Microsoft

Responsible AI Transparency Report

How we build, support our customers, and grow

Phone/WhatsApp: +86 185 1567 8796

May 2024

Contents

Phone/WhatsApp: +86 185 7567 8796
+1 (346) 638-0369

Foreword

In 2016, our Chairman and CEO, Satya Nadella, set us on a clear course to adopt a principled and human-centered approach to our investments in Artificial Intelligence (AI). Since then, we have been hard at work to build products that align with our values. As we design, build, and release AI products, six values—transparency, accountability, fairness, inclusiveness, reliability and safety, and privacy and security—remain our foundation and guide our work every day.

To advance our transparency practices, in July 2023, we committed to publishing an annual report on our responsible AI program, taking a step that reached beyond the White House Voluntary Commitments that we and other leading AI companies agreed to. This is our inaugural report delivering on that commitment, and we are pleased to publish it on the heels of our first year of bringing generative AI products and experiences to creators, non-profits, governments, and enterprises around the world.

As a company at the forefront of AI research and technology, we are committed to sharing our practices with the public as they evolve. This report enables us to share our maturing practices, reflect on what we have learned, chart our goals, hold ourselves accountable, and earn the public's trust. We've been innovating in responsible AI for eight years, and as we evolve our program, we learn from our past to continually improve. We take very seriously our responsibility to not only secure our own knowledge but to also contribute to the growing corpus of public knowledge, to expand access to resources, and promote transparency in AI across the public, private, and non-profit sectors.

In this inaugural annual report, we provide insight into 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. First, we provide insights into our development process, exploring how we map, measure, and manage generative AI risks. Next, we offer case studies to illustrate how we apply our policies and processes to generative AI releases. We also share details about how we empower our customers as they build their own AI applications responsibly. Lastly, we highlight how the growth of our responsible AI community, our efforts to democratize the benefits of AI, and our work to facilitate AI research benefit society at large.

There is no finish line for responsible AI. And while this report doesn't have all the answers, we are committed to sharing our learnings early and often and engaging in a robust dialogue around responsible AI practices. We invite the public, private organizations, non-profits, and governing bodies to use this first transparency report to accelerate the incredible momentum in responsible AI we're already seeing around the world.

Brad Smith
Vice Chair & President

Natasha Crampton
Chief Responsible AI Officer

Phone/WhatsApp: +86 185 7567 8796
+1 (310) 638-0569

Key takeaways

In this report, we share how we build generative applications responsibly, how we make decisions about releasing our generative applications, how we support our customers as they build their own AI applications, and how we learn and evolve our responsible AI program.

These investments, internal and external, continue to move us toward our goal—developing and deploying safe, secure, and trustworthy AI applications that empower people.

We created a new approach for governing generative AI releases, which builds on our Responsible AI Standard and the National Institute of Standards and Technology's AI Risk Management Framework. This approach requires teams to map, measure, and manage risks for generative applications throughout their development cycle.

30

We've launched 30 responsible AI tools that include more than 100 features to support customers' responsible AI development.

33

We've published 33 Transparency Notes since 2019 to provide customers with detailed information about our platform services like Azure OpenAI Service.

We continue to participate in and learn from a variety of multi-stakeholder engagements in the broader responsible AI ecosystem including the Frontier Model Forum, the Partnership on AI, MITRE, and the National Institute of Standards and Technology.

We support AI research initiatives such as the National AI Research Resource and fund our own Accelerating Foundation Models Research and AI & Society Fellows programs. Our 24 Microsoft Research AI & Society Fellows represent countries in North America, Eastern Africa, Australia, Asia, and Europe.

16.6%

In the second half of 2023, we grew our responsible AI community from 350 members to over 400 members, a 16.6 percent increase.

99%

We've invested in mandatory training for all employees to increase the adoption of responsible AI practices. As of December 31, 2023, 99 percent of employees completed the responsible AI module in our annual Standards of Business Conduct training.

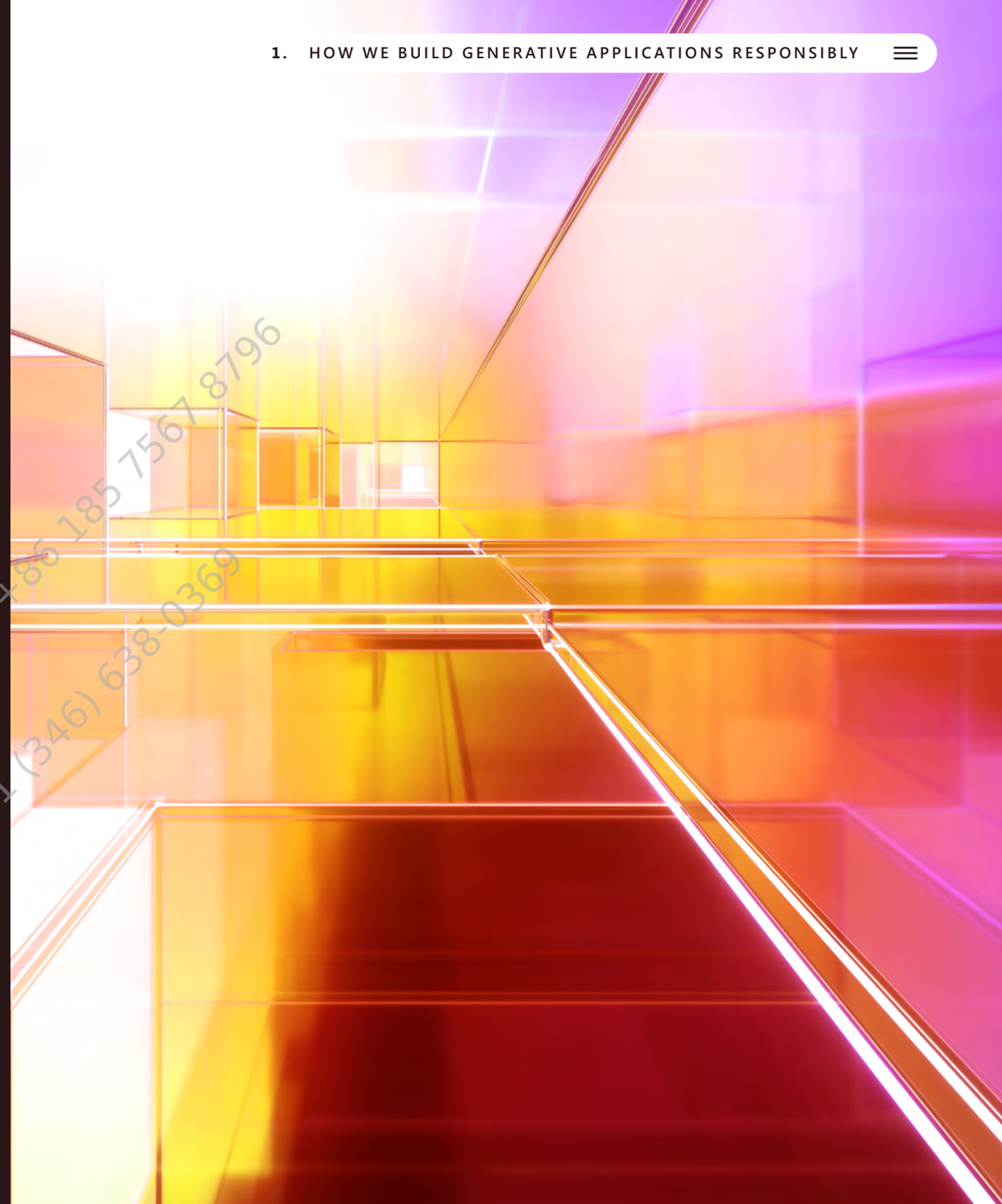
Phone/WhatsApp: +86 185 7567 8796
+1 (346) 638 0369

Section 1.

How we build generative applications responsibly

AI is poised to shape the future. Generative AI—artificial intelligence models and applications capable of creating original content, including text, image, and audio—has accelerated this transformation.

At Microsoft, we recognize our role in shaping this technology. We have released generative AI technology with appropriate safeguards at a scale and pace that few others have matched. This has enabled us to experiment, learn, and hone cutting-edge best practices for developing generative AI technologies responsibly. As always, we are committed to sharing our learnings as quickly as possible, and generative AI is no exception.



Phone/WhatsApp: +86 185 7567 8796
+1 (346) 638-0369

Govern: Policies, practices, and processes

In 2023, we regularly published resources to share best practices for developing generative applications responsibly. These include an overview of responsible AI practices for OpenAI models available through Azure OpenAI Service,[1] a Transparency Note[2] describing how to deploy Azure OpenAI models responsibly, examples relevant to generative AI in the HAX toolkit,[3] best practices[4] and a case study[5] for red teaming large language model (LLM) applications, and system message—or metaprompt—guidance.[6] In March 2024, we released additional tools our customers can use to develop generative applications more responsibly. This includes prompt shield to detect and block prompt injection attacks,[7] safety evaluation in Azure AI Studio to evaluate AI-generated outputs for content risks,[8] and risks & safety monitoring in Azure OpenAI Service to detect misuse of generative applications.[9]

In the following sections, we outline some of our recent innovations to map, measure, and manage risks associated with generative AI.

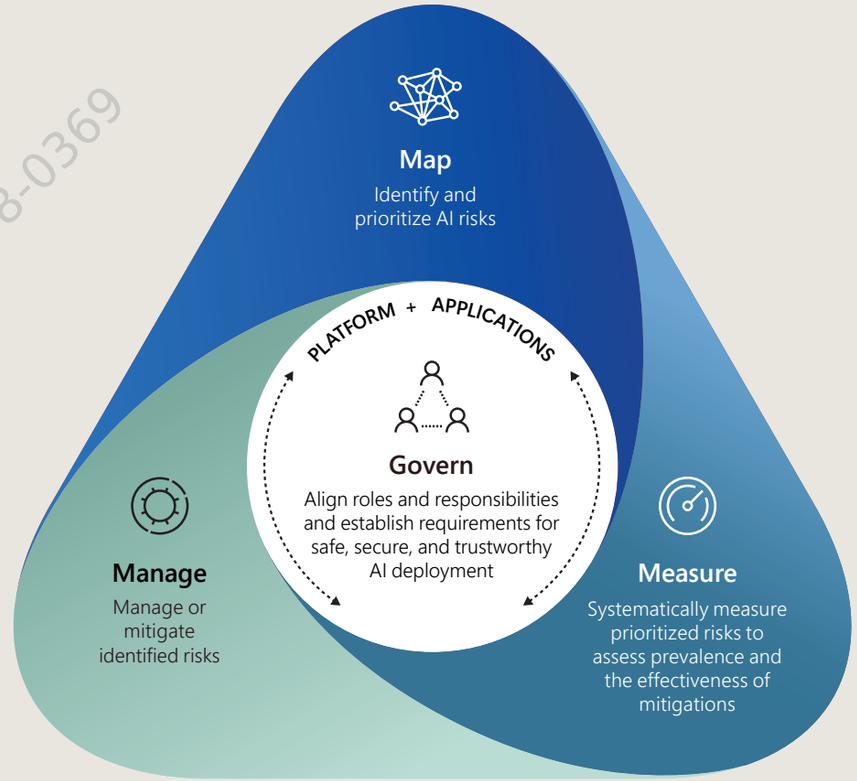
1. **First,** we cover specific requirements for generative applications, based on our Responsible AI Standard.
2. **Next,** we discuss how AI red teaming plays an important role in mapping generative AI risks at the model and application layers.
3. **Then,** we discuss the role of systematic measurement and how it provides metrics that inform decision making.
4. **Finally,** we describe some of our approaches to managing generative AI risks. This includes using technology to reinforce trust in democratic processes and manage generative AI's impact on the information ecosystem by implementing provenance tools to label AI-generated content.

Putting responsible AI into practice begins with our Responsible AI Standard. The Standard details how to integrate responsible AI into engineering teams, the AI development lifecycle, and tooling.

In 2023, we used our Responsible AI Standard to formalize a set of generative AI requirements, which follow a responsible AI development cycle. Our generative AI requirements align with the core functions of the National Institute for Standards and Technology (NIST) AI Risk Management Framework—govern, map, measure, and manage—with the aim of reducing generative AI risks and their associated harms.

