentation, and skilling must go hand-in-hand with AI governance to create tighter feedback loops on what is effective in practice. As we've learned through implementing Microsoft's Responsible AI Standard, while minimum guardrails provide an important starting point, we can learn much more about how to govern AI technology effectively in practice through AI deployment and experimentation. Moreover, as we've learned through establishing Microsoft's responsible AI program and developing role-based training, good governance includes investing in people so that they can take advantage of the capabilities that AI already demonstrates and strengthen their readiness for the AI capabilities that may emerge. If we get this technology into the hands of more people who can apply it to the local challenges that they uniquely understand, then we will not only have much greater impact in realizing opportunities but also a much broader feedback loop on governance.
- **Advancing scientific understanding to inform effective guardrails and practice**. Over the past 18 months, global stakeholders have come together to make significant progress in defining approaches to governing AI development and deployment. As we continue to work towards

interoperable global governance, we are also turning our attention towards building a deeper shared understanding of effective and easy to adopt risk management techniques and technical practices that can help realize the high-level goals of these governance frameworks. At Microsoft, we are continuing to invest in our own internal governance frameworks, learning from their implementation while also participating in multistakeholder governance entities. Partnerships with government research bodies, such as AI institutes, and collaborating on consensus-driving publications like the International AI Safety Report can help us close identified evidence gaps and synthesize research and applied learnings. Frameworks for voluntary reporting on governance practices, such as the Hiroshima AI Process Reporting Framework, can also help to deepen, align, and streamline shared expectations across jurisdictions.

- **Aligning expectations for guardrail implementation.** Today's AI systems often involve multiple models and components from different providers. For organizations deploying these systems to have confidence in each component and the system as a whole, it's important to align expectations for guardrail implementation across different actors in the supply chain. Building on developing industry norms, as Microsoft has done through its Frontier Governance Framework to address national security risks of highly capable models, will help accelerate progress. Where expectations for the behavior of AI applications diverge among jurisdictions, focusing on where guardrails can align—such as on expectations for model-level transparency—will yield significant benefits for AI adoption and innovation.

To make further progress in advancing AI governance, we must develop ecosystem-wide reference points for effective guardrails as well as tools that support implementation for ourselves and our customers. Just as internal investments in cybersecurity practices and tools have enabled us to support the broader cybersecurity ecosystem, so too have our years building an AI governance program readied us to help others seize AI opportunities. We will continue to make such investments and work through organizations like the Frontier Model Forum and MLCommons to develop industry practice reference points and AI evaluation tools.

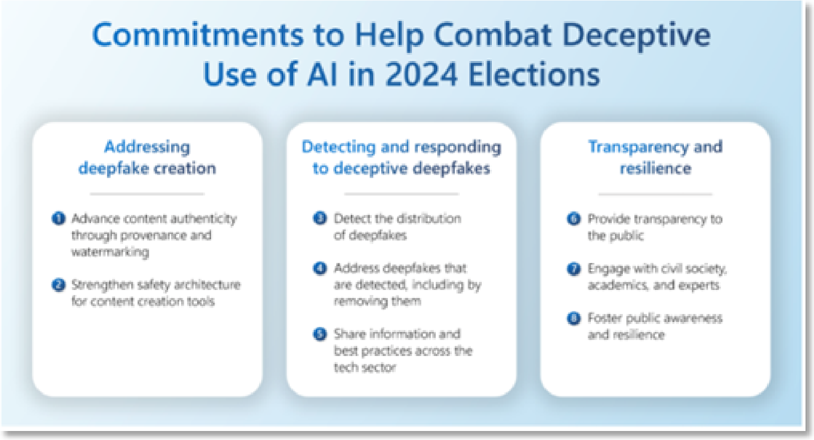
Section 5 - Content authentication & provenance mechanisms

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

As we continue to see advances with AI, the emergence of synthetic content—text, images, and videos generated by AI—and new uses of personal information, present both unprecedented opportunities and significant challenges. Synthetic content has the potential to revolutionize creative industries and enable new forms of artistic expression and innovation. In that context, our generative applications are designed to adhere to company policies, including our Responsible AI Standard's Transparency requirement to inform people that they are interacting with an AI system. We update these policies as needed, informed by regulatory developments and feedback from internal and external stakeholders.

