Microsoft’s AI Safety Policies

An update prepared for the UK AI Safety Summit

Introduction

Microsoft welcomes the opportunity to share information about how we are advancing responsible artificial intelligence (AI), including by implementing voluntary commitments that we and others made at the White House convening in July.[1] Visibility into our policies and how we put them into practice helps to inform and accelerate responsible technology development and deployment. It can also strengthen public-private partnerships driving progress on AI safety, security, and trust.

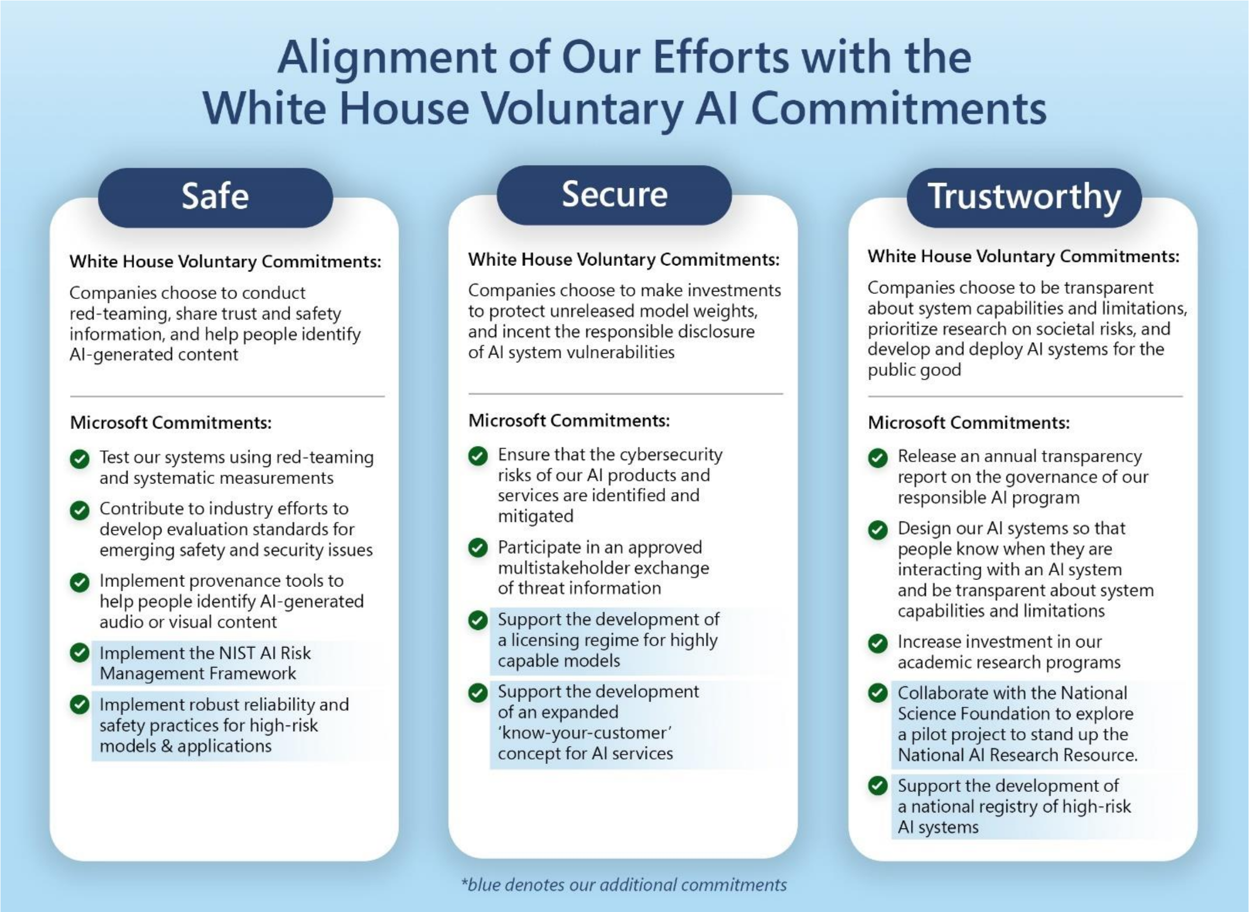
As a developer and deployer of AI models, API services, and applications, Microsoft works to map, measure, and manage risk and apply multi-layered governance that embeds robust checks on processes and outcomes. For frontier models specifically, Microsoft works closely with OpenAI.

Since 2019, Microsoft and OpenAI have been engaged in a long-term collaboration to develop advanced AI systems, underpinned by a shared commitment to responsible development and deployment practices. Microsoft’s efforts to deploy frontier models at scale build upon and complement OpenAI’s leading model development practices. For a comprehensive accounting of the model development and deployment practices that apply to OpenAI’s frontier models as deployed in Microsoft’s offerings, OpenAI’s and Microsoft’s responses to the UK Government’s AI Safety Policies Request should be read together.

The UK Government has requested information about nine areas of practice and investment, many of which relate to the voluntary commitments we published in July.[2] We have indicated these points of connection at the beginning of each section, distinguishing between the White House Voluntary Commitments and the additional independent commitments we made as Microsoft (denoted with blue), as also illustrated in the chart below.

[1] Our commitments to advance safe, secure, and trustworthy AI - Microsoft On the Issues

[2] Microsoft-Voluntary-Commitments-July-21-2023.pdf



Alignment of Our Efforts with the White House Voluntary AI Commitments

Safe

White House Voluntary Commitments:

Companies choose to conduct red-teaming, share trust and safety information, and help people identify AI-generated content

Microsoft Commitments:

- ✓ Test our systems using red-teaming and systematic measurements
- ✓ Contribute to industry efforts to develop evaluation standards for emerging safety and security issues
- ✓ Implement provenance tools to help people identify AI-generated audio or visual content
- ✓ Implement the NIST AI Risk Management Framework
- ✓ Implement robust reliability and safety practices for high-risk models & applications

Secure

White House Voluntary Commitments:

Companies choose to make investments to protect unreleased model weights, and incent the responsible disclosure of AI system vulnerabilities

Microsoft Commitments:

- ✓ Ensure that the cybersecurity risks of our AI products and services are identified and mitigated
- ✓ Participate in an approved multistakeholder exchange of threat information
- ✓ Support the development of a licensing regime for highly capable models
- ✓ Support the development of an expanded 'know-your-customer' concept for AI services

Trustworthy

White House Voluntary Commitments:

Companies choose to be transparent about system capabilities and limitations, prioritize research on societal risks, and develop and deploy AI systems for the public good

Microsoft Commitments:

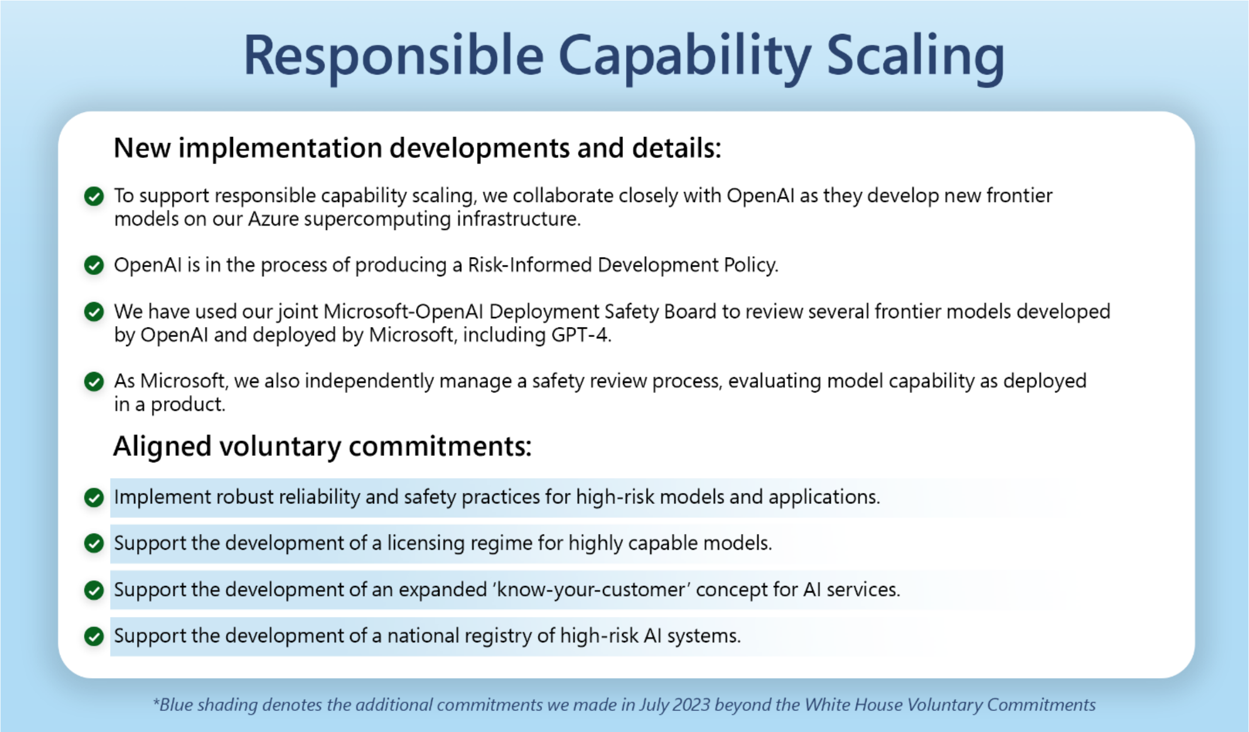
- ✓ Release an annual transparency report on the governance of our responsible AI program
- ✓ Design our AI systems so that people know when they are interacting with an AI system and be transparent about system capabilities and limitations
- ✓ Increase investment in our academic research programs
- ✓ Collaborate with the National Science Foundation to explore a pilot project to stand up the National AI Research Resource.
- ✓ Support the development of a national registry of high-risk AI systems

**blue denotes our additional commitments*

We also recognize that each of the nine areas of practice accrue to *mapping*, *measuring*, *managing*, and *governing* AI model development and deployment risk, the structure and terminology offered by the National Institute of Standards and Technology (NIST) AI Risk Management Framework (RMF).[3] To help provide context on how we are realizing our commitment to implement the NIST AI RMF, the terminology of "map, measure, manage, and govern" is used throughout this response to the UK Government's AI Safety Policies Request.