Phone/WhatsApp: +86 185 7567 8796
+1 (346) 638-0369

Govern, map, measure, manage: An iterative cycle



Govern

Governance contextualizes the map, measure, and manage processes. We've implemented policies and practices to encourage a culture of risk management across the development cycle.

- **Policies and principles:** Our generative applications are designed to adhere to company policies, including our responsible AI, security, privacy, and data protection policies. We update these policies as needed, informed by regulatory developments and feedback from internal and external stakeholders.
- **Procedures for pre-trained models:** For the use of pre-trained generative AI models, teams must review available information about the model, its capabilities, and its limitations, then map, measure, and manage relevant risks.
- **Stakeholder coordination:** Our policies, programs, and best practices include input from a diverse group of internal and external stakeholders. Cross-functional teams work together to map, measure, and manage risks related to generative applications.
- **Documentation:** We provide transparency materials to customers and users that explain the capabilities and limitations of generative applications, as well as guidelines to help them use generative applications responsibly.
- **Pre-deployment reviews:** We require teams to map, measure, and manage generative AI risks pre-deployment and throughout their development cycle. This includes identifying high-impact uses of generative AI for additional review by experts within the company.

Map

Mapping risks is a critical first—and iterative—step toward measuring and managing risks associated with AI, including generative AI. Mapping informs decisions about planning, mitigations, and the appropriateness of a generative application for a given context.

- **Responsible AI Impact Assessments:** The development of generative applications begins with an impact assessment as required by the Responsible AI Standard. The impact assessment identifies potential risks and their associated harms as well as mitigations to address them.
- **Privacy and security reviews:** Processes for identifying and analyzing privacy and security risks, like security threat modeling, inform a holistic understanding of risks and mitigations for generative applications.
- **Red teaming:** We conduct red teaming of generative AI models and applications to develop a deeper understanding of how the identified risks manifest and to identify previously unknown risks.

Measure

We've implemented procedures to measure AI risks and related impacts to inform how we manage these considerations when developing and using generative applications.

- **Metrics for identified risks:** We have established metrics to measure identified risks for generative applications.
- **Mitigations performance testing:** We measure how effective mitigations are in addressing identified risks.

Manage

We manage or mitigate identified risks at the platform and application levels. We also work to safeguard against previously unknown risks by building ongoing performance monitoring, feedback channels, processes for incident response, and technical mechanisms for rolling applications back. Finally, we release and operate the application. We've learned that a controlled release to a limited number of users, followed by additional phased releases, helps us map, measure, and manage risks that emerge during use. As a result, we can be confident the application is behaving in the intended way before a wider audience accesses it.

- **User agency:** We design our generative applications to promote user agency and responsible use, such as through user interfaces that encourage users to edit and verify AI-generated outputs.
- **Transparency:** We disclose the role of generative AI in interactions with users and label AI-generated visual content.
- **Human review and oversight:** We design generative applications so that users can review outputs prior to use. Additionally, we notify users that the AI-generated outputs may contain inaccuracies and that they should take steps to verify information generated by the tool.
- **Managing content risks:** We build generative applications to address potential content risks, such as by incorporating content filters and processes to block problematic prompts and AI-generated outputs.
- **Ongoing monitoring:** Our teams also implement processes to monitor performance and collect user feedback to respond when our applications don't perform as expected.
- **Defense in depth:** We use an approach to risk management that puts controls at every layer of the process, including platform- and application-level mitigations.

We map, measure, and manage generative AI risks throughout the development cycle to reduce the risk of harm.

Because there is no finish line for responsible AI, our framework is iterative. Teams repeat processes to govern, map, measure, and manage AI-related risks throughout the product development and deployment cycle.

As we expand and evolve our responsible AI program, each new improvement builds on the foundation of the Responsible AI Standard. For example, we recently updated our Security Development Lifecycle (SDL) to integrate Responsible AI Standard governance steps. We also enhanced internal guidance for our SDL threat modeling requirement, which integrates ongoing learnings about unique threats specific to AI and machine learning (ML).[10] Incorporating responsible AI requirements into existing security guidance embodies our unified approach to developing and deploying AI responsibly.

Threat modeling is key to mapping potential vulnerabilities, enabling us to measure and manage risks, and closely evaluate the impacts of mitigations.

Evolving our cybersecurity development cycle in the new age of AI

We've developed and deployed technology using our state-of-the-art cybersecurity practices for decades. In our efforts to develop robust and secure AI infrastructure, we build on our extensive cybersecurity experience and work closely with cybersecurity teams from across the company. Our holistic approach is based on thorough governance to shield AI applications from potential cyberattacks across multiple vectors. Defense strategies include governance of AI security policies and practices; identification of potential risks in AI applications, data, and supply chains; protection of applications and information; detection of AI threats; and response and recovery from discovered AI issues and vulnerabilities, including through rapid containment. We take valuable learnings from these strategies, customer feedback, and external researcher engagement to continuously improve our AI security best practices.

All Microsoft products are subject to Security Development Lifecycle (SDL) practices and requirements.[11] Teams must execute threat modeling to map potential vulnerabilities, measure and manage risks, and closely evaluate the impacts of mitigations. Central engineering teams and our Digital Security and Resilience team facilitate and monitor SDL implementation to verify compliance and secure our products. Additional layers of security include centralized and distributed security engineering, physical security, and threat intelligence. Security operations teams drive implementation and enforcement.

We synthesize and organize learnings about AI threats into security frameworks, such as:

- **The Adversarial Machine Learning Threat Matrix,** which we developed with MITRE and others.[12]
- **Our Aether[13] Security Engineering Guidance,** which added AI-specific threat enumeration and mitigation guidance to existing SDL threat modeling practices.[14]
- **Our AI bug bar,** which provides a severity classification for vulnerabilities that commonly impact AI and ML applications.[15]

Further, we apply SDL protection, detection, and response requirements to AI technology. Specifically for our products that leverage pre-trained models, model weights are encrypted-at-rest and encrypted-in-transit to mitigate the potential risk of model theft. We apply more stringent security controls for high-risk technology, such as for protecting highly capable models. For example, in our AI product environments where highly capable proprietary AI models are deployed, we employ strong identity and access control. We also use holistic security monitoring (for both external and internal threats) with rapid incident response and continuous security validation (such as simulated attack path analysis).

Map: Identifying risks

As part of our overall approach to responsible development and deployment, we identify AI risks through threat modeling,[16] responsible AI impact assessments,[17] customer feedback, incident response and learning programs, external research, and AI red teaming. Here, we discuss our evolving practice of AI red teaming.

Red teaming, originally defined as simulating real-world attacks and exercising techniques that persistent threat actors might use, has long been a foundational security practice at Microsoft.[18] In 2018, we established our AI Red Team. This group of interdisciplinary experts dedicated to thinking like attackers and probing AI applications for failures[19] was the first dedicated AI red team in industry.[20] Recently, we expanded our red teaming practices to map risks outside of traditional security risks, including those associated with non-adversarial users and those associated with responsible AI, like the generation of stereotyping content. Today, the AI Red Team maps responsible AI and security risks at the model and application layers:

- **Red teaming models.** Red teaming the model helps to identify how a model can be misused, scope its capabilities, and understand its limitations. These insights not only guide the development of platform-level evaluations and mitigations for use of the model in applications but can also be used to inform future versions of the model.
- **Red teaming applications.** Application-level AI red teaming takes a system view, of which the base model is one part. This helps to identify failures beyond just the model, by including the application specific mitigations and safety system. Red teaming throughout AI product development can surface previously unknown risks, confirm whether potential risks materialize in an application, and inform measurement and risk management. The practice also helps clarify the scope of an AI application's capabilities and limitations, identify potential for misuse, and surface areas to investigate further.

For generative applications we characterize as high-risk, we implement processes to ensure consistent and holistic AI red teaming by experts independent from the product team developing the application. We are also building external red teaming capacity to enable third-party testing before releasing highly capable models, consistent with our White House Voluntary Commitments.[21] Externally led red teaming for highly capable models will cover particularly sensitive capabilities, including those related to biosecurity and cybersecurity.

In 2018

We established the first dedicated AI red team in industry.

Phone/WhatsApp: +86 185 7567 8796 +1 (346) 638-0369

Measure: Assessing risks and mitigations

After mapping risks, we use systematic measurement to evaluate application and mitigation performance against defined metrics. For example, we can measure the likelihood of our applications to generate identified content risks, the prevalence of those risks, and the efficacy of our mitigations in preventing those risks. We regularly broaden our measurement capabilities. Some examples include:[22]

- Groundedness, to measure how well an application's generated answers align with information from input sources.
- Relevance, to measure how directly pertinent a generated answer is to input prompts.
- Similarity, to measure the equivalence between information from input sources and a sentence generated by an application.
- Content risks, multiple metrics through which we measure an application's likelihood to produce hateful and unfair, violent, sexual, and self-harm related content.
- Jailbreak success rate, to measure an application's resiliency against direct and indirect prompt injection attacks that may lead to jailbreaks.

We also share capabilities and tools that support measurement of responsible AI concepts and development of new metrics. We share some of these tools as open source on GitHub and with our customers via Azure AI, which includes Azure Machine Learning and Azure AI Studio.

Azure AI Content Safety

uses advanced language and vision models to help detect content risks such as hateful, sexual, violent, or self-harm related content.

Safety evaluations in Azure AI Studio

Many generative applications are built on top of large language models, which can make mistakes, generate content risks, or expose applications to other types of attacks. While risk management approaches such as safety system messages and content filters are a great start, it's also crucial to evaluate applications to understand if the mitigations are performing as intended.

With Azure AI Studio safety evaluations, customers can evaluate the outputs of generative applications for content risks such as hateful, sexual, violent, or self-harm related content. Additionally, developers can evaluate their applications for security risks like jailbreaks. Since evaluations rely on a robust test dataset, Azure AI Studio can use prompt templates and an AI-assisted simulator to create adversarial AI-generated datasets to evaluate generative applications. This capacity harnesses learning and innovation from Microsoft Research, developed and honed to support the launch of our own first-party Copilots, and is now available to customers in Azure as part of our commitment to responsible innovation.

Phone (WhatsApp: +86 185 7567 8796 +1 (346) 638 0365

Manage: Mitigating AI risks

Once risks have been mapped and measured, they need to be managed. We evaluate and improve our generative AI products across two layers of the technology to provide a defense in depth approach to mitigating risks.

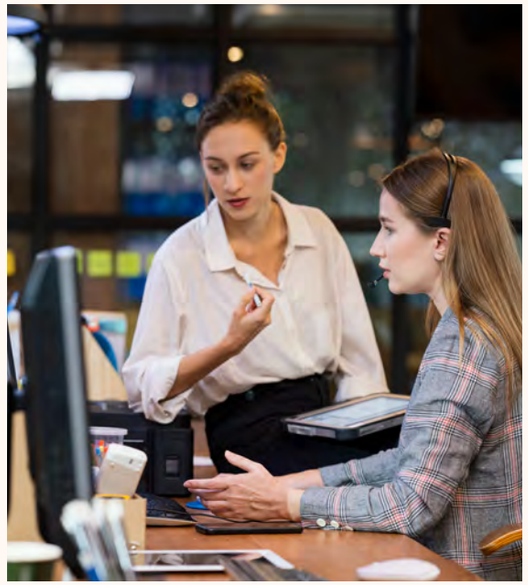
1. **Platform:** Based on the product's intended use, model-level mitigations can guide the application to avoid potential risks identified in the mapping phase. For example, teams can experiment with and fine-tune different versions of many generative AI models to see how potential risks surface differently in their intended use. This experimentation allows teams to choose the right model for their application. In addition, platform-level safety measures such as content classifiers reduce risks by blocking potentially harmful user inputs and AI-generated content. For example, Azure AI Content Safety provides API-level filters for content risks. Harmful user input or content risks generated by the AI model will be blocked when flagged by Azure AI Content Safety.