When users interact with Copilot in Bing, we provide in-product disclosure to inform users that they are interacting with an AI application and citations to source material to help users verify information in the responses and learn more. Other important notices may include disclaimers about the potential for AI to make errors or produce unexpected content. See below for screenshot example of in-product notice.

b. Does your organization use content provenance detection, labeling or watermarking mechanisms that enable users to identify content generated by advanced AI systems? If yes, how? Does your organization use international technical standards or best practices when developing or implementing content provenance?



Commitments to Help Combat Deceptive Use of AI in 2024 Elections

Addressing deepfake creation

1. Advance content authenticity through provenance and watermarking
2. Strengthen safety architecture for content creation tools

Detecting and responding to deceptive deepfakes

3. Detect the distribution of deepfakes
4. Address deepfakes that are detected, including by removing them
5. Share information and best practices across the tech sector

Transparency and resilience

6. Provide transparency to the public
7. Engage with civil society, academics, and experts
8. Foster public awareness and resilience

In 2024, more people voted in elections across the world than at any other time in history. As a leading technology company whose products are used to create AI-generated content, we know that we have a responsibility to take steps to prevent the creation and dissemination of deceptive content. Although generative AI poses a considerable risk in spreading disinformation, it did not play a central role in undermining the integrity of the 2024 election as originally anticipated. This was due in part to proactive measures implemented by governments, nonprofit organizations, and private sector companies globally.

In February 2024, Microsoft joined twenty-six other leading technology companies in signing the Tech Accord to Combat Deceptive Use of AI in 2024 Elections. The Accord consisted of eight specific commitments designed to make it more difficult for bad actors to use legitimate tools to create deepfakes, to bring the tech sector together to detect and respond to deepfakes in elections, and to help advance transparency and build societal resilience to deepfakes in elections.

Microsoft's work on living up to the commitments included in the Tech Accord started long before 2024. Three years prior, Microsoft co-founded the Coalition for Content Provenance and Authenticity (C2PA) to develop an open technical standard for establishing the provenance—the source and history—of digital content, including AI-generated images, audio and video. By 2023, leveraging the C2PA standard, we were automatically embedding cryptographically sealed provenance metadata, also known as Content Credentials, to all content generated by DALL-E series of models in Azure OpenAI Service, Bing Image Creator, Microsoft Copilot, Microsoft Designer, and Microsoft Paint.

By March 2024, Microsoft piloted Content Integrity Tools, which allowed users to add content credentials to their own authentic content. Designed as a pilot program primarily to support the 2024 election cycle and gather feedback about Content Credentials-enabled tools, the tools were available to political campaigns in the EU and U.S., as well as to elections authorities and select news media organizations in the EU and globally. These tools included a partnership and collaboration with fellow Tech Accord signatory, TruePic. Announced in April 2024, this collaboration leveraged TruePic's mobile camera SDK enabling campaign, election, and media participants to capture authentic images, videos and audio directly from a vetted and secure device. Called the "Content Integrity Capture App" (an app that makes it easy to directly capture images with C2PA enabled signing) launched for both Android and Apple and can be used by participants in the Content Integrity Tools pilot program.

In September 2024, Microsoft rolled out a new built-in feature in Azure OpenAI Service to add watermarks to all images generated using DALL·E, our flagship generative AI image generator. These watermarks, which are invisible to the human eye and can be identified by specialized tools, provide an additional layer of robustness for disclosing AI-generated content. These watermarks are resilient to common modifications such as resizing or cropping, helping ensure that the integrity of the watermark remains intact even when images are altered in various ways. They can also help recover contextually rich provenance information found in Content Credentials in cases where a C2PA manifest might have been removed from the media.

To further these goals, we created specific product implementation guidance for teams across Microsoft building AI systems, including use of certified candidate lists to mitigate the risk of deceptive AI generated images of candidates. If users sought election critical information—defined as factual aspects of political elections that could be proven to be true or false (e.g., where to vote, when to vote, and election results)—Microsoft generative AI tools were instructed to either refuse to answer, or to provide a demonstrably reliable answer directing users to specific authoritative sources, most commonly the election authority for that election.

In February 2024, we created a site for candidates and election authorities to report election-related deceptive AI content like deepfakes appearing on Microsoft consumer services. Beyond responding to reported events, we solicit the support of the Microsoft Threat Analysis Center (MTAC) whose mission is to detect, assess, and disrupt foreign cyber enabled influence threats to Microsoft, its customers, and democracies worldwide. MTAC also partnered with the AI for Good team to develop technical capabilities to better detect deepfakes. In 2024, MTAC publicly published eight reports focused on nation-state actors and election interference. The intelligence gathered by MTAC helps us see a broader view of the adversary threat landscape and enables us to proactively combat deceptive AI-generated content.