[3] AI Risk Management Framework | NIST

Responsible Capability Scaling



Responsible Capability Scaling

New implementation developments and details:

- To support responsible capability scaling, we collaborate closely with OpenAI as they develop new frontier models on our Azure supercomputing infrastructure.
- OpenAI is in the process of producing a Risk-Informed Development Policy.
- We have used our joint Microsoft-OpenAI Deployment Safety Board to review several frontier models developed by OpenAI and deployed by Microsoft, including GPT-4.
- As Microsoft, we also independently manage a safety review process, evaluating model capability as deployed in a product.

Aligned voluntary commitments:

- Implement robust reliability and safety practices for high-risk models and applications.
- Support the development of a licensing regime for highly capable models.
- Support the development of an expanded 'know-your-customer' concept for AI services.
- Support the development of a national registry of high-risk AI systems.

**Blue shading denotes the additional commitments we made in July 2023 beyond the White House Voluntary Commitments*

Microsoft is committed to responsible development and deployment of increasingly capable AI systems, including frontier models and the applications by which users access them. We have put in place and are continuously investing in a range of policies, practices, and partnerships to ensure we are appropriately mapping, measuring, and managing AI technology capability and risks across the AI product lifecycle, as further discussed below in the context of model evaluations and red teaming. Learnings from this process are fed back into policies for defining off-limits technology development or needed guardrails across release stages or deployment contexts.

To support responsible capability scaling, we collaborate closely with OpenAI as they are developing new frontier models on our Azure supercomputing infrastructure. OpenAI provides details of their risk mitigation practices and in-progress Risk-Informed Development Policy in their response to the UK Government's AI Safety Policies Request.

When it comes to frontier model deployment, Microsoft and OpenAI have together defined capability thresholds that act as a trigger to review models in advance of their first release or downstream deployment. The scope of a review, through our joint Microsoft-OpenAI Deployment Safety Board (DSB), includes model capability discovery. We established the joint DSB's processes in 2021, anticipating a need for a comprehensive pre-release review process focused on AI safety and alignment, well ahead of regulation or external commitments mandating the same.

We have exercised this review process with respect to several frontier models, including GPT-4. Using Microsoft's Responsible AI Standard and OpenAI's experience building and deploying advanced AI systems, our teams prepare detailed artefacts for the joint DSB review. Artefacts record the process by which our organizations have mapped, measured, and managed risks, including through the use of adversarial testing and third-party evaluations as appropriate. We continue to learn from, and refine, the joint DSB process, and we expect it to evolve over time.

As Microsoft, we also independently manage a subsequent safety review process. We evaluate model capability as deployed in a product – where additional safety mitigations can be implemented and measured for impact – to check for effective and appropriate mitigations prior to release.

Across all mapping, measurement, and management activities as well as product deployment decisions, governance is critical. Microsoft operates a multi-tiered approach to both top-down and distributed responsible AI governance, allowing us to set clear policies, convene leadership to make tough calls, and drive consistency rigor.

Our Responsible AI Council is a vital component of our governance structure. The Responsible AI Council, which is co-chaired by Brad Smith, our Vice Chair and President, and Kevin Scott, our Chief Technology Officer and EVP of AI, brings together accountable leaders as well as teams involved in research, engineering, and policy to confront difficult questions and ensure alignment and execution consistent with our responsible AI vision and commitments.[4]

Across products groups, designated Responsible AI Division Leads and Champs also work with their accountable Responsible AI Corporate Vice Presidents and Microsoft's Office of Responsible AI team to measure and continuously improve responsible AI practice implementation, including through shared learning and strategic investment in tools.

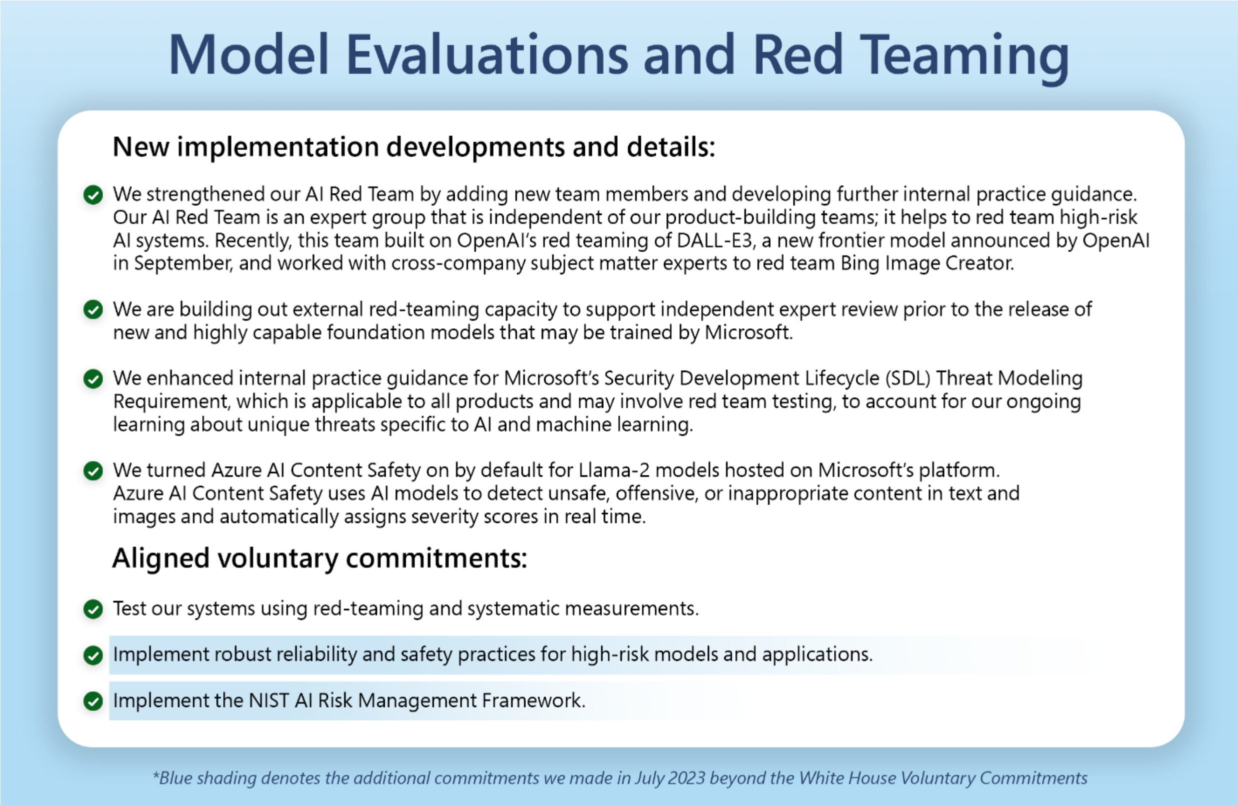
Microsoft also supports the formation of a globally coordinated licensing regime to govern the development and deployment of highly capable frontier models, enabling appropriate oversight into risks and mitigations.[5] Through such a regime, best practices for mapping risk (e.g., defining leading indicators of potential model risk) and for managing risk could regularly be assessed for impact and improved, and reliable processes for exchanging and using information about evaluations and measurements could be established. A framework for close coordination and information flows between licensees and their regulators is critical to ensure that developments material to the achievement of safety and security objectives can be acted upon swiftly.[6]

[4] Reflecting on our responsible AI program: Three critical elements for progress - Microsoft On the Issues

[5] How do we best govern AI? - Microsoft On the Issues

[6] Microsoft-Voluntary-Commitments-July-21-2023.pdf

Model Evaluations and Red Teaming



Model Evaluations and Red Teaming

New implementation developments and details:

- We strengthened our AI Red Team by adding new team members and developing further internal practice guidance. Our AI Red Team is an expert group that is independent of our product-building teams; it helps to red team high-risk AI systems. Recently, this team built on OpenAI's red teaming of DALL-E3, a new frontier model announced by OpenAI in September, and worked with cross-company subject matter experts to red team Bing Image Creator.
- We are building out external red-teaming capacity to support independent expert review prior to the release of new and highly capable foundation models that may be trained by Microsoft.
- We enhanced internal practice guidance for Microsoft's Security Development Lifecycle (SDL) Threat Modeling Requirement, which is applicable to all products and may involve red team testing, to account for our ongoing learning about unique threats specific to AI and machine learning.
- We turned Azure AI Content Safety on by default for Llama-2 models hosted on Microsoft's platform. Azure AI Content Safety uses AI models to detect unsafe, offensive, or inappropriate content in text and images and automatically assigns severity scores in real time.

Aligned voluntary commitments:

- Test our systems using red-teaming and systematic measurements.
- Implement robust reliability and safety practices for high-risk models and applications.
- Implement the NIST AI Risk Management Framework.

**Blue shading denotes the additional commitments we made in July 2023 beyond the White House Voluntary Commitments*

Responsible development and deployment of AI requires ongoing efforts to map, measure, and manage the potential for harms and misuse of systems. As we map potential harms and misuse through our Responsible AI Impact Assessment[7] and processes like red teaming, we also develop and implement practices to manage risk and measure effectiveness in reducing the potential for harms or misuse. We implement a layered approach, mapping, measuring, and managing risks of harm and misuse as AI is developed and deployed across the technology architecture, including at the model, API service, and application layers, working in collaboration with OpenAI, which provides Microsoft frontier models that we make available via platform services and leverage in applications. As products evolve or we learn more, we also continue to invest throughout the product lifecycle in mapping, measurement, and management and improving our overall approach.