2. **Application:** A number of mitigations implemented in a specific application can also further manage risks. For example, grounding a model's outputs with input data alongside safety system messages to limit the model within certain parameters helps the application align with our responsible AI Standard and user expectations. For example, a safety system message guides Microsoft Copilot in Bing to respond in a helpful tone and cite its sources.[23] Additionally, user-centered design is an essential aspect of our approach to responsible AI. Communicating what the technology is and is not intended to do shows the application's potential, communicates its limitations, and helps prevent misuse. For example, we include in-product disclosures of AI-generated content in our Copilots, FAQs on responsible AI for our applications like GitHub Copilot,[24] and Transparency Notes for our platform products such as Azure OpenAI Service.[25]

As part of our commitment to build responsibly and help our customers do so as well, we integrate content filtering across Azure OpenAI Service.[26] We regularly assess our content filtering systems to improve accuracy and to ensure they're detecting as much relevant content as possible. Over the past year, we expanded our detection and filtering capabilities to include additional risk categories, such as jailbreaks, and improved the performance of our text, image, multimodal, and jailbreak models. These improvements rely on expert human annotators and linguists who evaluate offline evaluation sets. We also anonymously sample online traffic to monitor for regressions while leveraging the at-scale annotation capabilities of OpenAI's GPT-4.

Importantly, we've made these detection and evaluation tools available to our customers with the October 2023 general release of Azure AI Content Safety. Customers can choose to use our advanced language and vision models to help detect hate, violent, sexual, and self-harm related content, plus added jailbreak protections. When problematic content is detected, the models assign estimated severity scores to help customers efficiently tackle prioritized items and take action to reduce potential harm.[27] The models are offered in Azure AI Content Safety as standalone APIs, and customers can configure the filters to detect content with defined severity scores to fit their specific goals and policies.

The application of AI in our safety systems empowers organizations to monitor and support product safety at a scale that would be impossible for humans alone. These same tools are also offered in Azure AI Studio, Azure Open AI, and Azure AI Content safety where customers can discover, customize, and operationalize large foundation models at scale.



Phone/WhatsApp: +86 185 7567 8796
+1 (346) 638-0369

A new jailbreak risk detection model

Because generative AI models have advanced capabilities, they can be susceptible to adversarial inputs that can result in safety system bypass. These could provoke restricted behaviors and deviations from built-in safety instructions and system messages. This kind of adversarial technique is called a "jailbreak attack," also known as a user prompt injection attack (UPIA). In October 2023, to increase the safety of large language model deployments, we released a new jailbreak risk detection model, now called prompt shield. Prompt shield was integrated with existing comprehensive content safety filtering systems across Azure OpenAI Service and made available in Azure AI Content Safety as an API. When a jailbreak attempt is detected, customers can choose to take a variety of steps best suited for their application, such as further investigations or banning users.

Types of jailbreak attacks

Prompt shield recognizes four different classes of UPIA.

Category	Description
→ **Attempt to change application rules**	This category includes requests to use a new unrestricted application without rules, principles, or limitations, or requests instructing the application to ignore, forget, or disregard rules, instructions, or previous turns.
→ **Embedding a conversation mockup to confuse the model**	This attack takes user-crafted conversational turns embedded in a single user query to instruct the application to disregard rules and limitations.
→ **Role-play**	This attack instructs the application to act as another persona without application limitations or assigns anthropomorphic human qualities to the application, such as emotions, thoughts, and opinions.
→ **Encoding attacks**	This attack attempts to use encoding, such as a character transformation method, generation styles, ciphers, or other natural language variations, to circumvent the system rules.

In March 2024, prompt shield was expanded to include protections against indirect prompt injection attacks, where a generative application processes malicious information not directly authored by the application developer or the user, which can result in safety system bypass.[28]

Limited Access for customized service safety settings

Because safety is a priority for us, our Azure OpenAI Service is offered with default content safety settings and safeguards. Customers must complete registration under our Limited Access policy framework[29] and attest to approved use cases to gain access to Azure OpenAI Service. Customized settings for content filters and abuse monitoring are only allowed for approved use cases, and access to the broadest range of configurability is limited to managed customers. Managed customers are those who are working directly in trusted partnerships with Microsoft account teams. Managed customers must also attest their use cases for customized content filtering and abuse monitoring. All customers must follow the Azure OpenAI Service Code of Conduct,[30] which outlines mitigations and content requirements that apply to all customer uses to support safe deployment.

In the next section, we use a specific example of an identified risk—information integrity risks in the age of generative AI—to illustrate how we manage risks by combining technological advancements with policies and programs.

Managing information integrity risks in the age of generative AI

Amid growing concern that AI can make it easier to create and share disinformation, we recognize that it is imperative to give users a trusted experience. As generative AI technologies become more advanced and prevalent, it is increasingly difficult to identify AI-generated content. An image, video, or audio clip generated by AI can be indistinguishable from real-world capture of scenes by cameras and other human-created media. As more creators use generative AI technologies to assist their work, the line between synthetic content created by AI tools and human-created content will increasingly blur.

Labeling AI-generated content and disclosing when and how it was made (otherwise known as provenance) is one way to address this issue. In May 2023, we announced our intent to build new media provenance capabilities that use cryptographic methods to mark and sign AI-generated content with metadata about its source and history. Since then, we've made significant progress on our commitment to deploy new state-of-the-art tools to help the public identify AI-generated audio and visual content. By the end of 2023, we were automatically attaching provenance metadata to images generated with OpenAI's DALL-E 3 model in our Azure OpenAI Service, Microsoft Designer, and Microsoft Paint. This provenance metadata, referred to as Content Credentials, includes important information such as when the content was created and which organization certified the credentials.

To apply Content Credentials to our products' AI-generated images, we use an open technical standard developed by the Coalition for Content Provenance and Authenticity (C2PA), which we co-founded in 2021. The industry has increasingly adopted the C2PA standard, which requires cryptographic methods to sign, seal, and attach metadata to the file with a trusted identity certificate. This means C2PA Content Credentials can deliver a high level of trust with information that is tamper-evident while also preserving privacy. Certification authorities issue identity certificates to vetted organizations, and individual sources within those organizations can be anonymized. The C2PA coalition and standard body builds on our early efforts to prototype and develop provenance technologies and our collaboration with the provenance initiative Project Origin,[31] which we founded alongside the British Broadcasting Corporation, the Canadian Broadcasting Corporation, and the New York Times to secure trust in digital media.

Beyond Microsoft, we continue to advocate for increased industry adoption of the C2PA standard. There are now more than 100 industry members of C2PA. In February 2024, OpenAI announced that they would implement the C2PA standard for images generated by their DALL-E 3 image model. This is in addition to completing pre-deployment risk mapping and leveraging red-teaming practices to reduce potential for harm—an approach similar to ours.

While the industry is moving quickly to rally around the C2PA standard, relying on metadata-based provenance or even watermarking approaches alone will be insufficient. It is important to combine multiple methods, such as embedding invisible watermarks, alongside C2PA Content Credentials and fingerprinting, to help people recover provenance information when it becomes decoupled from its content. Additionally, authentication, verification, and other forensic technologies allow people to evaluate digital content for generative AI contributions. No disclosure method is foolproof, making a stacked mitigation approach especially important.

We continue to test and evaluate combinations of techniques in addition to new methods altogether to find effective provenance solutions for all media formats. For example, text can easily be edited, copied, and transferred between file formats, which interferes with current technical capabilities that attach Content Credentials to a file's metadata. We remain committed to investing in our own research, sharing our learnings, and collaborating with industry peers to address ongoing provenance concerns.

Phone/WhatsApp: +86 185 1567 0706
+1 (310) 628-0360

In May 2023

we announced our intent to build new media provenance capabilities that use cryptographic methods to mark and sign AI-generated content with metadata about its source and history.

End of 2023

we began to automatically apply Content Credentials to AI-generated images from Microsoft Designer, Microsoft Paint, and DALL-E 3 in our Azure OpenAI Service.

This work is especially important in 2024, a year in which more people will vote for their elected leaders than any year in human history. A record-breaking elections year combined with the fast pace of AI innovation may offer bad actors new opportunities to create deceptive AI content (also known as "deepfakes") designed to mislead the public. To address this risk, we worked with 19 other companies, including OpenAI, to announce the new Tech Accord to Combat Deceptive Use of AI in 2024 Elections at the Munich Security Conference in February 2024.[32] These commitments include advancing provenance technologies, innovating robust disclosure solutions, detecting and responding to deepfakes in elections, and fostering public awareness and resilience.

Since signing the Tech Accord, we continue to make progress on our commitments. We recently launched a portal for candidates to report deepfakes on our services.[33] And in March, we launched Microsoft Content Integrity tools in private preview, to help political candidates, campaigns, and elections organizations maintain greater control over their content and likeness. The Content Integrity tools include two components: first, a tool to certify digital content by adding Content Credentials, and second, tools to allow the public to check if a piece of digital content has Content Credentials.[34]

In addition to engaging in external research collaborations[35] and building technical mitigations, it's equally important to consider policies, programs, and investments in the broader ecosystem that can further manage information integrity risks associated with generative AI.

We know that false or misleading information is more likely to spread in areas where there is limited or no local journalism. A healthy media ecosystem acts as a virtual town square where people gather reliable information and engage on the most pressing issues facing society. We support independent journalism to advance free, open coverage of important issues on a local and national scale. Our Democracy Forward Journalism Initiative provides journalists and newsrooms with tools and technology to help build capacity, expand their reach and efficiency, distribute trustworthy content, and ultimately provide the information needed to sustain healthy democracies.[36]

In addition to the commitments made in the Tech Accord, we continue to build on our existing programs to protect elections and advance democratic values around the world.

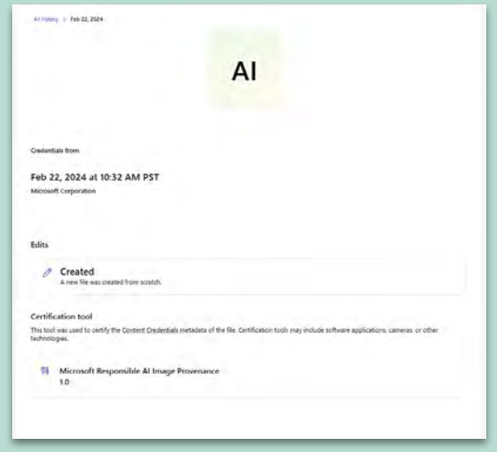
We support the rights of voters, candidates, political campaigns, and election authorities through a variety of programs and investments. These include our partnership with the National Association of State Election Directors, our endorsement of the Protect Elections from Deceptive AI Act in the United States, and our Elections Communications Hub.[37]

Case Study: Content Credentials for Microsoft Designer

Microsoft Designer allows users to input text prompts to generate images, such as the one below which was generated using the prompt: "people using an AI system for farming in a corn field."

Each image, like this one generated by Designer, is automatically marked and signed with Content Credentials, displayed in the right pane of the product interface. The Content Credentials indicate the date and time of creation using AI. Because Content Credentials are cryptographically signed and sealed as part of the image file's metadata, this information is tamper-evident and can be examined with tools such as Content Authenticity Initiative's open source Verify tool[38] and our Content Integrity Check tool.[39] Both tools are available as websites where users can upload files to check Content Credentials. For example, the image to the right shows that when examined with Verify or Check, Content Credentials for images generated by Designer indicate that they were created by a Microsoft product and the date of generation.