While the 2024 elections are behind us, the threat that deceptive AI content could pose to elections around the world is far from over. At Microsoft, we recognize that we need to take a whole-of-society approach to address the risk of bad actors using AI and deepfakes to deceive the public. This is why we invest in open standards like C2PA and share the insights we gather through MTAC openly. It's why we also launched a $2 million fund with OpenAI to increase AI education among voters and vulnerable communities. Microsoft will continue to develop our technology and policies, as well as work with other stakeholders globally, to ensure that we uphold the foundational principle of free expression for citizens in the United States and around the world.

Section 6 - Research & investment to advance AI safety & mitigate societal risks

a. How does your organization advance research and investment related to the following: security, safety, bias and disinformation, fairness, explainability and interpretability, transparency, robustness, and/or trustworthiness of advanced AI systems?

Microsoft researchers work to advance the state of the science of responsible AI with the aim of enhancing our understanding of AI, creating new model architectures with novel capabilities, achieving societal benefit, transforming scientific discovery, and extending human capabilities. In 2023, Microsoft formed the AI & Society research network, which addresses the many bidirectional relationships between AI technologies and people, groups, organizations, and society as a whole.

The network aims to bring together a diverse and multidisciplinary community of researchers to explore and shape the social and technical aspects of AI. In addition to pursuing the aims above, the research network explores topics that include sociotechnical evaluation and alignment of AI and equitable AI.

In alignment with the efforts of this research network, Microsoft launched the Microsoft Research AI & Society Fellows program to catalyze research collaboration between Microsoft researchers and eminent scholars and experts across a range of disciplines at the intersection of AI and its impact on society. This investment has resulted in 13 distinct, ambitious research collaborations bringing together Microsoft researchers and 24 esteemed fellows across academic and industrial disciplines.These research collaborations address some of the pressing challenges facing organizations who aim to advance responsible AI practices, including research focused on advancing the science of AI risk measurement as described in the Measurement section earlier in this report.

Other research collaborations within the AI & Society Fellows program have a broader focus. For example, one of the research collaborations aims to address how to most effectively regulate AI in light of the challenges of doing responsible AI in practice, exploring the challenges faced by front-line practitioners working on a range of responsible AI issues, including fairness, privacy, security, and more. The collaboration aims to take stock of existing social scientific insights into the difficulties faced by regulated entities seeking to comply with existing and forthcoming regulations, complete original empirical studies to fill identified gaps in the existing scholarship on responsible AI in practice, and channel all of these findings into the ongoing debates about how to craft effective regulations of AI.

Microsoft also established the AI Frontiers lab in 2024 to invest in the core technologies that push the frontier of what AI systems can do in terms of capability, reliability, and efficiency. Researchers at our

AI Frontiers Lab are not only interested in how well these systems work—they also want to ensure that we build in sociotechnical solutions that can make these systems work in a responsible way.

Our research teams work in close collaboration with our policy and engineering teams to inform our approach to responsible AI. Throughout 2024, Microsoft researchers pushed the frontiers of our understanding of mapping, measuring, and managing AI risks. We summarize some of their research contributions below.

Advancing the science of mapping risks. Microsoft researchers conducted studies and developed new frameworks to identify emerging AI risks, dive deeper into known risks, and interrogate the tools we use to identify them. Advancements in generative AI systems have enabled the development of agentic AI systems that are capable of autonomously executing actions and collaborating with other agents to achieve user-specified goals. Building on prior research, researchers across Microsoft support our understanding of the emerging risks associated with the development and use of agentic AI systems, including the wide range of failure modes with human-agent communication. Microsoft researchers, in collaboration with a broader set of researchers across academia and the technology industry, contributed new approaches to advance existing responsible AI tools and practices such as impact assessments and AI documentation frameworks. This includes a study that explores the use of impact assessments by industry researchers and contributes 10 design considerations to facilitate the effective design, development, and adaptation of an impact assessment template for use in industry research settings and beyond. In May 2024, researchers introduced the CLeAR (Comparable, Legible, Actionable, and Robust) framework that aims to help practitioners consider the complexities and tradeoffs required when developing documentation for datasets, models, and AI systems throughout their lifecycle.