Red teaming, including by simulating real-world attacks and exercising techniques that persistent threat actors might use, has long been a foundational security practice at Microsoft.[8] In 2018, we established our AI Red Team: a group of interdisciplinary experts dedicated to thinking like attackers and probing AI systems for failures.[9] Through research, we have also expanded our red teaming practices to map risks outside of traditional security, including those associated with benign usage scenarios and responsible AI. Today, for example, a red team might probe a Large Language Model (LLM) or an LLM-backed

[7] Development of an Impact Assessment is required by Goal A1 of our Responsible AI standard, and use of the Assessment is prompted through other Requirements of the standard. Microsoft Responsible AI Standard v2 General Requirements

[8] https://download.microsoft.com/download/C/1/9/C1990DBA-502F-4C2A-848D-392B93D9B9C3/Microsoft_Enterprise_Cloud_Red_Teaming.pdf

[9] Microsoft AI Red Team building future of safer AI | Microsoft Security Blog

feature for prompt injection attacks (where content submitted to the LLM by the user or by a third party results in unintended actions), content harms (where malicious or benign usage of the system results in harmful or inappropriate AI-generated content), and privacy harms (where an LLM leaks correct or incorrect personal information about an individual), among others. In the case of Bing Chat, AI red teaming focuses not only on how threat actors could subvert the system using security techniques but also on how the system could generate harmful or otherwise problematic content when non-malicious users interact with it.[10]

Because AI red teaming can surface previously unknown harms, confirm whether suspected harms are observable in a product, and inform measurement and risk management, iterative red teaming is critical at the base model level and throughout AI product development and deployment. Red teaming an AI model helps identify how it might be misused and the scope of its capabilities and limitations, improving the model development process and informing analysis of applications for which a model is suitable. Application-level AI red teaming takes a broader and more applied view, mapping model or application failures that persist despite different model- or application-level safety mitigations. Moreover, because AI systems are constantly evolving, we pursue multiple rounds of AI red teaming to look for vulnerabilities and attempt to measure their pervasiveness as well as mitigate them, both prior to product shipment and as an ongoing practice. Red teaming generative AI systems also requires multiple attempts; because a prompt may not lead to failure in one instance but could in another (as the probabilistic nature of generative AI allows for a wider range in creative output), we perform multiple rounds of red teaming in the same operation.[11]

To strengthen our internal governance of AI threat modeling, which may include red teaming, and reflect our ongoing AI threat research and learnings, we’ve updated internal practice guidance for Microsoft’s Security Development Lifecycle (SDL) Threat Modeling requirement, which is applicable to all products, to account for our ongoing learnings about unique threats specific to AI and machine learning.

For all generative AI products characterized as high risk, we are also implementing processes to ensure consistent and holistic AI red teaming by our AI Red Team, an expert group independent to base model or product groups. We are also building out external red-teaming capacity to ensure our readiness to organize red team testing by one or more independent experts prior to the release of new and highly capable foundation models that may be trained by Microsoft, consistent with our July commitment.[12] The topics covered by such red team testing will include testing of dangerous capabilities, including related to biosecurity and cybersecurity.

While red teaming is useful for mapping risks, systematic measurement is important for understanding the prevalence of risks and the effectiveness of risk mitigations. Through systematic measurement, we evaluate model performance against specific metrics, and the range of issues we are systematically measuring for all products is regularly expanding. Some examples of metrics include:[13]

- Groundedness, through which we measure how well a model’s generated answers align with information from the input source. Answers are verified as claims against context in the user-

[10] Microsoft AI Red Team building future of safer AI | Microsoft Security Blog

[11] Microsoft AI Red Team building future of safer AI | Microsoft Security Blog; https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW14Gtw

[12] Microsoft-Voluntary-Commitments-July-21-2023.pdf

[13] Monitoring evaluation metrics descriptions and use cases (preview) - Azure Machine Learning | Microsoft Learn

defined ground truth source, and even if answers are factually correct, if not verifiable against source text, then they're scored as ungrounded.

- Relevance, through which we measure the extent to which a model's generated answers are pertinent and directly related to the given questions.
- Similarity, through which we measure the equivalencies between a "ground-truth" answer and a prediction sentence generated by an AI model.

We also share responsible AI capabilities and tools that support measurement of these and other metrics open source on GitHub and via Azure Machine Learning, empowering platform and application developers to access model explanations and error and fairness evaluations.[14] For example, in May, we announced prompt flow, which allows users to create prompt workflows that connect to various language models and data sources and assess the quality of their workflows against metrics such as groundedness, ultimately enabling them to choose the best prompt for their use case.[15]

We also systematically measure the impact of mitigations for potentially unsafe model outputs, such as content filtering systems that run prompts and completions though models aimed at detecting and preventing the output of harmful content. As part of our commitment to both building responsible AI systems and helping others do so, we integrate content filtering across Azure OpenAI[16] and have also prioritized work on tools for customers. Azure AI Content Safety uses AI models to detect unsafe, offensive, or inappropriate content in text and images and automatically assigns severity scores in real time, enabling customers to efficiently and in a prioritized manner review flagged items and take informed action.[17] We've also turned this safety system on by default for Llama-2 models hosted on Microsoft's platform, mitigating intentional misuse as well as potential mistakes by the model.[18]

Our process of launching Bing Chat, which runs on a variety of advanced Microsoft and OpenAI technologies, including OpenAI's GPT-4 model, provides a product-specific example of how we have approached mapping, measuring, and managing risk at the model and application layers, including through red teaming and model evaluations.[19]

- *Map.* At the model level, our work began with extensive red teaming in collaboration with OpenAI, and a multidisciplinary team of experts also conducted numerous rounds of application-level red teaming before our limited release preview, helping us better understand how the system could be exploited and improve our mitigations. Non-adversarial red teamers also extensively probed the new application for risks that could arise in benign usage scenarios. Post release, red teamers from different regions and backgrounds continue to attempt to compromise the system, and their findings are used to expand the datasets that Bing Chat uses to improve the system.
- *Measure.* To better understand and address potential harms, we developed additional metrics specific to new AI experiences like jailbreaks, harmful content, and ungrounded content. We

[14] Responsible AI dashboard | Microsoft AI Lab, Microsoft Responsible AI Toolbox - Microsoft Responsible AI, Use Responsible AI scorecard (preview) in Azure Machine Learning - Azure Machine Learning | Microsoft Learn

[15] Microsoft Build brings AI tools to the forefront for developers - The Official Microsoft Blog; What is Azure Machine Learning prompt flow (preview) - Azure Machine Learning | Microsoft Learn

[16] Azure OpenAI Service content filtering - Azure OpenAI | Microsoft Learn

[17] Azure AI Content Safety – AI Content Moderation | Microsoft Azure

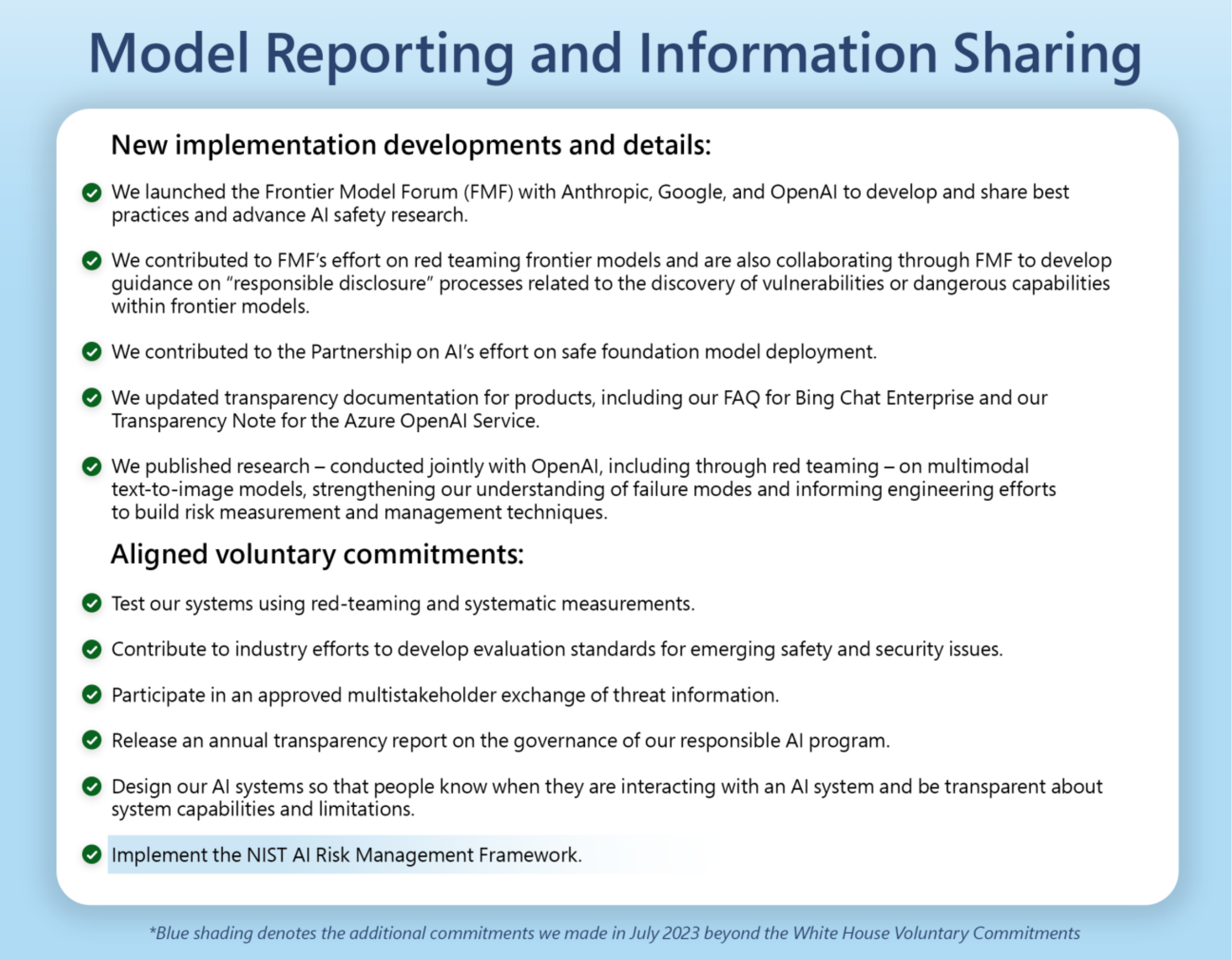
[18] Introducing Llama 2 on Azure (microsoft.com)

[19] The new Bing - Our approach to Responsible AI

also enabled measurement at scale through partially automated measurement pipelines (that then allowed us to build tools for third parties). Each time the product changes, existing mitigations are updated, or new mitigations are proposed, we update our measurement pipelines to assess both product performance and responsible AI metrics. As we identify new issues during a preview period and ongoing red teaming, we also expand measurement sets to assess additional harms.