We continue to build provenance capabilities into our products, including most recently in



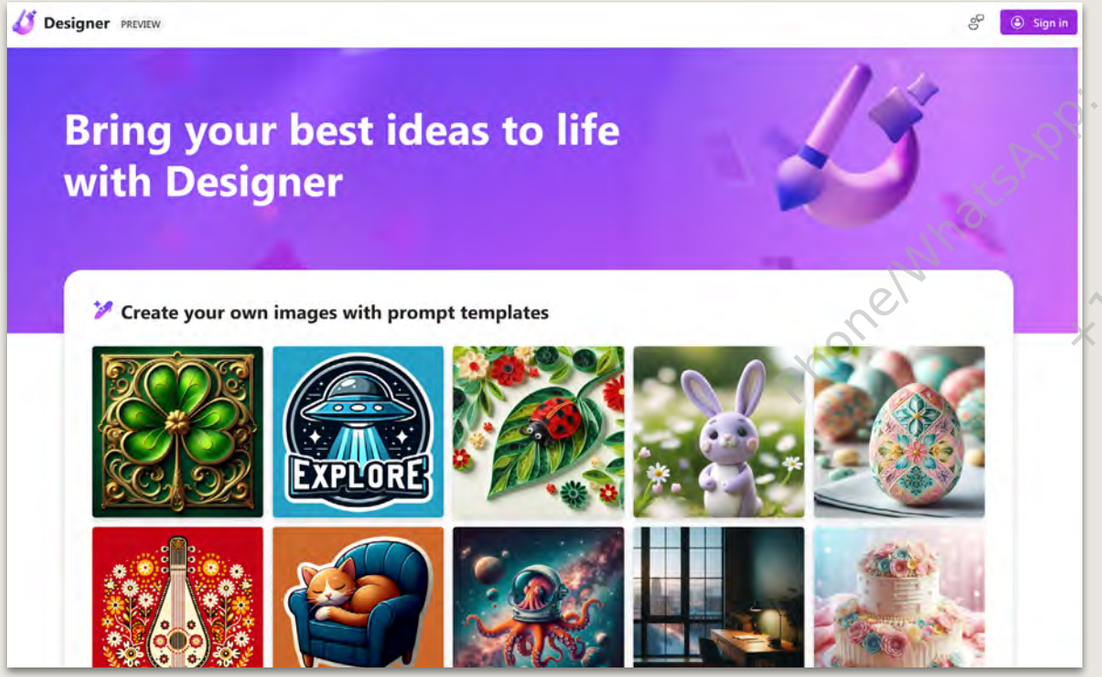
the DALL-E 3 series models hosted in Azure OpenAI Service. AI-generated images from Azure OpenAI Service using DALL-E 3 now include provenance information that attach source and generation date through Content Credentials.[40]

Page/WhatsApp: +86 185 7567 8796 +1 (346) 638-0369

Third-party evaluation of Microsoft Designer

The work of AI risk management cannot be done by companies alone. This is why we are committed to learning from stakeholders in academia, civil society, and government whose perspectives, evaluations, and concerns we consider as we build. Below is an example of how we've exercised this commitment through an external assessment of Microsoft Designer.

Designer is a general use text-to-image generative AI tool. Its many uses can make it vulnerable to adversarial use, including the generation of images that can be used for information operations. While we can't control what happens to images generated by our applications once they leave our platform, we can mitigate known risks at the user input and image output stages. This is why we've put in place safeguards to restrict what the application will generate, including deceptive images that could further information operations.



To better understand the risk of misleading images, reduce potential harms, and promote information integrity, we partnered with NewsGuard to evaluate Designer. NewsGuard is an organization of trained journalists that scores news sources for adherence to journalistic principles. As part of their analysis, NewsGuard prompted Designer to create visuals that reinforced or portrayed prominent false narratives related to politics, international affairs, and elections. Of the 275 images created:

- Mitigations worked in 88 percent of the attempts and the output images contained no problematic content.
- 12 percent of the output images contained problematic content.

To enhance performance related to information integrity, we regularly improve prompt blocking, content filters, and safety system message mitigations. Following the mitigation improvements, we input the same prompts developed by NewsGuard which had previously resulted in problematic content and reevaluated the images generated by Designer.

96.4%

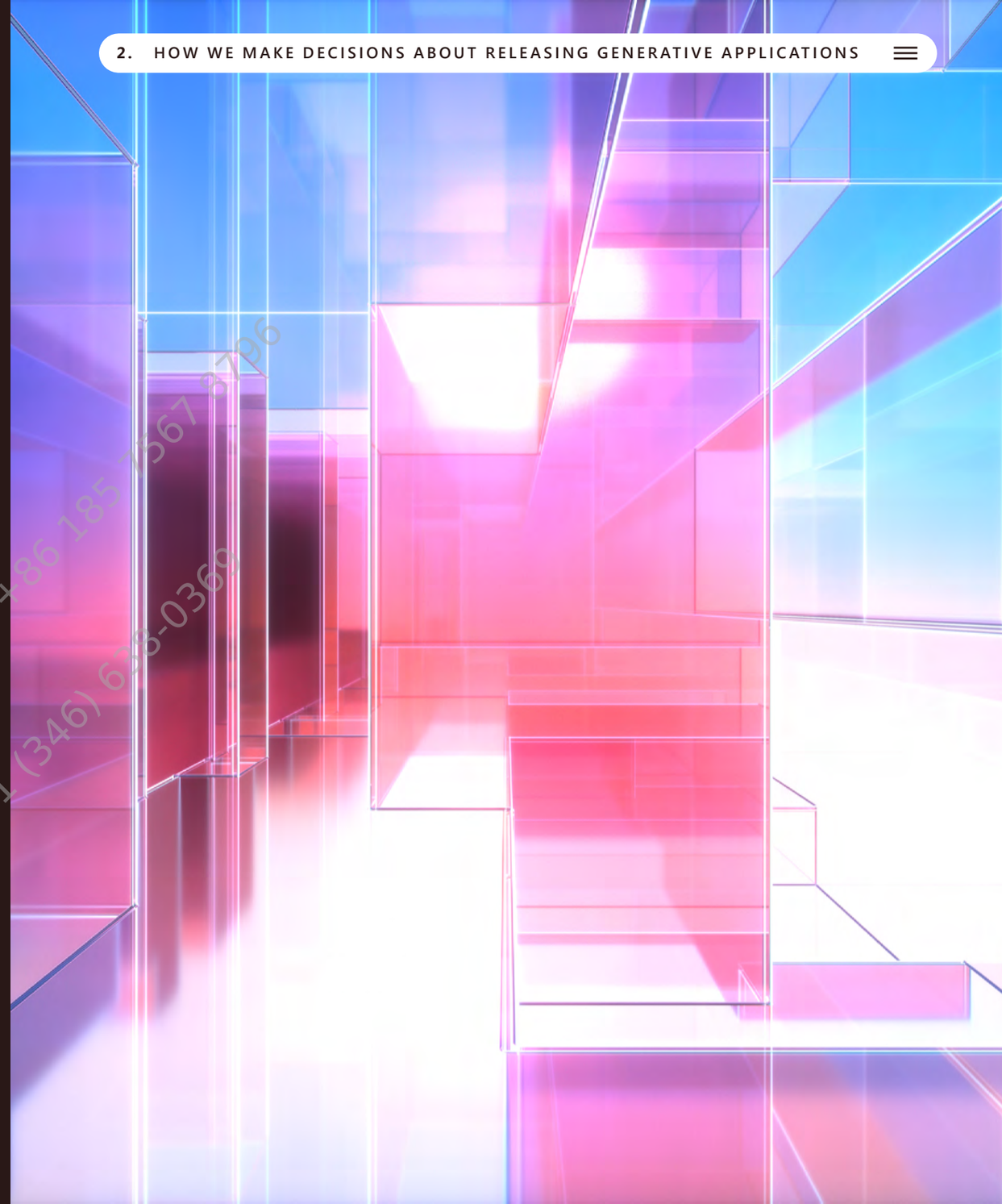
of test prompts successfully mitigated following improvements to the Designer safety system.

We found that only 3.6 percent of the output images contained problematic content.

NewsGuard's analysis and our mitigation improvements were steps in the right direction, but there is more work to be done. Evaluating and managing risks of our generative applications is an ongoing process and inherently iterative in nature. As new issues surface—identified by our internal responsible AI practices, external evaluations, and feedback submitted by users—we take action to address them. In addition, the question of how to build scaled measurement for evaluating information integrity in images is still an open research question. To address this challenge, we are excited to announce a new collaboration with researchers from Princeton's Empirical Studies and Conflict Project to advance this research.

Section 2.

How we make decisions about releasing generative applications



Phone/WhatsApp: +86 185 7367 3796
+1 (346) 638-0369

Deployment safety for generative AI applications

At each stage of the map, measure, and manage process for generative AI releases, we've built best practices, guidelines, and tools that reflect our learnings from the last year of releasing generative applications.

For example, when teams are asked to evaluate the potential for generative applications to produce ungrounded content, they are provided with centralized tools to measure that risk alongside patterns and best practices to guide their design of specific mitigations for their generative application.

After teams complete their initial analysis, senior experts review the evaluations and mitigations and make any further recommendations or requirements before products are launched. These reviews ensure that we apply a consistent and high bar to the design, build, and launch of our generative AI applications. When gaps are identified, these experts dive deep with product teams and leaders to assess the problems and agree on further mitigations and next steps. This oversight by senior leadership provides important touch points throughout a product's development cycle to manage risks across the company. This process improves each AI product and generates important lessons to apply to our map, measure, and manage approach.

As we learn more about how generative AI is used, we continue to iterate on our requirements, review processes, and best practices.

In this section, we share what we've learned through two examples. These short case studies demonstrate how we've improved our products through our work to map, measure, and manage risks associated with generative AI.

Our collaboration with OpenAI

While we often build and deploy state-of-the-art AI technology, we also partner with other companies building advanced AI models, including OpenAI.

Our most recent agreement with OpenAI[41] extends our ongoing collaborations in AI supercomputing and research. The agreement also allows both OpenAI and Microsoft to independently commercialize any advanced technologies developed by OpenAI. We deploy OpenAI's models across several consumer and enterprise products, including Azure OpenAI Service, which enables developers to build cutting-edge AI applications through OpenAI models with the benefit of Azure's enterprise-grade capabilities.

Microsoft and OpenAI share a commitment to building AI systems and products that are trustworthy and safe. OpenAI's leading research on AI Alignment,[42] their preparedness framework,[43] and our Responsible AI Standard establish the foundation for the safe deployment of our respective AI technologies and help guide the industry toward more responsible outcomes.

Phone/WhatsApp: +86 185 7567 8706
+1 (346) 638-0569

Case Study: Safely deploying Copilot Studio

In 2023, we released Copilot Studio, which harnesses generative AI to enable customers without programming or AI skills to build their own copilots.[44] Natural language processing enables the cloud-based platform to interpret customer prompts and create interactive solutions that customers can then deploy to their users. It also enables customers to test, publish, and track the performance of copilots within the platform so they remain in control of the experience. As with all generative applications, the Copilot Studio engineering team mapped, measured, and managed risks according to our governance framework prior to deployment.

Map. As part of their process, the engineering team mapped key risks associated with the product in their Responsible AI Impact Assessment as well as security and privacy reviews, including the potential for copilots to provide ungrounded responses to user prompts.

Measure and Manage. The Copilot Studio team worked with subject matter experts to measure and manage key risks iteratively throughout the development and deployment process. To mitigate AI-generated content risks, the Copilot Studio team included safety system message mitigations and leveraged Azure OpenAI Service's content filtering capabilities to direct copilots to generate only acceptable content. One of the key risks for this product is groundedness, or potential for AI-generated output to contain information that is not present in the input sources. By improving groundedness mitigations through metaprompt adjustments, the Copilot Studio team significantly enhanced in-domain query responses, increasing the in-domain pass rate from 88.6 percent to 95.7 percent. This means that when a user submits a question that is in-domain—or topically appropriate—copilots built with Copilot Studio are able to respond more accurately. This change also resulted in a notable 6 percent increase in answer rate within just one week of implementation. In other words, the improved groundedness filtering also reduced the number of queries that copilots declined to respond to, improving the overall user experience.

The team also introduced citations, so copilot users have more context on the source of information included in AI-generated outputs. By amending the safety system message and utilizing content filters, the Copilot Studio team improved citation accuracy from 85 percent to 90 percent.

Following the map, measure, and manage framework and supported by robust governance processes, the Copilot Studio team launched an experience where customers can build safer and more trustworthy copilots.

Case Study: Safely deploying GitHub Copilot

GitHub Copilot is an AI-powered tool designed to increase developer productivity through a variety of features, including code suggestions and chat experience to ask questions about code.[45] Code completion is a feature that runs in the integrated development environment (IDE), providing suggested lines of code as developers work on projects. GitHub Copilot Chat can be used in different environments, including the IDE and on GitHub.com, and provides a conversational interface for developers to ask coding questions. GitHub Copilot runs on a variety of advanced Microsoft and OpenAI technologies, including OpenAI's GPT models. In developing the features for GitHub Copilot, the team worked with their Responsible AI Champions—responsible AI experts within their organization—to map, measure, and manage risks associated with using generative AI in the context of coding.