Researchers also probed the potential risks associated with the use of conversational AI systems for social and emotional support. Through this study, researchers have developed a taxonomy and framework that advance our understanding of AI behaviors, psychological impacts and the contexts in which these impacts may manifest. Using insights from their work, they recommend emphasizing AI interaction disclosure, specifically focusing on emphasizing the non-human nature of the system and when sensitive topics are discussed, enabling the system to gracefully redirect the user to accessible and actionable resources. These research insights are critical inputs to Microsoft's AI policy development and oversight efforts. For example, the Sensitive Uses and Emerging Technologies team collaborates closely with researchers studying the risks of using conversational AI systems for emotional support. These collaborations inform how the Sensitive Uses and Emerging Technologies team crafts early-steer guidance aimed at providing engineering teams with actionable recommendations when building conversational AI systems.

Advancing the science of measuring risks. Researchers across industry and academia are uniquely positioned to make meaningful contributions to advance the science of AI risk measurement. This would involve developing frameworks and tools to build reliable and repeatable measurement approaches that are adaptable across different types of AI models, systems, use cases and deployment scenarios. Within Microsoft Research, the Sociotechnical Alignment Center (STAC) continues to produce meaningful research and thought leadership in this area. Building on the four-part measurement framework discussed earlier in the report, the STAC team has published papers that extend this framework beyond concepts to include the systematization and operationalization of amounts, populations, and instances, and a set of general dimensions that capture critical choices involved in GenAI evaluation design. Lastly, STAC has published work to help bridge risk mapping and measurement by investigating whether red teaming can produce measurements that enable meaningful comparisons of systems, and how red teaming can evolve to more effectively support principled e valuations. These research contributions advance our ability to better understand, interrogate and compare different evaluations.

Other Microsoft Research teams made additional contributions that advanced the science of AI risk evaluation in 2024, ranging from studies that focus on the use of synthetic data, including in AI evaluation tasks and studies that focus on improving benchmark evaluation practices. The AI Frontiers lab has provided meaningful thought leadership and scalable tools to fill existing gaps in benchmark evaluation practices across the industry, which we highlighted earlier in this report. Taking a closer look at the potential risks of enabling AI systems with code execution capabilities, the team published a paper and open-sourced an evaluation platform with benchmarks intended to provide practical evaluation tools to assess the safety of AI systems that can execute code.

Advancing the science of managing risks. Researchers across Microsoft continue to explore novel strategies to manage the risks associated with the use and misuse of AI. These range from mechanisms to protect against indirect prompt injection attacks; reduce the risk of overreliance on AI outputs; and develop novel approaches to steer model behavior. Identifying novel strategies to defend against indirect prompt injection attacks (XPIA), an attack mechanism where threat actors embed hidden malicious instructions in a grounding data source to circumvent safety guardrails, continues to be an area of interest for both practice and research. In 2025, Microsoft researchers, in collaboration with researchers in academia, developed a benchmark to assess the risk of XPIA vulnerabilities in LLMs, now available on GitHub. In their research, they also identified promising defense mechanisms to defend against XPIA, including boundary awareness to help LLMs differentiate between user prompts and external content, and explicit reminders for the LLM not to execute instructions embedded in external content.

Exploring ways to reduce inappropriate reliance on AI-generated output continued to be a theme in research throughout 2024. Researchers across Microsoft contributed to a growing body of work on fostering appropriate reliance. One study explored the impact of LLM's uncertainty expression, comparing first-person with general perspective. Another paper explored the risk of overreliance on generative AI and identified emerging mitigation techniques such as uncertainty highlighting, cognitive forcing functions, contrastive explanations, and AI critiques. Researchers also explored how to steer model behavior towards safer and more reliable outputs while preserving model performance. In one study, researchers explored inference time interventions where they first identify features that mediate refusals and assess whether amplifying these features improves robustness to challenging multi-turn jailbreak attacks while preserving model performance. Another body of work examines the relationship between AI model's risks and values, offers an evaluation framework that is designed for the and proposes various alignment strategies that offer alternatives to traditional AI alignment efforts such as reinforcement learning using human feedback (RLHF).