- *Manage.* As we map harms and measure them, we manage an ongoing process of developing and measuring the impact of mitigations. For Bing Chat, some of the steps that we've taken to manage risks include: an incremental release strategy (to allow us to mitigate emerging issues before broader release); inclusion of references to source websites for search results (to mitigate the risk that users may over-rely on ungrounded generated content); the use of classifiers and content filters (which may stop flagged generated content from being returned to the user); the use of metaprompting (which gives instructions to a model to guide its behavior); and limitations on user-Bing reply exchanges per session (to limit conversational drift).

Model Reporting and Information Sharing



Model Reporting and Information Sharing

New implementation developments and details:

- We launched the Frontier Model Forum (FMF) with Anthropic, Google, and OpenAI to develop and share best practices and advance AI safety research.
- We contributed to FMF's effort on red teaming frontier models and are also collaborating through FMF to develop guidance on "responsible disclosure" processes related to the discovery of vulnerabilities or dangerous capabilities within frontier models.
- We contributed to the Partnership on AI's effort on safe foundation model deployment.
- We updated transparency documentation for products, including our FAQ for Bing Chat Enterprise and our Transparency Note for the Azure OpenAI Service.
- We published research – conducted jointly with OpenAI, including through red teaming – on multimodal text-to-image models, strengthening our understanding of failure modes and informing engineering efforts to build risk measurement and management techniques.

Aligned voluntary commitments:

- Test our systems using red-teaming and systematic measurements.
- Contribute to industry efforts to develop evaluation standards for emerging safety and security issues.
- Participate in an approved multistakeholder exchange of threat information.
- Release an annual transparency report on the governance of our responsible AI program.
- Design our AI systems so that people know when they are interacting with an AI system and be transparent about system capabilities and limitations.
- Implement the NIST AI Risk Management Framework.

**Blue shading denotes the additional commitments we made in July 2023 beyond the White House Voluntary Commitments*

Transparency is a foundational responsible AI principle to which Microsoft committed in 2019, and as our Responsible AI Standard captures, communication to stakeholders about the capabilities and limitations of the AI systems they use is key to realizing that principle.[20] Driving implementation of all six of our responsible AI principles and working through sensitive use cases has also reinforced for us the importance of providing context to our customers to empower them to deploy their AI systems responsibly.[21]

What specifically needs to be communicated and to whom as part of Microsoft's transparency commitment is defined by our Responsible AI Standard.[22] Conceptually, we recognize that different stakeholders have different needs and goals when it comes to transparency. For instance, the approach to transparency that works for an end-user interacting with a model like GPT-4 through a specific application like Bing Chat is likely to be different from the approach that works for application developers using an offering like the Azure OpenAI Service to incorporate GPT-4 into their own AI

[20] Responsible AI Principles and Approach | Microsoft AI; Responsible AI Standard v2 Goal T2: Communication to Stakeholders
[21] The building blocks of Microsoft's responsible AI program - Microsoft On the Issues
[22] Microsoft Responsible AI Standard v2 General Requirements

systems. Even a particular end-user may have different transparency needs when using GPT-4 to research a medical procedure versus to caption vacation photos.

We take a human-centered approach in which stakeholders and their goals inform the development and evaluation of layered transparency measures. For products where Microsoft designs and develops the end-to-end system, we rely on a combination of product features and documentation to achieve transparency that meets the various needs of stakeholders. For example, Bing Chat's user-centered design incorporates user experience interventions in the interface itself to disclose that it's powered by AI and to help users understand the capabilities and limitations of the system. Product FAQs can be a resource for application users that need more context on key issues. For example, our GitHub Copilot FAQ gives context on limitations in the functionality and security of the code it generates as well as some of the human oversight, privacy, and fairness implications of its use,[23] and a Bing Chat Enterprise FAQ provides context on uses and limitations.[24] Other stakeholders, like regulators, may have broader questions and seek deeper context on our approach to developing AI applications; those are best answered outside of the application in stand-alone documentation, such as our Bing Chat whitepaper.[25]

For our platform systems like the Azure OpenAI Service, where Microsoft makes models available to customers but doesn't design an AI system end-to-end, we rely on documentation to communicate information to customers to enable them to integrate those models responsibly. For example, Transparency Notes enable us to communicate the purposes, capabilities, and limitations of AI systems so our customers can understand when and how to deploy our platform technology. Our Azure OpenAI Transparency Note provides context on text, image, and speech models available through that service, describing the techniques the models employ, the use cases for which they're envisioned, and the limitations and potential biases in their behavior.[26] This Transparency Note builds on the system card documentation that OpenAI itself produces for models like GPT-4.[27]

Beyond any single platform or product, we believe transparency in research and corporate practice can be effective in helping the public understand the state of the art and in driving organizational accountability. In September, for instance, we released the results of a study, conducted jointly with OpenAI, to explore multimodal text-to-image models, including through red teaming, to understand failure modes, inform engineering efforts to build measurement and mitigation techniques, and reflect on longer-term fairness harms.[28] In July, we committed to releasing an annual transparency report about our policies, systems, progress, and performance in managing AI responsibly and safely. Specifically, our forthcoming, inaugural responsible AI transparency report will address the functioning and ongoing development of our governance systems in addition to providing case studies on the implementation of responsible AI measures.[29]

To ensure the latest context informs our responsible AI practices and that we share learnings with others, Microsoft also leverages multiple processes to exchange information. For example, along with Anthropic, Google, and OpenAI, we launched the Frontier Model Forum (FMF) to share best practices

[23] GitHub Copilot · Your AI pair programmer · GitHub
[24] Frequently asked questions about Bing Chat Enterprise | Microsoft Learn
[25] The new Bing - Our approach to Responsible AI
[26] Transparency Note for Azure OpenAI - Azure AI services | Microsoft Learn
[27] gpt-4-system-card.pdf (openai.com)
[28] Frontiers of multimodal learning: A responsible AI approach - Microsoft Research
[29] Microsoft-Voluntary-Commitments-July-21-2023.pdf

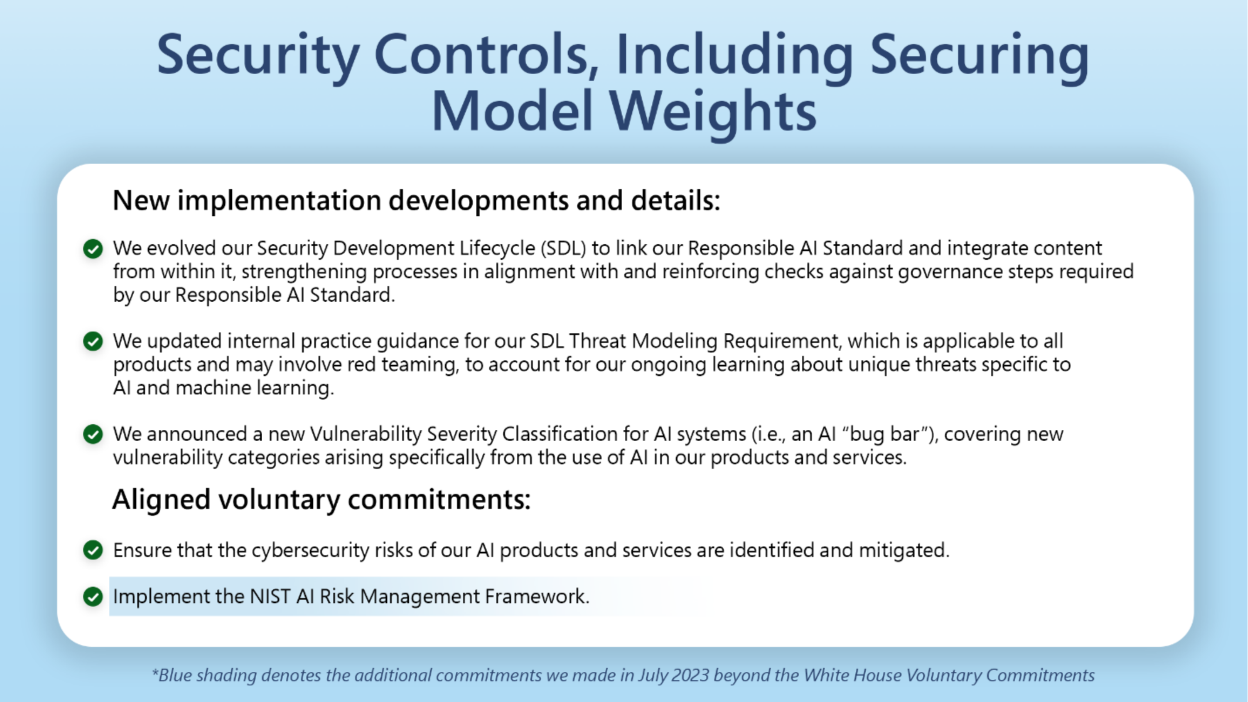
and advance AI safety research.[30] We contributed to FMF's effort to share case studies on red teaming frontier models,[31] and we are collaborating through FMF to develop guidance on "responsible disclosure" processes related to the discovery of vulnerabilities or dangerous capabilities within frontier models. Through the Partnership on AI, we have also contributed to the development of guidance, released for public comment in October, for safe foundation model deployment.[32] In a more security-specific context, Microsoft Threat Intelligence, which tracks and helps defend against the most sophisticated threat actors impacting our customers, also exchanges threat intel information with those best positioned to protect and use it.