Map. The team completed their Responsible AI Impact Assessment as well as security and privacy reviews to map different risks associated with the product. These risks included 1) the generation of code that may appear valid but may not be semantically or syntactically correct; 2) the generation of code that may not reflect the intent of the developer; and 3) more fundamentally, whether GitHub Copilot was actually increasing developer productivity. The last category, generally referred to as fitness for purpose, is an important concept for establishing that an AI application effectively addresses the problem it's meant to solve.

Measure. In addition to assessing performance, like the quality of responses, and running measurement sets to evaluate risks like insecure code or content risks, the GitHub Copilot team set out to understand if Copilot improved developer productivity. Research on how 450 developers at Accenture used the GitHub Copilot code completion feature over six months found that:[46]

- **94** percent of developers reported that using GitHub Copilot helped them remain in the flow and spend less effort on repetitive tasks.
- **90** percent of developers spent less time searching for information.
- **90** percent of developers reported writing better code with GitHub Copilot.
- **95** percent of developers learned from Copilot suggestions.

Case Study: Safely deploying GitHub Copilot, cont.

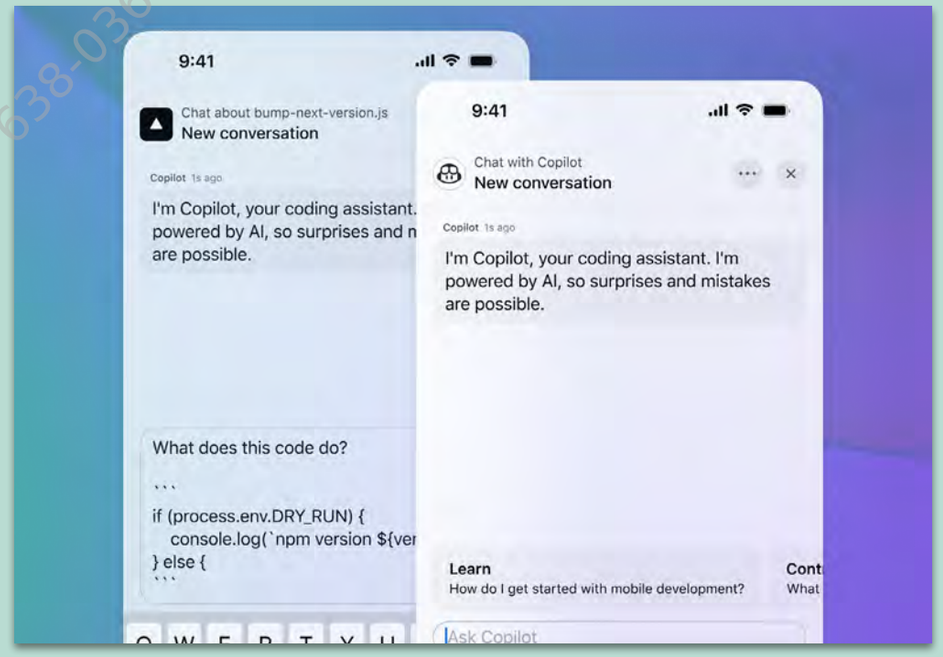
In a follow-on study of GitHub Copilot Chat, the team saw similar improvements in productivity.[47] In this study, researchers recruited 36 participants that had between five and ten years of coding experience. Participants were asked to a) write code, being randomly assigned to use or not use GitHub Copilot Chat and b) review code, being randomly assigned to review code that was authored with assistance from GitHub Copilot Chat or not. The researchers created a framework to evaluate code quality, asking participants to assess if code was readable, reusable, concise, maintainable, and resilient. In analyzing the data, researchers found that:

- **85** percent of developers felt more confident in their code quality when authoring code with GitHub Copilot and GitHub Copilot Chat.
- Code reviews were more actionable and completed **15** percent faster than without GitHub Copilot Chat.
- **88** percent of developers reported maintaining flow state with GitHub Copilot Chat because they felt more focused, less frustrated, and enjoyed coding more.

This research indicates that not only is GitHub Copilot making developers more productive, it also increases developer satisfaction.

Manage. As we've shared throughout the report, risks often need to be mitigated at multiple levels, and mitigations often work to manage multiple risks.

- **Human oversight.** Responsible AI is a shared responsibility, and our goal is to empower users to use GitHub Copilot in a safe, trustworthy, and reliable way. To support developer oversight of AI-generated code, GitHub Copilot was designed to offer suggested lines of code, which a developer reviews and accepts. In GitHub Copilot Chat, developers can review and copy code suggestions generated in the chat window into their coding environment. Importantly, the developer remains in control of the code they're writing.[48]
- **Staying on topic.** While the models behind GitHub Copilot can generate a wide range of content, one key approach to mitigate AI-generated content risks is to keep conversations limited to coding. The GitHub Copilot team built a classifier to reduce the number of off-topic conversations on the platform to keep conversations on topic and to protect users. In addition to the off-topic classifier, GitHub Copilot runs a variety of content filters, including to block content related to self-harm, violence, and hate speech.
- **Transparency.** The transparency documentation we provide on our GitHub Copilot features provides developers with important information about how best to use the features responsibly.[49,50] We also bake transparency right into the experience. GitHub Copilot discloses that code suggestions are AI-generated and may contain mistakes, empowering developers to make informed decisions about how best to use GitHub Copilot features.



Sensitive Uses program in the age of generative AI

The generative AI release process integrates with existing responsible AI programs and processes, such as our Sensitive Uses program, established in 2017 to provide ongoing review and oversight of high-impact and higher-risk uses of AI.

Employees across the company must report AI uses to our Sensitive Uses program for in-depth review and oversight if the reasonably foreseeable use or misuse of AI could have a consequential impact on an individual's legal status or life opportunities, present the risk of significant physical or psychological injury, or restrict, infringe upon, or undermine the ability to realize an individual's human rights. Particularly high-impact use cases are also brought before our Sensitive Uses Panel. Professionals from across our research, policy, and engineering organizations with expertise in human rights, social science, privacy, and security lend their expertise to the Sensitive Uses team and the Sensitive Uses Panel to help address complex sociotechnical issues and questions.

After review and consultation, the Sensitive Uses team delivers directed, concrete guidance and mitigation requirements tailored to the project.

Since 2019, the Sensitive Uses team has received over 900 submissions, including 300 in 2023 alone. In 2023, nearly 70 percent of cases were related to generative AI.

The increase in generative AI cases led to new insights about emerging risks, such as the capability of generative applications to make ungrounded inferences about a person. In some scenarios, our teams observed that a chatbot could provide realistic sounding but incorrect responses to questions that were outside the scope of their grounding data. Depending on the context, these ungrounded responses could misattribute actions or information about individuals or groups.

For example, a chatbot designed to answer questions about workplace benefits shouldn't answer questions about employee performance when that information is not included in the grounding data. Some of the mitigations that can prevent ungrounded inferences include safety system message provisions to guide which questions chatbots should respond to, ensuring that application responses are grounded in the right source data, and isolating private data. Once risks are assessed by the Sensitive Uses team, guidance is given to product teams on the mitigations for the use case.

Phone/WhatsApp: +86 185 7567 8796
+1 (346) 638-0369

900

submissions have been received by the Sensitive Uses team since 2019, including 300 in 2023 alone.

Nearly 70%

of cases in 2023 were related to generative AI.

Sensitive Uses in action: Microsoft Copilot for Security

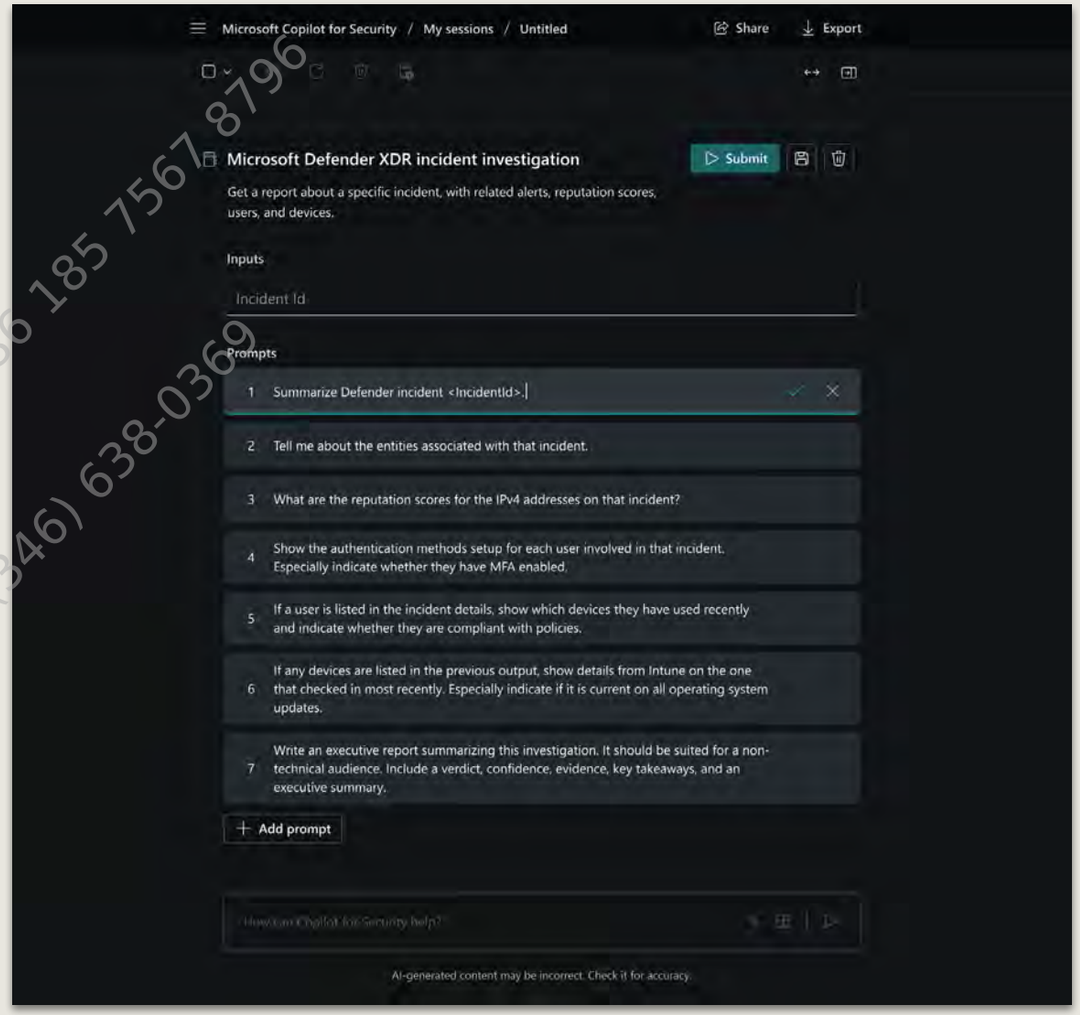
One example of a product that underwent Sensitive Uses review is Copilot for Security,[51] an AI-powered tool that helps security professionals respond to threats and assess risk exposure faster and more accurately. Copilot for Security uses generative AI to investigate analysts' digital environments, flag suspicious activity or content, and improve analysts' response to incidents. It generates natural language insights and recommendations from complex data, which helps analysts catch threats they may have otherwise missed and helps organizations potentially prevent and disrupt attacks at machine speed.

Through a Responsible AI Impact Assessment and with the support of their Responsible AI Champion, the Copilot for Security team identified that the project could meet the threshold for Sensitive Uses. They submitted the project to the Sensitive Uses program as part of early product development. The Sensitive Uses team confirmed that Microsoft Copilot for Security met the criteria for a Sensitive Uses review. They then worked with the product team and Responsible AI Champion to map key risks associated with the product, including that analysts could be exposed to potential content risks as part of their work.