Frontier research on mitigating risks of abusive AI-generated content

Microsoft AI technology is built with safeguards to help prevent malicious actors from creating and disseminating abusive AI-generated content, including deepfakes, intended to defraud, abuse or manipulate people. As part of our efforts to continually iterate upon and improve these safeguards, our Aether Committee is leading a study to explore the effectiveness of media integrity technologies used to disclose synthetic image, video, and audio content. This includes exploring the rough edges of the technologies and the opportunities to boost resilience when it comes to technical attacks and sociotechnical robustness. Microsoft's AI for Good team is conducting complementary research on synthetic image and video detection capabilities, the public's ability to identify AI-generated content, and the robustness of detection tools.

We're also engaging in a number of ongoing efforts to improve our understanding of existing and emergent harms and thereby develop more effective mitigations. For example, Microsoft's Digital Crimes Unit is co-leading a project with the US Secret Service and the German Federal Criminal Police, funded by Europol, to evaluate and address the threat caused by cybercriminals' misuse of AI services, including synthetic media and fake content. Microsoft's Digital Safety Unit is supporting research through the Tech Coalition's Safe Online Research Fund in collaboration with the End Violence Against Children Partnership to advance our understanding of the patterns of online child sexual abuse and exploitation, including how generative AI can be abused. Microsoft's Threat Analysis Center is also researching the use and reach of AI-generated media in influence operations by nation-state actors.

As threats continue to evolve, we will continue researching and implementing mitigations that keep pace with new technology and keep our users protected.

Optimizing prompts through PromptWizard

Large language models (LLMs) rely on prompts, carefully crafted inputs that guide them to produce relevant and meaningful outputs. Creating prompts that can help with complex tasks is a time-intensive and expertise-heavy process that often involves months of trial and error. To address this challenge, we developed PromptWizard (PW), a research framework that automates and streamlines the process of prompt optimization. In December 2024, we released an open-source version of the PromptWizard codebase to foster collaboration and innovation within the research and development community.

PromptWizard (PW) is designed to automate and simplify prompt optimization. It combines iterative feedback from LLMs with efficient exploration and refinement techniques to create highly effective prompts within minutes by optimizing both the instruction and the in-context learning examples. Central to PW is its self-evolving and self-adaptive mechanism, where the LLM iteratively generates, critiques, and refines prompts and examples in tandem. This process ensures continuous improvement through feedback and synthesis, achieving a holistic optimization tailored to the specific task at hand. By evolving both instructions and examples simultaneously, PW ensures significant gains in task performance.

After conducting rigorous evaluation of PromptWizard on over 45 tasks, spanning both general and domain-specific challenges, our researchers found that PW consistently outperformed competitors in accuracy, efficiency, and adaptability. PW excels in conditions where training data is limited, requiring as few as five examples to produce effective prompts. Across five diverse datasets, PW demonstrated an average accuracy drop of only 5%, from 87% to 81.9%, when using five examples compared to 25 examples. PromptWizard also reduces computational costs by using smaller LLMs for prompt generation, reserving more powerful models for inference.

Advancing AI research beyond Microsoft. In addition to conducting our own research, we also support academic research that might otherwise lag behind research conducted by private companies due to lack of resources. For example, in January 2024, we announced our support of the National AI Research Resource (NAIRR) pilot led by the National Science Foundation, which provides high-quality data, computational resources, and educational support to make cutting-edge AI research possible for more U.S. academic institutions and non-profits. Microsoft has committed $20 million worth of Azure compute credits to support researchers with high-performance computing resources and access to leading-edge models. Our commitment to the NAIRR pilot also includes collaborative opportunities with Microsoft's scientists and engineers and resources to accelerate domain-specific research such as innovative tools for chemistry and materials science research via Azure Quantum Elements, and tools for research and development on AI fairness, accuracy, reliability, and interpretability.

To date, we have provided 38 grants for researchers across the US to access critical resources for AI research and development on Azure. These grants are in support of support of research projects at both academic institutions and non-profit organizations focused on molecular biology and protein design, healthcare and drug discovery, sustainability and earth sciences, personalized education and accessibility, agriculture, human-AI collaboration, and new approaches for AI privacy, safety, and security.

Accelerating Foundation Models Research

Foundation models continue to fuel a fundamental shift in computing research, natural sciences, social sciences, and computing education itself. The Accelerating Foundation Models Research (AFMR) initiative was created by Microsoft Research to work together with the broader academic research community to enable AI advances and nurture a vibrant and diverse AI research ecosystem by providing access to state-of-the-art foundation models hosted on Microsoft Azure through Microsoft Azure AI services. The goal of AFMR is to foster more collaborations across disciplines, institutions, and sectors to unleash the full potential of AI for a wide range of research questions, applications, and societal contexts.