[30] Microsoft, Anthropic, Google, and OpenAI launch Frontier Model Forum - Microsoft On the Issues

[31] FMF-AI-Red-Teaming.pdf (frontiermodelforum.org)

[32] Partnership on AI Releases Guidance for Safe Foundation Model Deployment, Takes the Lead to Drive Positive Outcomes and Help Inform AI Governance Ahead of AI Safety Summit in UK -

Security Controls, Including Securing Model Weights



Security Controls, Including Securing Model Weights

New implementation developments and details:

- ✓ We evolved our Security Development Lifecycle (SDL) to link our Responsible AI Standard and integrate content from within it, strengthening processes in alignment with and reinforcing checks against governance steps required by our Responsible AI Standard.
- ✓ We updated internal practice guidance for our SDL Threat Modeling Requirement, which is applicable to all products and may involve red teaming, to account for our ongoing learning about unique threats specific to AI and machine learning.
- ✓ We announced a new Vulnerability Severity Classification for AI systems (i.e., an AI "bug bar"), covering new vulnerability categories arising specifically from the use of AI in our products and services.

Aligned voluntary commitments:

- ✓ Ensure that the cybersecurity risks of our AI products and services are identified and mitigated.
- ✓ Implement the NIST AI Risk Management Framework.

**Blue shading denotes the additional commitments we made in July 2023 beyond the White House Voluntary Commitments*

Responsible AI is a commitment that extends across the product lifecycle and enabling infrastructure. Microsoft's decades-long and ongoing investments in developing and implementing state-of-the-art cybersecurity practices are integrated with our defense-in-depth efforts for AI systems and information, including model weights. A holistic approach is critical, including governance of AI security policies and practices; identification of AI systems, data, and supply chains as well as potential risks; protection of systems and information; detection of AI threats; and response to and recovery from discovered AI issues and vulnerabilities, including through rapid containment and continuous improvement processes.

Governance is foundational, and along with structures and processes to bring together centralized and distributed security engineering, physical security, threat intelligence, and security operations teams that own implementation and enforcement,[33] SDL[34] requirements, tools, and verification processes are key to our approach. Through SDL implementation, facilitated and monitored by central engineering system teams and our Digital Security & Resilience team led by our CISO among others, all Microsoft products are responsible for SDL practices, including threat modeling to map potential vulnerabilities and measure and manage risk, including through mapped and measured mitigations.

[33] Some of the teams include Azure Security (responsible for continuously improving the built-in security posture of Azure at all layers: the datacenter, physical infrastructure, and cloud products and services); Cyber Defense Operations Center (a fusion center that brings together incident responders, data scientists, and security engineers to provide around-the-clock protection to our corporate infrastructure and cloud infrastructure that customers use); Digital Security & Resilience (organization led by our CISO and dedicated to enabling Microsoft to build the most trusted devices and services while keeping our company and customers protected); Identity and Network Access (identity platform security and defense); Microsoft Defender Experts and Microsoft Defender Threat Intelligence (product-focused security researchers, applied scientists, and threat intelligence analysts); Microsoft Security Response Center (vulnerability research and response); and Microsoft Threat Intelligence Center (team dedicated to identifying and tracking the most sophisticated adversaries impacting Microsoft customers).

[34] Microsoft Security Development Lifecycle

For AI technology, we've updated our SDL threat modeling requirement to explicitly account for our ongoing learnings about unique AI threats, which we have been in a leader in researching and developing frameworks to systematically organize. For example, along with MITRE and others, we helped create the Adversarial Machine Learning Threat Matrix.[35] Our AI and Ethics in Engineering and Research (AETHER)[36] Security Engineering Guidance added AI-specific threat enumeration and mitigation guidance to existing SDL threat modeling practices,[37] and our AI bug bar provides a severity classification for vulnerabilities commonly impacting AI and machine learning systems.[38] Internal trainings also provide context on threat modeling for AI.

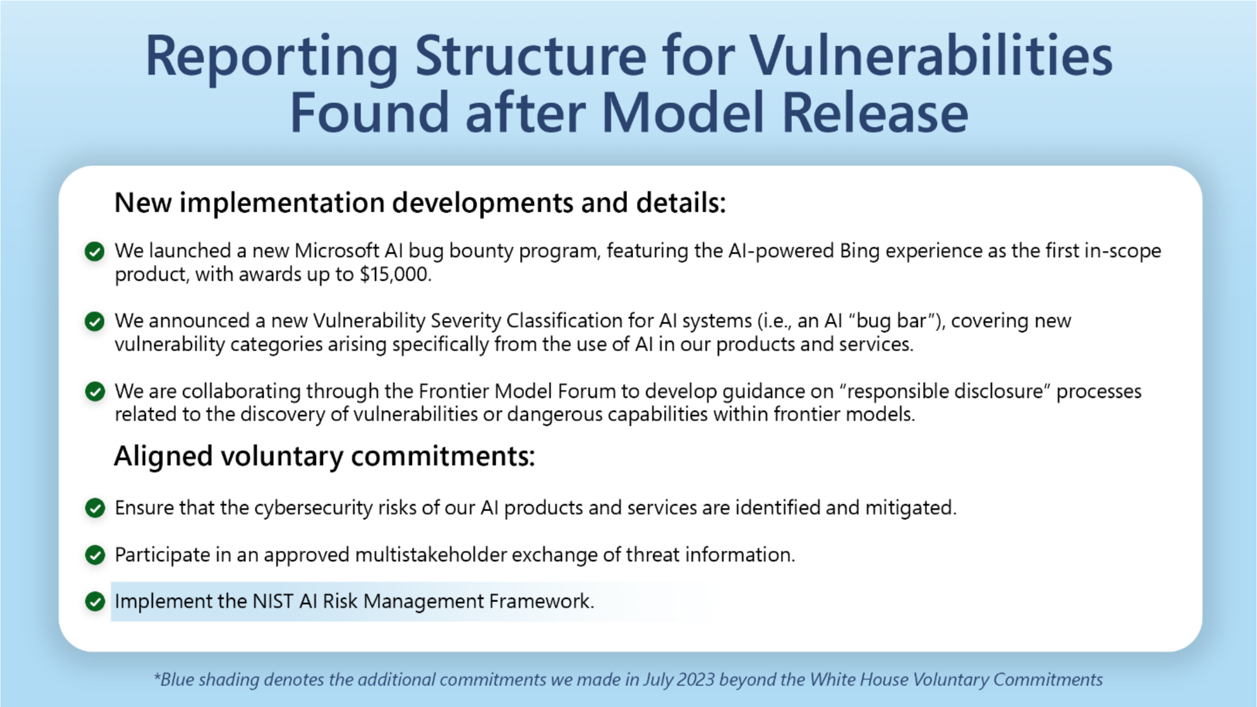
SDL protection, detection, and response requirements also apply to AI technology. For instance, Microsoft employs strong identity and access control, holistic security monitoring (for both external and internal threats) with rapid incident response, and continuous security validation (such as simulated attack path analysis) for our AI environments.[39] Model weights are encrypted-at-rest and encrypted-in-transit where applicable 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.[40]

Robust physical, operational, and network security measures, including for supplier management, identity and access management, and insider threat monitoring, also protect cloud infrastructure.[41] Supplier security and privacy are governed by our Supplier Security and Privacy Assurance program.[42] 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.[43] We log and retain access requests and analyze data to detect anomalies and prevent and detect unnecessary or unauthorized access.[44] We also employ multiple strategies for securing the network boundary.[45]

Since making voluntary commitments at the White House convening in July,[46] we have taken key steps to further invest in governance and implementation. We have linked our Responsible AI Standard and content from it within our SDL, strengthening processes for having responsible AI risks inform secure development processes. Our robust integration also reinforces checks against governance steps required by our Responsible AI Standard. (Our Responsible AI Standard also continues to reference SDL, ensuring that cybersecurity risks inform AI risk management. [47])

[35] Cyberattacks against machine learning systems are more common than you think | Microsoft Security Blog
[36] Satya Nadella email to employees: Embracing our future: Intelligent Cloud and Intelligent Edge - Stories (microsoft.com)
[37] Threat Modeling AI/ML Systems and Dependencies - Security documentation | Microsoft Learn
[38] Microsoft Vulnerability Severity Classification for Artificial Intelligence and Machine Learning Systems
[39] Microsoft-Voluntary-Commitments-July-21-2023.pdf
[40] Microsoft-Voluntary-Commitments-July-21-2023.pdf
[41] What is Cloud Infrastructure? | Microsoft Azure
[42] Supplier management overview - Microsoft Service Assurance | Microsoft Learn
[43] Datacenter physical access security - Microsoft Service Assurance | Microsoft Learn
[44] Datacenter physical access security - Microsoft Service Assurance | Microsoft Learn
[45] Network security - Microsoft Service Assurance | Microsoft Learn
[46] Our commitments to advance safe, secure, and trustworthy AI - Microsoft On the Issues
[47] Microsoft Responsible AI Standard v2 General Requirements (see Goal PS2)

Reporting Structure for Vulnerabilities Found after Model Release



Reporting Structure for Vulnerabilities Found after Model Release

New implementation developments and details:

- ✓ We launched a new Microsoft AI bug bounty program, featuring the AI-powered Bing experience as the first in-scope product, with awards up to $15,000.
- ✓ We announced a new Vulnerability Severity Classification for AI systems (i.e., an AI "bug bar"), covering new vulnerability categories arising specifically from the use of AI in our products and services.
- ✓ We are collaborating through the Frontier Model Forum to develop guidance on "responsible disclosure" processes related to the discovery of vulnerabilities or dangerous capabilities within frontier models.

Aligned voluntary commitments:

- ✓ Ensure that the cybersecurity risks of our AI products and services are identified and mitigated.
- ✓ Participate in an approved multistakeholder exchange of threat information.
- ✓ Implement the NIST AI Risk Management Framework.