The team landed on an innovative approach to address risks, which, due to the nature of routine security work, are different than those for consumer solutions. For example, security professionals may encounter offensive content or malicious code in source information. To allow analysts to stay in control of when they encounter potentially harmful content, the team made sure that Microsoft Copilot for Security surfaces these risks when requested by the security professionals. While Microsoft Copilot for Security suggests next steps, analysts ultimately decide what to do based on their organization's unique needs.

The Copilot for Security team worked closely with subject matter experts to validate their approach and specific mitigations. They improved in-product messaging to avoid overreliance on AI-generated outputs. They also refined metrics for grounding to improve the product's generated content. Through required ongoing monitoring of the product over the course of its phased releases, the team triaged and addressed responsible AI issues weekly.

This process led to a more secure, transparent, and trustworthy generative AI product that empowers security professionals to protect their organizations and customers, furthering our pursuit of the next generation of cybersecurity protection.[52]



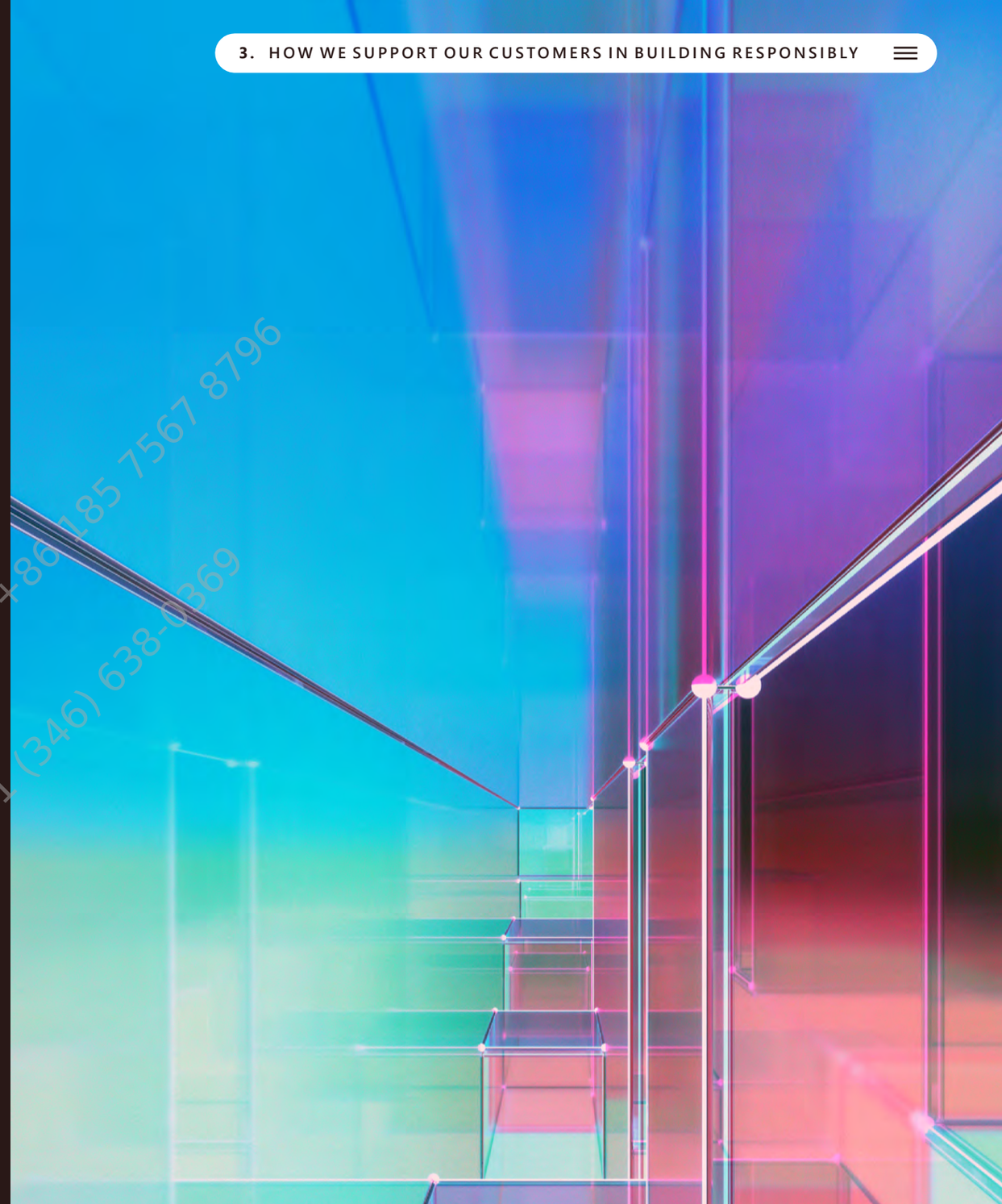
+1 (346) 638-0369 +86 185 7567 8796 WhatsApp

Section 3.

How we support our customers in building AI responsibly

In addition to building our own AI applications responsibly, we empower our customers with responsible AI tools and features. We invest in our customers' responsible AI goals in three ways:

1. We stand behind our customers' deployment and use of AI through our AI Customer Commitments.
2. We build responsible AI tools for our customers to use in developing their own AI applications responsibly.
3. We provide transparency documentation to customers to provide important information about our AI platforms and applications.



Phone/WhatsApp: +86 185 7567 8796
+1 (346) 638-0369

AI Customer Commitments

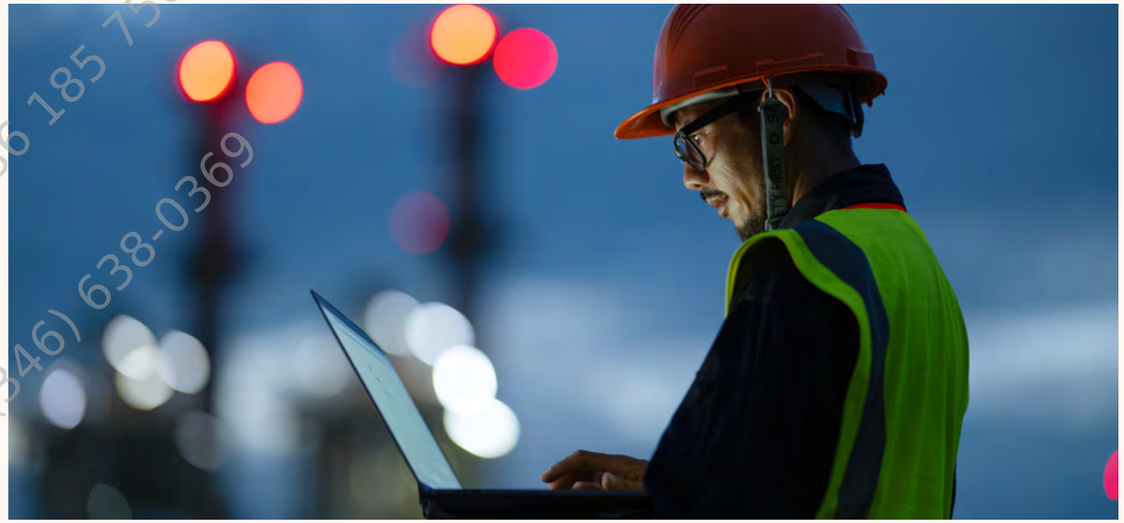
In June 2023, we announced our AI Customer Commitments,[53] outlining steps to support our customers on their responsible AI journey.

We recognize that ensuring the right guardrails for the responsible use of AI will not be limited to technology companies and governments. Every organization that creates or uses AI applications will need to develop and implement governance systems. We made the following promises to our customers:

- We created an AI Assurance Program to help customers ensure that the AI applications they deploy on our platforms meet the legal and regulatory requirements for responsible AI. This program includes regulator engagement support, along with our promise to attest to how we are implementing the NIST AI Risk Management Framework.
- We continue to engage with customer councils, listening to their views on how we can deliver the most relevant and compliant AI technology and tools.
- We created a Responsible AI Partner program for our partner ecosystem and 11 partners have joined the program so far. These partners have created comprehensive practices to help customers evaluate, test, adopt, and commercialize AI solutions.[54]
- We announced, and later expanded, the Customer Copyright Commitment[55] in which Microsoft will defend commercial customers who are sued by a third party for copyright infringement for using Azure OpenAI Service, our Copilots, or the outputs they generate and pay any resulting adverse judgments or settlements, as long as the customer met basic conditions such as not attempting to generate infringing content and using our required guardrails and content filters.[56]

Ultimately, we know that these commitments are an important start, and we will build on them as both the technology and regulatory conditions evolve. We are excited by this opportunity to partner more closely with our customers as we continue on our responsible AI journey together.

+1 (346) 638-0369 WhatsApp: +86 185 7567 8796



11

partners have joined since we created the Responsible AI Partner program.

Tools to support responsible development

To empower our customers, we've released 30 responsible AI tools that include more than 100 features to support customers' responsible AI development. These tools work to map and measure AI risks and manage identified risks with novel mitigations, real-time detection and filtering, and ongoing monitoring.

Tools to map and measure risks

We are committed to developing tools and resources that enable every organization to map, measure, and manage AI risks in their own applications. We've also prioritized making responsible AI tools open access. For example, in February 2024, we released a red teaming accelerator, Python Risk Identification Tool for generative AI (PyRIT).[57] PyRIT enables security professionals and machine learning engineers to proactively find risks in their generative applications. PyRIT accelerates a developer's work by expanding on their initial red teaming prompts, dynamically responding to AI-generated outputs to continue probing for content risks, and automatically scoring outputs using content filters. Since its release on GitHub, PyRIT has received 1,100 stars 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.

After identifying risks with a tool like PyRIT, customers can use safety evaluations in Azure AI Studio to conduct pre-deployment assessments of their generative application's susceptibility to generate low-quality or unsafe content, as well as to monitor trends post-deployment. For example, in November 2023 we released a limited set of generative AI evaluation tools in Azure AI Studio to allow customers to assess the quality and safety of their generative applications.[58] The first pre-built metrics offered customers an easy way to evaluate their applications for basic generation quality metrics such as groundedness, which measures how well the model's generated answers align with information from the input sources. In March 2024, we expanded our offerings in Azure AI Studio to include AI-assisted evaluations for safety risks across multiple content risk categories such as hate, violence, sexual, and self-harm, as well as content that may cause fairness harms and susceptibility to jailbreak attacks.[59]

Recognizing that evaluations are most effective when iterative and contextual, we've continued to invest in the Responsible AI Toolbox (RAI Toolbox).[60] This open-source tool, which is also integrated with Azure Machine Learning, offers support for computer vision and natural language processing (NLP) scenarios. The RAI Toolbox brings together a variety of model understanding and assessment tools such as fairness analysis, model interpretability, error analysis, what-if exploration, data explorations, and causal analysis. This enables ML professionals to easily flow through different stages of model debugging and decision-making. As an entirely customizable experience, the RAI Toolbox can be deployed for various functions, such as holistic model or data analysis, comparing datasets, or explaining individual instances of model predictions. On GitHub, the RAI Toolbox has received 1,200 stars with more than 4,700 downloads per month.

Tools to manage risks

Just as we measure and manage AI risks across the platform and application layers of our generative products, we empower our customers to do the same. For example, Azure AI Content Safety helps customers detect and filter harmful user inputs and AI-generated content in their applications. Importantly, Azure AI Content Safety provides options to detect content risks along multiple categories and severity levels to enable customers to configure settings to fit specific needs. Another example is our system message framework and templates, which support customers as they write effective system messages—sometimes called metaprompts—which can improve performance, align generative application behavior with customer expectations, and help mitigate risks in our customers' applications.[61]

In October 2023, we made Azure AI Content Safety generally available. Since then, we've continued to expand its integration across our customer offerings, including its availability in Azure AI Studio, a developer platform designed to simplify generative application development, and across our Copilot builder platforms, such as Microsoft Copilot Studio. We continue to expand customer access to additional risk management tools that detect risks unique to generative AI models and applications, such as prompt shield and groundedness detection. Prompt shield detects and blocks prompt injection attacks, which bad actors use to insert harmful instructions into the data processed by large language models.[62] Groundedness detection finds ungrounded statements in AI-generated outputs and allows the customer to implement mitigations such as triggering rewrites of ungrounded statements.[63]

In March 2024, we released risks & safety monitoring in Azure OpenAI Service, which provides tools for real-time harmful content detection and mitigation, offering insights into content filter performance on actual customer traffic and identifying users who may be abusing a generative application.[64] Customers can use these insights to fine-tune content filters to align with their safety goals. Additionally, the potentially abusive user detection feature

analyzes trends in user behavior and flagged content to generate reports for our customers to decide whether to take further action in Azure AI Studio. The report includes a user ranking and an abuse report, enabling customers to take action when abuse is suspected.