To date, the AFMR research community has published over 300 papers co-authored by computer scientists and researchers outside computer science supporting over 123 institutions in 19 countries. This depth and breadth of expertise across disciplines, cultures, and languages has contributed meaningfully to our ability to use AI to address some of the world's greatest challenges around the following three goals:

- **Aligning AI with shared human goals, values, and preferences via research on models** to enhance safety, robustness, sustainability, responsibility, and transparency, while also exploring new evaluation methods to measure the rapidly growing capabilities of new models.
- **Improving human-AI interactions via sociotechnical research,** which enables AI to extend human ingenuity, creativity and productivity, while also working to reduce inequities of access and working to ensure positive benefits for people and societies worldwide.
- **Accelerating scientific discovery in natural sciences** through proactive knowledge discovery, hypothesis generation, and multiscale multimodal data generation.

Working together as a global research community is essential to realizing the promise of AI to benefit each individual, organization, and society as a whole. AFMR is one means by which we make progress towards these goals.

b. How does your organization collaborate on and invest in research to advance the state of content authentication and provenance?

Microsoft actively collaborates with other leading AI companies and stakeholders to advance the state of the art in content provenance and authentication practices and to help ensure disclosure and transparency mechanisms can be scaled and integrated across platforms.

As discussed above, Microsoft was a founding member of the Coalition for Content Provenance and Authenticity (C2PA), which is been the primary industry organization focused on developing content provenance and authentication specifications that can set and advance an industry baseline for these practices. The formation of the C2PA brought together founding members of the Adobe-led Content Authenticity Initiative (CAI) and the Microsoft- and BBC-led Project Origin, unifying technical specifications under a single entity. C2PA member organizations work together to develop content provenance specifications for common asset types and formats to enable publishers, creators and consumers to trace the origin and evolution of a piece of media, including images, videos, audio and documents. Microsoft's remains an active member of the C2PA, and its commitment to the organization's mission to advance the state of the art in cryptographically signed content provenance methods is reflected in the company's holding of the C2PA's co-chair position since the organization's founding.

In addition, the investment into the watermarking feature in the Azure OpenAI service discussed above, is part of a broader initiative Microsoft has undertaken to further industry-wide best practices for disclosing AI-generated content, validating and displaying those disclosures, and detecting AI generated content when disclosures are missing.

Microsoft's Aether Committee is also leading a study to explore the robustness of individual and valuable combinations of media integrity technologies available today for image, video, and audio generation systems.The study will identify and address potential rough edges and downsides of the technologies as they are being used today and opportunities to boost resilience when it comes to technical attacks and sociotechnical robustness.

c. Does your organization participate in projects, collaborations, and investments in research that support the advancement of AI safety, security, and trustworthiness, as well as risk evaluation and mitigation tools?

Microsoft researchers work to advance the state of the science of responsible AI with the aim of enhancing our understanding of AI, creating new model architectures with novel capabilities, achieving societal benefit, transforming scientific discovery, and extending human capabilities.

In 2023, Microsoft formed the AI & Society research network, which addresses the many bidirectional relationships between AI technologies and people, groups, organizations, and society as a whole. The network aims to bring together a diverse and multidisciplinary community of researchers to explore and shape the social and technical aspects of AI.

In alignment with the efforts of this research network, Microsoft launched the Microsoft Research AI & Society Fellows to catalyze research collaboration between Microsoft researchers and eminent scholars and experts across a range of disciplines at the intersection of AI and its impact on society. This investment has resulted in 13 distinct, ambitious research collaborations bringing together Microsoft researchers and 24 esteemed fellows across academic and industrial disciplines.These research collaborations address some of the pressing challenges facing organizations who aim to advance responsible AI practices, including research focused on advancing the science of AI risk measurement as described in the Measurement section earlier in this report.