**Blue shading denotes the additional commitments we made in July 2023 beyond the White House Voluntary Commitments*

Microsoft is an industry leader in Coordinated Vulnerability Disclosure (CVD), a process whereby vendors receive information from external finders about potential vulnerabilities affecting their products and services and work with those finders to investigate and mitigate confirmed vulnerabilities and release related information publicly in a way that minimizes risk to users. We have published our CVD policy externally, and we have also established a clear process by which we receive vulnerability reports from external finders and work with them throughout the processes of investigating, remediating, and providing public information about confirmed vulnerabilities, including by giving credit to finders.[48] The Microsoft Security Response Center (MSRC) receives all such reports from external finders and manages coordination throughout the CVD process, working with others internally to investigate and remediate as appropriate.[49]

External finders may also be eligible for financial reward as part of our Bug Bounty Programs.[50] Our bounty range varies by product, including up to $100,000 among cloud programs and up to $250,000 among platform programs.[51] In October, we launched a new Microsoft AI bug bounty program reflecting key recent investments and learnings, including from an AI research challenge and the process of updating our vulnerability severity classification for AI systems.[52] This new bounty program, with awards up to $15,000, features the AI-powered Bing experience as the first in-scope product.[53]

[48] microsoft.com/en-us/msrc/cvd
[49] MSRC Researcher Portal (microsoft.com)
[50] https://microsoft.com/msrc/bounty
[51] Microsoft Bounty Programs | MSRC
[52] Introducing the Microsoft AI Bug Bounty Program featuring the AI-powered Bing experience | MSRC Blog | Microsoft Security Response Center
[53] Microsoft AI Bounty | MSRC

Upon confirmation of a vulnerability received from an external finder, MSRC works with the external finder and relevant internal product team(s) to develop, test, and release a mitigation, often involving a software update. In doing so, Microsoft uses vulnerability severity to prioritize rapid mitigation work, focusing on the most critical issues first (rather than, for example, mitigating issues in the order in which they were received or confirmed).

To strengthen transparency with customers and security researchers regarding our approach to risk-based approach to prioritizing rapid mitigation work, we provide vulnerability classifications and severity ratings, including for different classes of products. For example, MSRC maintains a vulnerability severity classification for online services.[54] Recently, we also issued a new vulnerability severity classification for AI systems (i.e., AI bug bar), covering new vulnerability categories arising specifically from the use of AI in our products and services.[55] MSRC also continues to maintain a security update severity rating system that is aligned with our severity classification systems for classes of products (i.e., Low, Moderate, Important, and Critical severity ratings), supporting customers with understanding risk and prioritizing patching.[56]

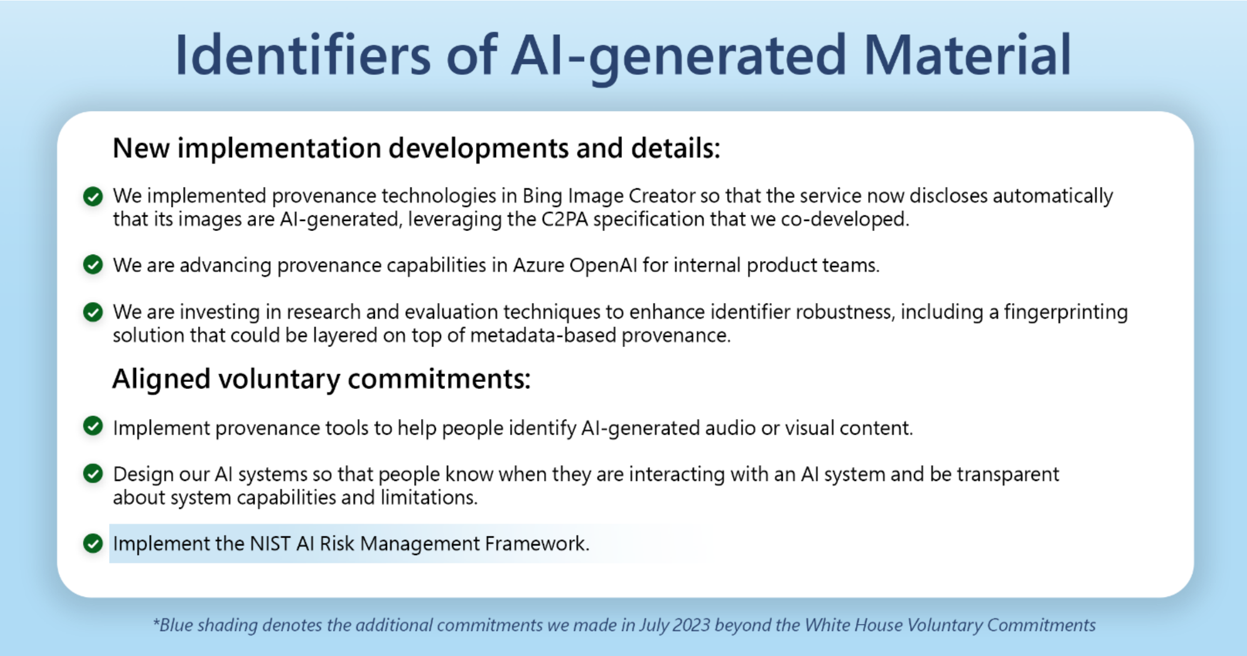
Microsoft is also collaborating with others in industry through the Frontier Model Forum to scope a new “responsible disclosure” process through which providers can receive and share information related to the discovery of vulnerabilities or dangerous capabilities within frontier models.

[54] Microsoft Vulnerability Severity Classification for Online Services

[55] Microsoft Vulnerability Severity Classification for Artificial Intelligence and Machine Learning Systems

[56] Security Update Severity Rating System (microsoft.com)

Identifiers of AI-generated Material



Identifiers of AI-generated Material

New implementation developments and details:

- ✓ We implemented provenance technologies in Bing Image Creator so that the service now discloses automatically that its images are AI-generated, leveraging the C2PA specification that we co-developed.
- ✓ We are advancing provenance capabilities in Azure OpenAI for internal product teams.
- ✓ We are investing in research and evaluation techniques to enhance identifier robustness, including a fingerprinting solution that could be layered on top of metadata-based provenance.

Aligned voluntary commitments:

- ✓ Implement provenance tools to help people identify AI-generated audio or visual content.
- ✓ Design our AI systems so that people know when they are interacting with an AI system and be transparent about system capabilities and limitations.
- ✓ Implement the NIST AI Risk Management Framework.

**Blue shading denotes the additional commitments we made in July 2023 beyond the White House Voluntary Commitments*

To indicate that a piece of content is AI-generated, there are two common approaches used in Microsoft products, depending on the AI-generated media format: watermarking and metadata-based provenance. Watermarks can be visible (e.g., a logo on an AI-generated image to indicate that it's AI-generated) or invisible (and thus rely on a detection tool); to establish provenance, information can be included in metadata attached to AI-generated content. A third, less common approach is fingerprinting.

In 2021, Microsoft co-founded the Coalition for Content Provenance and Authenticity (C2PA) alongside Adobe, Arm, BBC, Intel, and Truepic and co-developed the C2PA technical specification, the leading open standard upon which an interoperable provenance ecosystem can be built.[57] Much like we treat mailing envelopes or boxes as containers for physical content, digital assets like images and videos are placed inside their own containers. The C2PA specification defines how to embed a cryptographically sealed, verifiable unit of provenance information called a "C2PA Manifest" inside a digital asset's container. For example, when a user creates an image with C2PA-enabled software, it will generate a provenance manifest and cryptographically bind it to the JPEG file. Verification tools can then be used to view the manifest attached to the image since it is embedded in the JPEG file structure, or "container," itself.[58] C2PA is also designed for use as the final signing step before content is shared or published.

Microsoft has been working along with others in C2PA to implement the specification, improving transparency and helping to drive the broader ecosystem forward. In May 2023, we announced new media provenance capabilities and plans to use C2PA to mark and sign AI-generated images produced by Microsoft Designer and Bing Image Creator.[59]

[57] C2PA Explainer :: C2PA Specifications
[58] https://verify.contentauthenticity.org/
[59] Microsoft Build brings AI tools to the forefront for developers - The Official Microsoft Blog

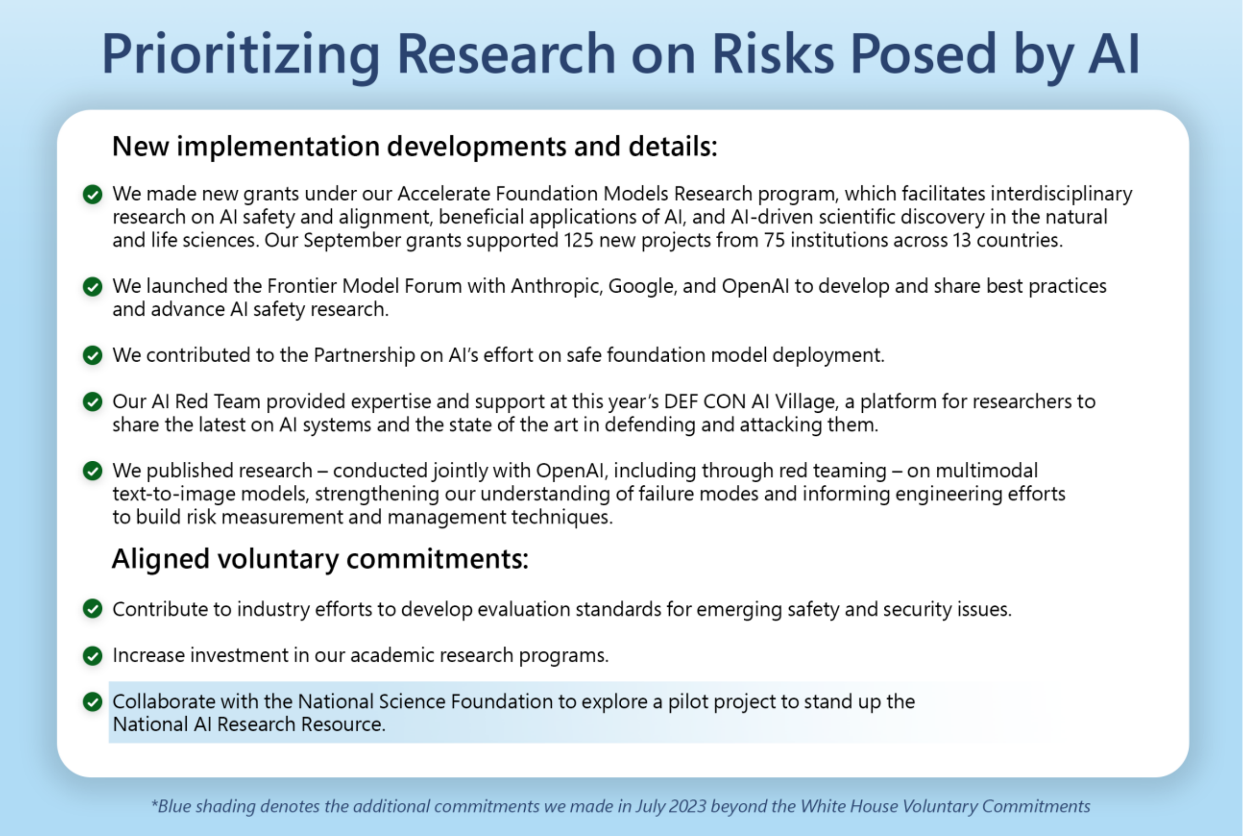
Bing Image Creator now discloses content as AI generated automatically. We are also advancing provenance capabilities in Azure OpenAI for internal product teams.