As we continue to improve our tools to map, measure, and manage generative AI risks, we make those tools available to our customers to enable an ecosystem of responsible AI development and deployment.

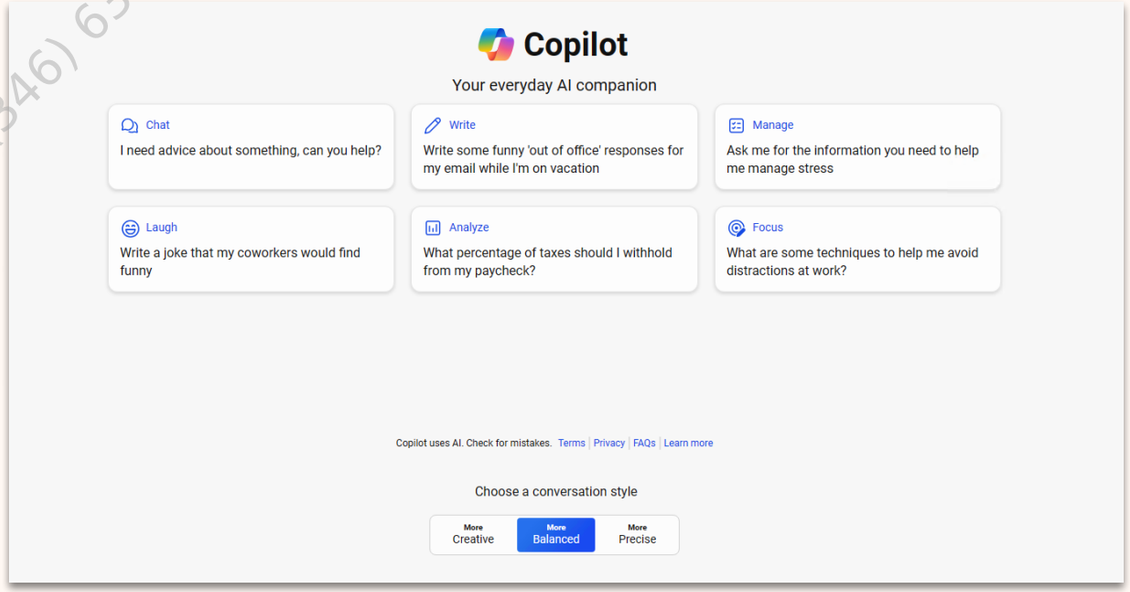
Transparency to support responsible development and use by our customers

Beginning in 2019, we've regularly released Transparency Notes—documentation covering responsible AI topics—for our platform services which customers use to build their own AI applications.

Since then, we've published 33 Transparency Notes. Required for our platform services, these follow a specific template to provide customers with detailed information about capabilities, limitations, and intended uses to enable responsible integration and use. Some examples include Transparency Notes for Azure AI Vision Face API,[65] Azure OpenAI Service,[66] and Azure Document Intelligence.[67]

In 2023, we expanded our transparency documentation beyond Transparency Notes. We now require our non-platform services, such as our Copilots, to publish Responsible AI Frequently Asked Questions (FAQs) and include user-friendly notices in product experiences to provide important disclosures. For example, Copilot in Bing provides users with responsible AI documentation[68] and FAQs[69] that detail our risk mapping, measurement, and management methods. In addition, 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. These user-friendly transparency documents and product integrated notices are especially important in our Copilot experiences, where users are less likely to be developers.

Transparency documentation and in-product transparency work together to enable our customers to build and use AI applications responsibly. And as with our other responsible AI programs, we anticipate that the ways we provide transparency for specific products will evolve as we learn.



Phone/WhatsApp: +86 185 1567 0796
+1 (346) 638-0360

Section 4.

How we learn, evolve, and grow

As we've prioritized our company-wide investments in responsible AI over the last eight years, people remain at the center of our progress. From our growing internal community to the global responsible AI ecosystem, the individuals and communities involved continue to push forward what's possible in developing AI applications responsibly. In this section, we share our approach to learning, evolving, and growing by bringing outside perspectives in, sharing learnings outwards, and investing in our community.

Phone/WhatsApp: +86 135 7567 8796
+1 (346) 638-0363

Governance of responsible AI at Microsoft: Growing our responsible AI community

At Microsoft, no one team or organization can be solely responsible for embracing and enforcing the adoption of responsible AI practices.

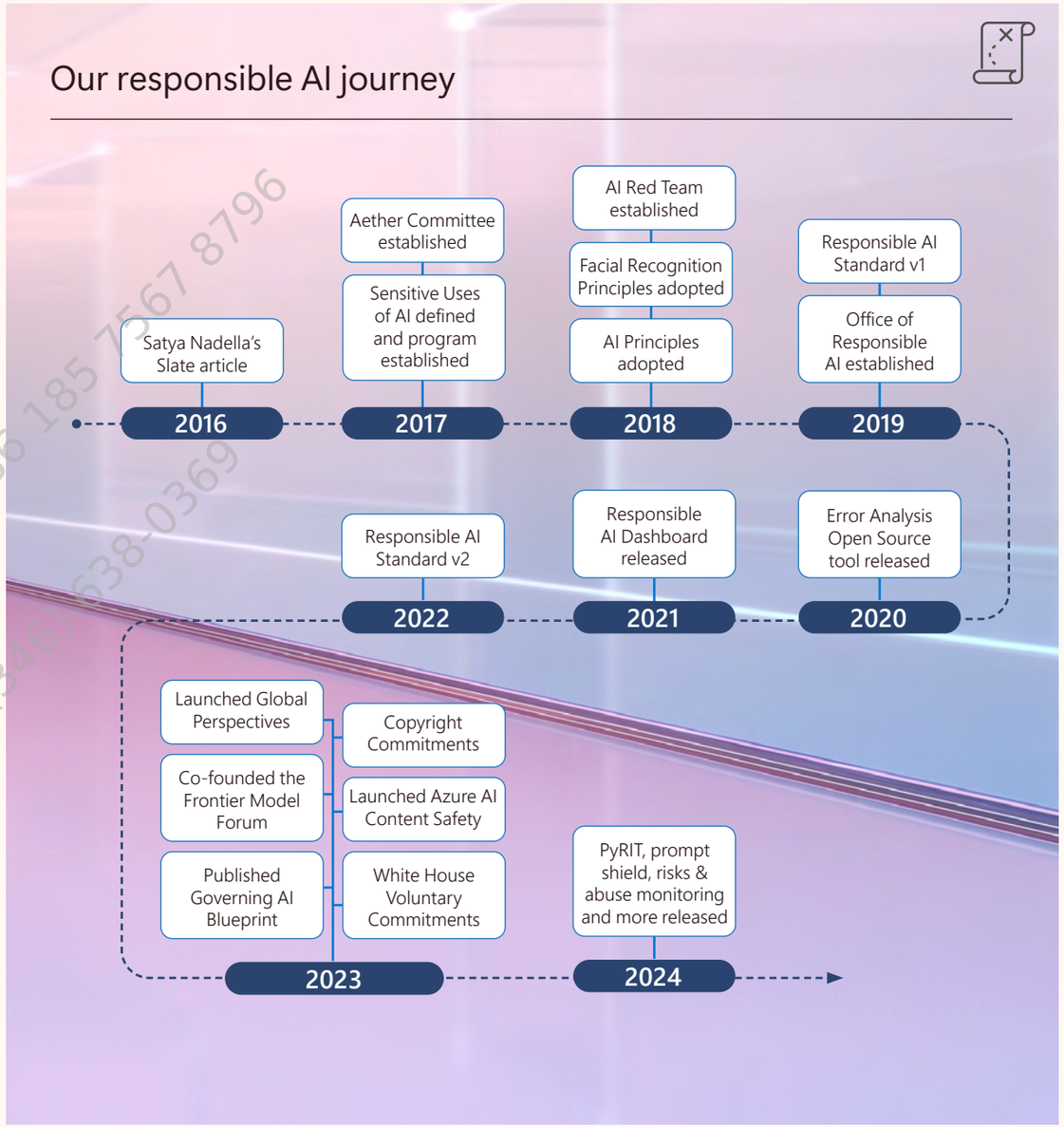
Rather, everyone across every level of the company must adhere to these commitments in order for them to be effective. We developed our Responsible AI Standard to communicate requirements and guidance so all teams can uphold our AI principles as they develop AI applications.

Specialists in research, policy, and engineering combine their expertise and collaborate on cutting-edge responsible AI practices. These practices ensure we meet our own commitments while also supporting our customers and partners as they work to build their own AI applications responsibly.

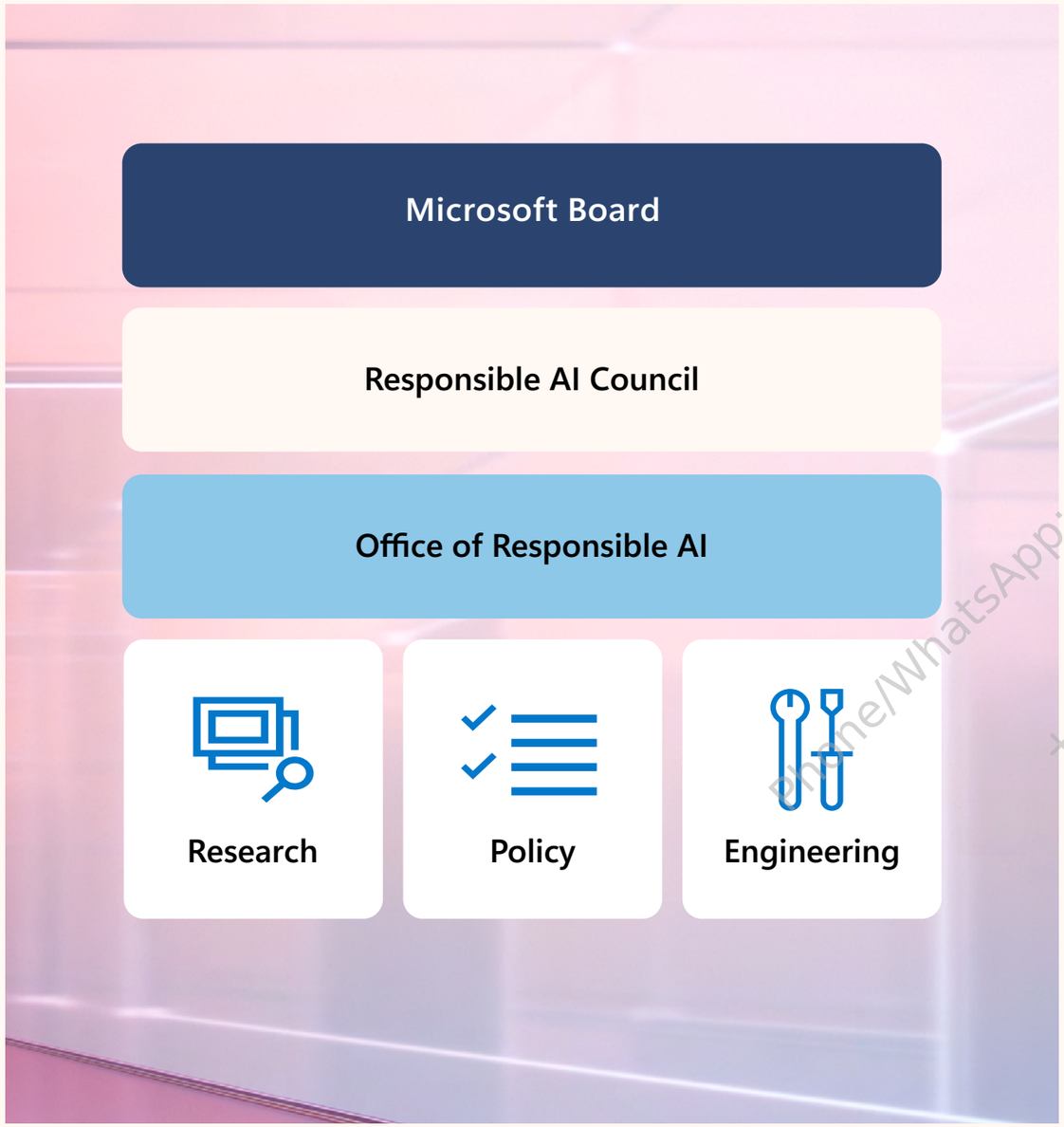
- **Research:** Researchers in Aether,[70] Microsoft Research,[71] and our engineering teams keep the responsible AI program on the leading edge of issues through thought leadership. They conduct rigorous AI research, including on transparency, fairness, human-AI collaboration, privacy, security, safety, and the impact of AI on people and society. Our researchers actively participate in broader discussions and debates to ensure that our responsible AI program integrates big-picture perspectives and input.