Other research collaborations within the AI & Society Fellows program have a broader focus. For example, one of the research collaborations aims to address how to most effectively regulate AI in light of the challenges of doing responsible AI in practice, exploring the challenges faced by front-line practitioners working on a range of responsible AI issues, including fairness, privacy, security, and more. The collaboration aims to take stock of existing social scientific insights into the difficulties faced by regulated entities seeking to comply with existing and forthcoming regulations, complete original empirical studies to fill identified gaps in the existing scholarship on responsible AI in practice, and channel all of these findings into the ongoing debates about how to craft effective regulations of AI.

Microsoft also established the AI Frontiers lab in 2024 to invest in the core technologies that push the frontier of what AI systems can do in terms of capability, reliability, and efficiency. Researchers at our AI Frontiers Lab are not only interested in how well these systems work—they also want to ensure that we build in sociotechnical solutions that can make these systems work in a responsible way.

Our research teams work in close collaboration with our policy and engineering teams to inform our approach to responsible AI. Throughout 2024, Microsoft researchers pushed the frontiers of our understanding of mapping, measuring, and managing AI risks. We summarize some of their research contributions below.

Microsoft also participates in various projects, collaborations, and investments in research that support the advancement of AI safety, security, and trustworthiness. Our investments in research within Microsoft and beyond Microsoft are described in detail in our response to 6A. Some additional examples include:

- Upholding democratic values and respecting human rights
- Fundamental Rights - AI for Good - Microsoft Research
- Advancing AI trustworthiness: Updates on responsible AI research - Microsoft Research
- Multimodal Generative AI: the Next Frontier in Precision Health - Microsoft Research
- Protecting children and vulnerable groups
- Understanding the Representation and Representativeness of Age in AI Data Sets - Microsoft Research
- Tech Coalition | Tech Coalition Hosts EU Briefing and Announces Research on Generative AI
- Safeguarding IP rights
- CATER: Intellectual Property Protection on Text Generation APIs via Conditional Watermarks - Microsoft Research
- Safeguarding privacy
- The Crossroads of Innovation and Privacy: Private Synthetic Data for Generative AI - Microsoft Research
- Avoiding harmful bias
- Measurement and Fairness - Microsoft Research
- Explaining CLIP's performance disparities on data from blind/low vision users - Microsoft Research
- PARIKSHA: A Scalable, Democratic, Transparent Evaluation Platform for Assessing Indic Large Language Models - Microsoft Research
- Avoiding mis- and disinformation and information manipulation
- Generative AI and Plural Governance: Mitigating Challenges and Surfacing Opportunities - Microsoft Research

Please note that this is not an exhaustive list of Microsoft’s research investments to support the advancements of AI safety, security, and trustworthiness.

d. What research or investment is your organization pursuing to minimize socio-economic and/or environmental risks from AI?

In January 2025, we released an updated version of our AI and sustainability playbook, reflecting the targeted actions needed to unlock the full potential of AI for accelerating sustainability progress globally. This report highlights Microsoft's innovations and actions to advance each of the five plays. Examples of our efforts across the five plays include:

- **Play 1: Invest in AI for sustainability:** Microsoft is investing in building AI tools, such as MatterGen and MatterSim, which enable researchers to design and test materials with tenfold greater accuracy and significantly faster performance, while also predicting global weather and atmospheric processes with increased accuracy and at speeds up to 5,000 times greater than current forecasting systems. We are also building AI-enabled tools to empower stakeholders to more effectively and efficiently manage agriculture and water resources and to expedite the licensing process for carbon-free electricity.

- **Play 2: Develop digital and data infrastructure for the inclusive use of AI for sustainability :** We are creating tools to fill critical data gaps, which can enhance AI models for better measuring and predicting complex systems such as biodiversity and climate. For instance, SPARROW captures images and acoustic recordings to gather data on biodiversity and ecosystem health in remote areas. Additionally, we are partnering with G42 on a $1 billion digital ecosystem initiative in Kenya.

- **Play 3: Minimize resource use, expand access to carbon-free electricity, and support local communities:** Microsoft is innovating datacenter development with low-carbon materials like cross-laminated timber. Through an agreement with Brookfield, we aim to add 10.5 gigawatts (GW) of renewable energy to the grid.

- **Play 4: Advance AI policy principles and governance for sustainability :** We advocated for policies that accelerate grid decarbonization, including Federal Energy Regulatory Commission (FERC) transmission rules and provisions in the Inflation Reduction Act in the United States. In addition, we continue to advance AI governance within Microsoft and globally.