As part of our broader AI safety strategy of taking an iterative approach to risk management and driving continuous improvement, we are also investing in research and evaluation of techniques to enhance robustness. For example, as a potential layered mitigation beyond metadata-based provenance, we are also exploring a fingerprinting solution to help identify if an image was AI generated.

Microsoft and others are also continuing to invest in developing and promoting C2PA. For instance, while the specification to date can only be used with some digital assets, C2PA is continually expanding the standard to support new media formats. In April, the latest specification update added support for many new formats, including MPF, WebP, AIFF, AVI, and GIF.[60]

[60] C2PA Technical Specification :: C2PA Specifications

Prioritizing Research on Risks Posed by AI



Prioritizing Research on Risks Posed by AI

New implementation developments and details:

- ✓ We made new grants under our Accelerate Foundation Models Research program, which facilitates interdisciplinary research on AI safety and alignment, beneficial applications of AI, and AI-driven scientific discovery in the natural and life sciences. Our September grants supported 125 new projects from 75 institutions across 13 countries.
- ✓ We launched the Frontier Model Forum with Anthropic, Google, and OpenAI to develop and share best practices and advance AI safety research.
- ✓ We contributed to the Partnership on AI's effort on safe foundation model deployment.
- ✓ Our AI Red Team provided expertise and support at this year's DEF CON AI Village, a platform for researchers to share the latest on AI systems and the state of the art in defending and attacking them.
- ✓ We published research – conducted jointly with OpenAI, including through red teaming – on multimodal text-to-image models, strengthening our understanding of failure modes and informing engineering efforts to build risk measurement and management techniques.

Aligned voluntary commitments:

- ✓ Contribute to industry efforts to develop evaluation standards for emerging safety and security issues.
- ✓ Increase investment in our academic research programs.
- ✓ Collaborate with the National Science Foundation to explore a pilot project to stand up the National AI Research Resource.

**Blue shading denotes the additional commitments we made in July 2023 beyond the White House Voluntary Commitments*

Across multiple teams within Microsoft, we are investing in both internal and external efforts to accelerate research on AI safety, security, and societal impact and to increase access to AI resources. Internally, Microsoft Research significantly invests in AI, focusing efforts on 1) understanding general AI, taking inspiration from the study of human intelligence and from the prediction and observation of natural phenomena; 2) driving model innovation, in pursuit of more capable and aligned forms of AI; 3) ensuring societal benefit through trustworthy AI that supports human flourishing; 4) transforming scientific discovery, including with our recently established AI4Science organization; and 5) extending human capabilities, incubating novel AI applications in industries such as agriculture and healthcare.[61]

Microsoft Research is also expanding and diversifying its collaborators network and working to foster a vibrant global AI research community. Microsoft Research has recently launched Accelerate Foundation Models Research (AFMR), a research grant program through which we aim to facilitate interdisciplinary research on aligning AI with human goals, values, and preferences; improving human interactions via sociotechnical research; and accelerating scientistic discovery in natural sciences.[62] After managing a pilot phase that launched earlier this year, we expanded the program and have now selected 125 new projects from 75 institutions across 13 countries. The focus of our first open call for proposals was on aligning AI systems with human goals and preferences; advancing beneficial applications of AI; and accelerating scientific discovery in the natural and life sciences. As we continue to expand the breadth of

[61] AI and Microsoft Research - Microsoft Research

[62] Accelerate Foundation Models Research: Supporting a global academic research ecosystem for AI - Microsoft Research

our reach with academic partnerships, we will also continue to expand the depth of our research, including in areas like AI evaluation and measurement.

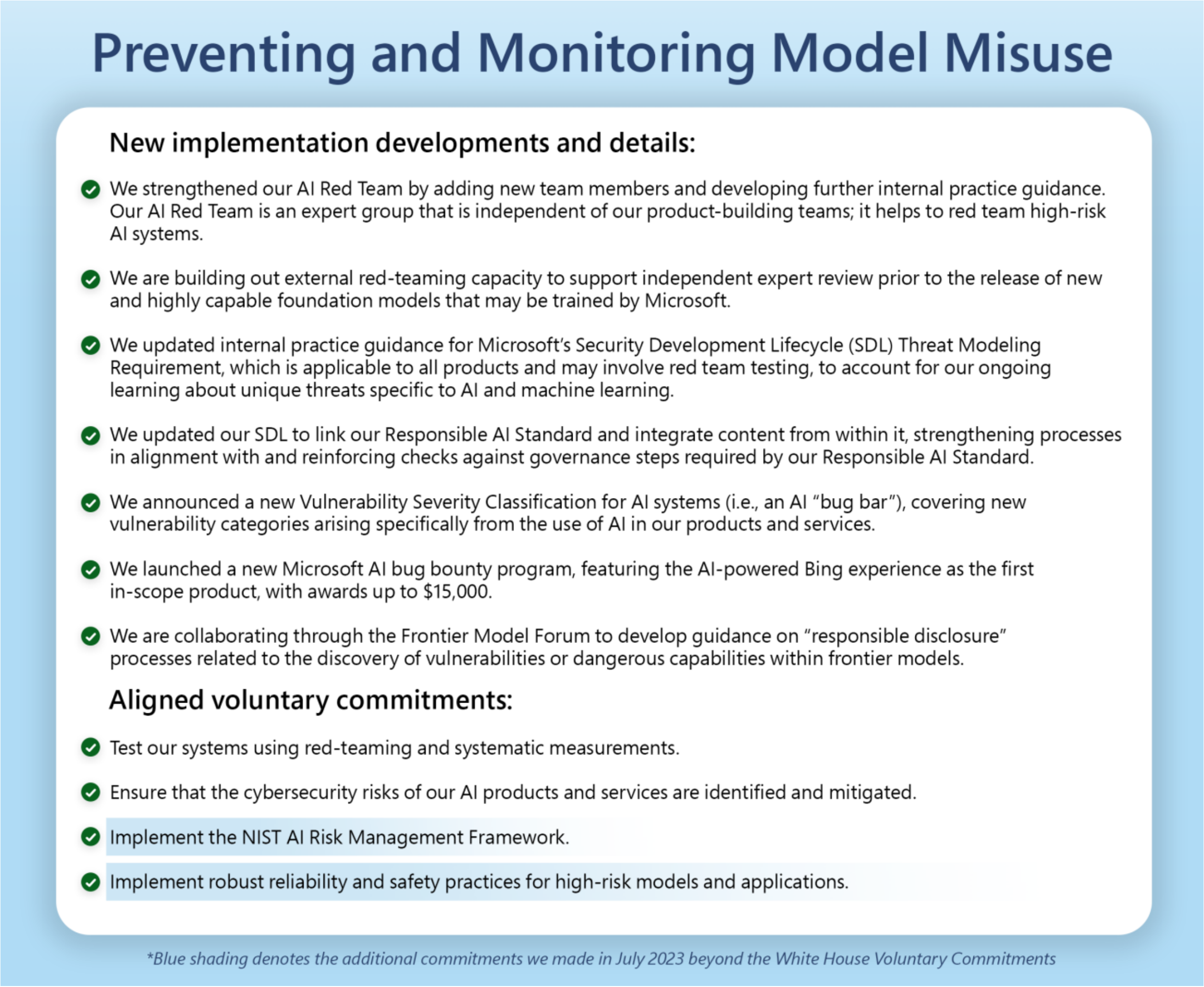
Microsoft Research and other teams also collaborate with researchers, practitioners, and other experts in industry and civil society organizations to advance knowledge of safety risks and best practices for safety and security techniques.

- In July, Microsoft worked with three other leading AI companies to launch the Frontier Model Forum, which is bringing together technical experts to define consensus best practices for the responsible development and deployment of frontier models.[63]
- In August, our AI Red Team, an expert group independent to our product groups, participated in this year's DEF CON AI Village, a platform for researchers to share the latest on AI systems and the state of the art in defending and attacking them.[64]
- In September, we released the results of a study, conducted jointly with OpenAI, to explore multimodal text-to-image models, including through red teaming to understand failure modes and engineering efforts to build measurement and mitigation techniques.[65]
- In October, Partnership on AI released for public comment guidance to which we contributed on safe foundation model deployment, aiming to provide tools and knowledge to foster responsible development and deployment of AI models with a focus on safety for society and adaptability to support evolving capabilities.[66] We've previously contributed to other Partnership on AI multi-stakeholder group efforts, including on the development of practices for AI-generated media disclosure.