- **Play 5: Build workforce capacity to use AI for sustainability:** Microsoft Philanthropies' Skills for Social Impact program trained over 14 million people in digital and AI skills to support a workforce ready to deploy AI for sustainability. As the window for achieving global sustainability goals narrows, the urgency for action intensifies. The world needs every tool at its disposal, and the potential of AI to accelerate sustainability is already being realized. Sustainability is not a journey that can be taken alone, and unlocking the full potential of AI for climate progress requires continued partnerships to combine expertise, technology, and innovation.

Microsoft is deeply committed to advancing fairness in AI through various research initiatives aimed at minimizing quality of service disparities, allocation, and representational harms. One of the key projects in this area is Fairlearn, an open-source toolkit developed by Microsoft and released in 2020. Fairlearn empowers data scientists and developers to assess and improve the fairness of their AI systems. It includes an interactive visualization dashboard and unfairness mitigation algorithms designed to help navigate trade-offs between fairness and model performance. The toolkit focuses on identifying and mitigating two specific types of harms: allocation harms, which occur when AI systems extend or withhold opportunities, resources, or information, and quality-of-service harms, which involve disparities in the performance of AI systems across different demographic groups.

Section 7 - Advancing human and global interests

a. What research or investment is your organization pursuing to maximize socio-economic and environmental benefits from AI? Please provide examples.

Microsoft is committed to making AI accessible to everyone—individuals, organizations, and industries—at every stage of adoption. Through its Airband Initiative, Microsoft is working to close the digital divide, particularly in the Global South. According to the International Telecommunication Union, a third of the world's population remains offline, including two-thirds of the population in Africa. The Airband Initiative aims to bridge this gap by expanding internet access to 250 million unserved and underserved people by 2025, including 100 million in Africa.

Microsoft is actively driving digital access across Latin America, Asia, and Africa. To date, we have provided internet coverage to more than 96 million people worldwide. In addition to infrastructure investments, we are investing billions in countries such as Kenya, South Africa, Indonesia, Thailand, Malaysia, and the Philippines to strengthen cybersecurity capacity, digital skilling, nonprofit support, and local research and development (R&D). Our digital skilling initiatives have already trained 14.1 million people globally in essential digital and AI skills, equipping them with valuable certifications. Moving forward, our AI skilling programs will empower 26 million more people to use and develop AI tools, including 5 million from underserved communities, with a focus on youth, women, rural areas, and the Global South. Over the past year, Microsoft has provided $4.7 billion in grants, discounted software, and services to support these efforts. Additionally, the Microsoft AI for Good Lab is expanding to Abu Dhabi, where a team of local data scientists will drive AI innovation to address critical societal challenges.

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

Microsoft invests in creating skilling resources and working with partners around the world to deliver them, including resources on AI technology in general and on Microsoft's AI technology specifically (for more information, see Microsoft launches new AI Skills Initiative and grant - Microsoft Stories India; Get skilled up and ready on Microsoft AI | Microsoft Learn). In addition, Microsoft has worked with organizations like UNESCO's AI Business Council to build a repository of AI skilling resources from Microsoft and others (Business Council for Ethics of AI | UNESCO).

c. Does your organization prioritize AI projects for responsible stewardship of trustworthy and human-centric AI in support of the UN Sustainable Development Goals? Please provide examples.

When used responsibly, and in a manner that is consistent with fundamental rights and freedoms, AI allows for unparalleled potential to overcome many of the obstacles we face. The Microsoft AI for Good Lab is fueled by a passionate commitment to leverage AI's transformative power for the greater good of humanity. This means identifying and prioritizing the most significant challenges first. We focus on areas where AI can make a tangible difference such as healthcare, education, and inequality. We provided some specific examples in our September 2023 Microsoft and the SDGs Report.

d. Does your organization collaborate with civil society and community groups to identify and develop AI solutions in support of the UN Sustainable Development Goals and to address the world's greatest challenges? Please provide examples.

Microsoft's AI For Good Lab works with a wide range of civil society and community groups to harness the power of AI to solve some of our world's most pressing challenges. For example, we've partnered with Planet and the Institute for Health Metrics and Evaluation (IHME) to employ satellite imagery and AI to produce high-resolution population maps for better climate-event resilience; conducted a machine learning study with Catholic Relief Services to predict food insecurity in Malawi; and worked with SEEDS to develop AI models that provide early warnings of extreme heat risks (for more information, see: Fundamental Rights - AI for Good - Microsoft Research).