We are also supportive of national and global efforts to establish AI computing resources for academic research. In May, Microsoft advocated not only for the establishment of a National AI Research Resource in the United States as introduced by 2020 legislation but also for an extension to accommodate access by academic institutions in allied nations abroad, including the European Union, Japan, the United Kingdom, and other like-minded countries.[67] An important complement to providing such access is the development of governance best practices for the academic community engaged in frontier research on applications and the safety and security of highly capable models, and Microsoft would also welcome the opportunity to help develop such practices within a collaborative multi-stakeholder group.

[63] Microsoft, Anthropic, Google, and OpenAI launch Frontier Model Forum - Microsoft On the Issues
[64] AI Village at DEF CON announces largest-ever public Generative AI Red Team - AI Village
[65] Frontiers of multimodal learning: A responsible AI approach - Microsoft Research
[66] Partnership on AI Releases Guidance for Safe Foundation Model Deployment, Takes the Lead to Drive Positive Outcomes and Help Inform AI Governance Ahead of AI Safety Summit in UK -
[67] Governing AI: A Blueprint for the Future (microsoft.com)

Preventing and Monitoring Model Misuse



Preventing and Monitoring Model Misuse

New implementation developments and details:

- We strengthened our AI Red Team by adding new team members and developing further internal practice guidance. Our AI Red Team is an expert group that is independent of our product-building teams; it helps to red team high-risk AI systems.
- We are building out external red-teaming capacity to support independent expert review prior to the release of new and highly capable foundation models that may be trained by Microsoft.
- We updated internal practice guidance for Microsoft's Security Development Lifecycle (SDL) Threat Modeling Requirement, which is applicable to all products and may involve red team testing, to account for our ongoing learning about unique threats specific to AI and machine learning.
- We updated our SDL to link our Responsible AI Standard and integrate content from within it, strengthening processes in alignment with and reinforcing checks against governance steps required by our Responsible AI Standard.
- We announced a new Vulnerability Severity Classification for AI systems (i.e., an AI "bug bar"), covering new vulnerability categories arising specifically from the use of AI in our products and services.
- We launched a new Microsoft AI bug bounty program, featuring the AI-powered Bing experience as the first in-scope product, with awards up to $15,000.
- We are collaborating through the Frontier Model Forum to develop guidance on "responsible disclosure" processes related to the discovery of vulnerabilities or dangerous capabilities within frontier models.

Aligned voluntary commitments:

- Test our systems using red-teaming and systematic measurements.
- Ensure that the cybersecurity risks of our AI products and services are identified and mitigated.
- Implement the NIST AI Risk Management Framework.
- Implement robust reliability and safety practices for high-risk models and applications.

**Blue shading denotes the additional commitments we made in July 2023 beyond the White House Voluntary Commitments*

At Microsoft, maintaining AI products that adhere to our responsible AI commitments throughout their product lifecycle means that they are subject to an iterative cycle of mapping, measuring, and managing risk pre and post deployment. This means that policies and practices across multiple areas of inquiry on which we have provided context in this update, including model evaluation and red teaming, security controls, and data input controls, must be implemented iteratively as appropriate. (Reference those sections for fuller context on our policies and practices for preventing and monitoring for model misuse.) Oversight of these ongoing processes through our robust governance structures is just as critical for product monitoring and maintenance – and continuous learning and improvement – as it is in advance of a product launch.

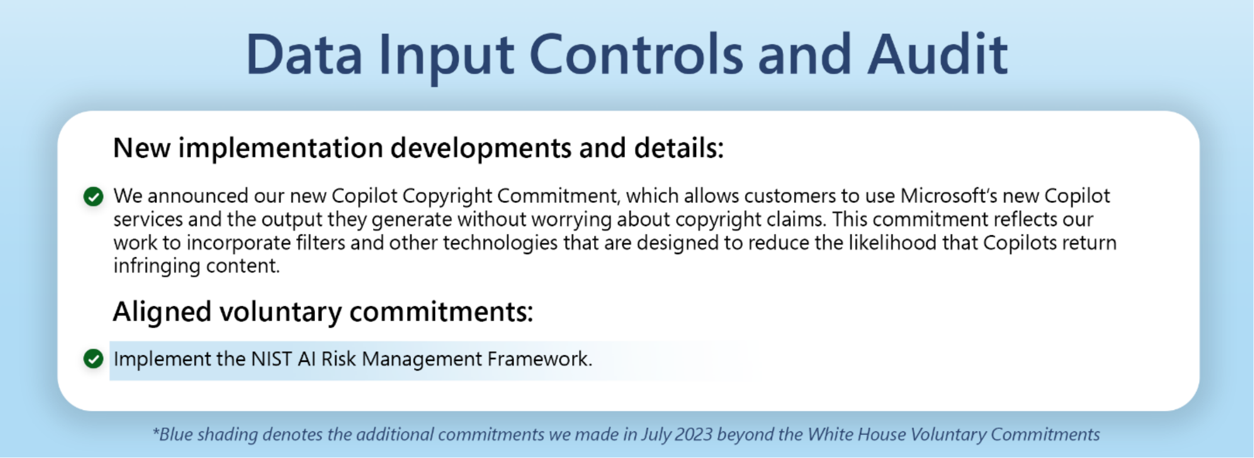
As we learn about new patterns of misuse or incidents, through our own ongoing research or internal red teaming or from external reports, we must be prepared to act to contain issues, enhance products, implement new mitigations, and otherwise protect customers. To strengthen readiness, we continuously research and engage with partners to improve mitigation techniques, and we regularly test and as necessary adjust plans to respond to incidents or detections of new potential patterns of misuse.

We also regularly assess and as necessary adjust our policies to strengthen security controls and customer transparency.

In addition, our AI products have built-in or add-on capabilities particularly focused on monitoring for patterns of misuse post deployment, and learnings feed back into product development to strengthen ongoing preventative efforts grounded in mapping, measuring, and mitigating practices (as described in detail in the Model Evaluations and Red Teaming section). For example, Azure OpenAI abuse monitoring detects and mitigates instances of recurring content and/or behaviors that suggest use of the service in a manner that may violate our code of conduct or other applicable product terms.[68]

[68] Data, privacy, and security for Azure OpenAI Service - Azure AI services | Microsoft Learn. To detect and mitigate abuse, Azure OpenAI stores all prompts and generated content security for up to 30 days (unless a customer is approved for and elects to configure abuse monitoring off, which requires meeting our Limited Access eligibility criteria and attesting to restricting use to specific use cases. Limited Access features for Azure AI services - Azure AI services | Microsoft Learn

Data Input Controls and Audit



Data Input Controls and Audit

New implementation developments and details:

✓ We announced our new Copilot Copyright Commitment, which allows customers to use Microsoft's new Copilot services and the output they generate without worrying about copyright claims. This commitment reflects our work to incorporate filters and other technologies that are designed to reduce the likelihood that Copilots return infringing content.

Aligned voluntary commitments:

✓ Implement the NIST AI Risk Management Framework.

**Blue shading denotes the additional commitments we made in July 2023 beyond the White House Voluntary Commitments*

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 impacting all our Responsible AI principles.[69] As a result, 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 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.[70] 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.[71] Established policies dictate that training data and fine-tuned models are available exclusively for use by the customer, are stored within the same region as the Azure OpenAI resource, 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.[72] All applicable data processed by AI products is also subject to Microsoft's General Data Protection Regulation (GDPR)[73] and other legal commitments and data privacy and security compliance offerings.[74]

Microsoft supports several existing technical mechanisms that content providers can use to restrict access to their data. This can be done by a number of technical measures, including putting content behind a paywall or using other means to technically restrict access. Content providers may also

[69] Microsoft Responsible AI Standard v2 General Requirements
[70] Data, privacy, and security for Azure OpenAI Service - Azure AI services | Microsoft Learn
[71] Data, privacy, and security for Azure OpenAI Service - Azure AI services | Microsoft Learn
[72] Data, privacy, and security for Azure OpenAI Service - Azure AI services | Microsoft Learn
[73] General Data Protection Regulation - Microsoft GDPR | Microsoft Learn
[74] Compliance offerings for Microsoft 365, Azure, and other Microsoft services. | Microsoft Learn

implement mechanisms to indicate that they do not intend the content that they make publicly accessible to be scaped by using machine-readable means, such as the robots.txt web standard.

Guardrails in our Copilots also help respect authors' copyrights.[75] We have incorporated filters and other technologies that are designed to reduce the likelihood that Copilots return infringing content. These build on and complement our work to protect digital safety, security, and privacy, based on a broad range of capabilities and techniques, many of which were introduced above, including classifiers, metaprompts, content filtering, and operational monitoring and abuse detection.

Just as we've leveraged our industry-leading compliance efforts to support cloud customers in meeting GDPR and other obligations, earlier this year, we announced commitments to inform and advance customers' responsible AI governance and compliance.[76] In September, we built on these commitments by announcing our new Copilot Copyright Commitment, which allows customers to use Microsoft's new Copilot services and the output they generate without worrying about copyright claims. Specifically, if a third party sues a commercial customer for copyright infringement for using one of Microsoft's Copilots or the output they generate, then we will defend the customer and pay the amount of any resulting adverse judgments or settlements as long as the customer has used the guardrails and content filters that we have built into our products.[77]

[75] Microsoft announces new Copilot Copyright Commitment for customers - Microsoft On the Issues
[76] Announcing Microsoft's AI Customer Commitments - The Official Microsoft Blog
[77] Microsoft announces new Copilot Copyright Commitment for customers - Microsoft On the Issues

Conclusion

Microsoft appreciates the opportunity to contribute to the UK's AI Safety Summit and to provide information in response to its inquiry about safety policies. Ongoing public-private dialogue is critical to rapidly advancing a shared understanding of effective practices and evaluation techniques. We look forward to the UK's next steps in convening the upcoming Summit, carrying forward efforts to strengthen global coordination of AI safety evaluations, and supporting greater international collaboration on research and codes of practice through the G7, Organization for Economic Co-operation and Development (OECD), and other multi-lateral and multi-stakeholder forums.