- **Policy:** The Office of Responsible AI (ORA) collaborates with stakeholders and policy teams across the company to develop policies and practices to uphold our AI principles when building AI applications. ORA defines roles and responsibilities, establishes governance systems, and leads Sensitive Use reviews to help ensure our AI principles are upheld in our development and deployment work. ORA also helps to shape the new laws, norms, and standards needed to ensure that the promise of AI technology is realized for the benefit of society at large.

- **Engineering:** Engineering teams create AI platforms, applications, and tools. They provide feedback to ensure policies and practices are technically feasible, innovate novel practices and new technologies, and scale responsible AI practices throughout the company. Our engineering teams draw on interactions with customers and user research to address stakeholder concerns in the development of our AI applications.



Phone/WhatsApp: +86 185 7567 8796
+1 (346) 638-0369



Applying lessons from previous efforts to address privacy, security, and accessibility, we've built a dedicated responsible AI program to guide our company-wide efforts.[72] We combine a federated, bottom-up approach with strong top-down support and oversight by company leadership to fuel our policies, governance, and processes. From a governance perspective, the Environmental, Social, and Public Policy Committee of the Board of Directors provides oversight and guidance on responsible AI policies and programs. Our management of responsible AI starts with CEO Satya Nadella and cascades across the senior leadership team and all of Microsoft. At the senior leadership level, the Responsible AI Council provides a forum for business leaders and representatives from research, policy, and engineering. The council, co-led by Vice Chair and President Brad Smith and Chief Technology Officer Kevin Scott, meets regularly to grapple with the biggest challenges surrounding AI and to drive progress in our responsible AI policies and processes. Executive leadership and accountability are key drivers to ensure that responsible AI remains a priority across the company.

At the community level, we've nurtured a unique Responsible AI Champion program that engages our engineering and global field teams in our responsible AI work. The Responsible AI Champion program is guided by a defined structure, with clear roles and responsibilities that empower our Champions and enable a culture of responsible AI across the company. Senior leaders accountable for responsible AI identify members of their organization to serve as Responsible AI Champions. These Champions enable their organizations to carry out our AI commitments by working together and learning from one another's expertise. This company-wide network troubleshoots problems, offers guidance, and advises on how to implement the Responsible AI Standard.

Our combined bottom-up and top-down approach empowers individuals, teams, and organizations and facilitates a culture of responsible AI by design. The collaborative and multidisciplinary structure embedded in our responsible AI program leverages the incredible diversity of the company[73] and amplifies what we can achieve. Our engineering, policy, and research teams bring a wealth of passion, experience, and expertise, which enables us to develop and deploy safe, secure, and trustworthy AI.

A dedicated responsible AI program

Our combined bottom-up and top-down approach empowers individuals, teams, and organizations and facilitates a culture of responsible AI by design.

Phone/WhatsApp: +86 185 7567 8796 +1 (346) 638-0369

Growing a responsible AI community

We strongly believe that AI has the potential to create ripples of positive impact across the globe. Over the years, we have matched that belief with significant investments in new engineering systems, research-led incubations, and, of course, people. We continue to grow and now have over 400 people working on responsible AI, more than half of whom focus on responsible AI full-time. In the second half of 2023, we grew our responsible AI community 16.6 percent across the company. We increased the number of Responsible AI Champions across our engineering groups and grew the number of full-time employees who work on centralized responsible AI infrastructure, AI red teaming, and assessing launch readiness of our products.

We continue to grow a diverse community to fulfill our commitments to responsible AI and to positively impact the products and tools that millions of people use every day. Some responsible AI community members provide direct subject matter expertise, while others build out responsible AI practices and compliance motions. Our community members hold positions in research, policy, engineering, sales, and other core functions, touching all aspects of our business. They bring varied perspectives rooted in their diverse professional and academic backgrounds, including liberal arts, computer science, international relations, linguistics, cognitive neuroscience, physics, and more.

11,000

attendees welcomed at SkillUp AI events.

30,000

employees reached through more than 40 events by our AI/ML connected community.

Supporting our responsible AI community through training

We have an enormous opportunity to integrate AI throughout the products and services we offer—and are dedicated to doing so responsibly. This journey begins with educating all of our employees.

The 2023 version of our Standards of Business Conduct training, a business ethics course required companywide, covers the resources our employees use to develop and deploy AI safely. As of December 31, 2023, 99 percent of all employees completed this course, including the responsible AI module.

For our responsible AI community members, our training goes even deeper. We provide extensive training for our over 140 Responsible AI Champions, more than 50 of whom joined the program in 2023. In turn, Responsible AI Champions help scale the implementation of responsible AI practices by training other members of the responsible AI community in their respective Microsoft organizations. This cascade strengthens peer relationships, builds trust among employees, and enables Champions to customize instruction to their specific organizations or divisions.

We continue to refine our training to keep pace with rapid developments in AI, particularly for responsible AI-focused professionals. We've developed learning sessions and resources for educating employees on responsible AI skills, our internal processes, and our approach to responsible AI. Ongoing education helps to keep our responsible AI subject matter experts current so they can disseminate up-to-date best practices throughout the company.

We also provide training on general AI topics so our employees can improve their knowledge and abilities as AI becomes more important for both our society and our business.

Throughout 2023, more than 100,000 employees attended conferences and events such as the AI/ML Learning Series and Hackathon, which has incubated more than 11,000 AI-focused projects. At these events, they learn the latest technologies and ways to apply responsible AI principles through company communities and channels. Our employees also lead by sharing their experiences and expertise. For example, our AI/ML connected community. reached nearly 30,000 employees through more than 40 events in 2023, and our SkillUp AI events welcomed more than 11,000 attendees.

Phone/WhatsApp: +86 185 7567 8706
+1 (346) 638-0369

Building safe and responsible frontier models through partnerships and stakeholder input

AI sits at the exciting intersection of technological breakthrough and real-world application. We are continually discovering new ways to push the limits with AI, innovating solutions to address society's biggest problems. Frontier models, highly capable AI models that go beyond today's state-of-the-art technologies, offer significant opportunities to help people be more productive and creative as well as hold major potential to address global challenges.

Alongside these benefits, they also present risks of harm. That is why we are engaging in a number of partnerships across industry, civil society, and academia to share our learnings and learn from others.

An important example of how we can lead through partnership is our co-founding of the Frontier Model Forum alongside Anthropic, Google, and OpenAI. The Frontier Model Forum is an industry non-profit dedicated to the safe and secure development of frontier AI models by sharing information, developing best practices, and advancing research in frontier AI safety.

By leveraging the expertise of the founding members, as well as other organizations committed to developing and deploying frontier models safely, the Frontier Model Forum works toward four priorities:

1. **Advance AI safety research.** We must question, investigate, and collaborate on how to responsibly develop and deploy frontier models to address the challenges of frontier AI. We will collaborate to develop standardized approaches to enable independent evaluations of models' capabilities and safety, where appropriate.
2. **Contribute to the development and refinement of best practices.** We are working with partners to identify best practices for the responsible development and deployment of frontier models and how to improve our practices as we continuously learn.
3. **Share information and seek input.** We work with policymakers, academics, civil society organizations, and the private sector to share knowledge about safety risks and continuously earn trust. We strongly believe that AI will touch virtually every aspect of life, so frontier models will need input from all corners of society to operate responsibly.
4. **Support efforts to develop applications that can help meet society's greatest challenges.** Innovation must play a central role in tackling complex and persistent issues, from human-caused climate change to global health.

We continue to support the work of the Frontier Model Forum as it advances understanding of how to address frontier AI risks in a way that benefits organizations and communities around the world.

Phone/WhatsApp: +86 185 7567 8796
+1 (346) 638-0369

We are also a founding member of the multi-stakeholder organization Partnership on AI (PAI). We consistently contribute to workstreams across its areas of focus, including safety-critical AI; fair, transparent, and accountable AI; AI, labor, and the economy; and AI and media integrity. In June 2023, we joined PAI's Framework for Collective Action on Synthetic Media.[74] This set of practices guides the responsible development, creation, and sharing of media created with generative AI. We also participated in PAI's process to develop Guidance for Safe Foundation Model Deployment.[75] We shared insights from our research on mapping, measuring, and mitigating foundation model risks and benefited from a multi-stakeholder exchange on this topic.

We continually look for ways to engage with stakeholders who represent specific concerns. Since early 2023, we have been actively engaging with news publishers and content creators globally, including in the Americas, Europe, and Australia. We listen to feedback from creators and publishers to learn how creative industries are using generative AI tools and to understand concerns from creative communities. We have engaged in public and private consultations, roundtables, and events. For example, we participated in Creative Commons' community meetings on generative AI and creator empowerment. We also sponsored the Creative Commons Global Summit on AI and the Commons held in Mexico City.[76] This summit brought together a diverse set of artists, civil society leaders, technology companies, and academics to address issues related to AI and creative communities. We participated in the inaugural Centre for News Technology and Innovation roundtable on Defining AI in News. Attendees included news organizations, technology companies, and academics from the United States, the United Kingdom, Brazil, and Nigeria. A report from the event highlights areas of opportunity and further multi-stakeholder collaboration.[77] In Australia, we also participated in government-led roundtables engaging with content and creative industries on addressing the issue of copyright and AI.

We support creators by actively engaging in consultations with sector-specific groups to obtain feedback on our tools and incorporate their feedback into product improvements. For example, news publishers expressed hesitation around their content being used to train generative AI models. However, they did not want any exclusion from training datasets to affect how their content appeared in search results. In response to that feedback, we launched granular controls to allow web publishers to exercise greater control over how content from their websites is accessed and used.[78]

We are committed to responsibly scaling AI to empower every person on the planet to achieve more. Our engagements with the broader community of concerned artists, civil society organizations, and academics reflect our investment in learning as we evolve our approach to responsible AI.

+1 (346) 638-0369 phone/whatsapp: +86 185 7567 8796

Using consensus-based safety frameworks

Technology sector-led initiatives comprise one important force to advance responsible AI. Industry and others stand to significantly benefit from the key role that governments can also play.

From within the U.S. Department of Commerce, the National Institute for Standards and Technology (NIST) built and published a voluntary framework to develop AI applications and mitigate related risks. Extensive consultation with industry, civil society organizations, and academic stakeholders helped NIST refine this AI Risk Management Framework (AI RMF). We contributed to NIST consultations and have applied learnings from NIST's work ourselves, including our application of the NIST RMF in our generative AI requirements.

To implement its tasks in Executive Order (EO) 14110 (on the Safe, Secure, and Trustworthy Development and Use of AI), NIST will consult with stakeholders to develop additional guidance, such as a generative AI-specific version of the AI RMF. Federal agencies and their AI providers can leverage the NIST AI RMF and NIST's additional reference materials to meet obligations required by the implementation of EO 14110. The NIST-led AI Safety Institute Consortium (AISIC), which we have joined, has launched five working groups.[79] These working groups will contribute further guidance, datasets, frameworks, and test environments to advance the field of AI safety.

Governments, industry, and other stakeholders can also partner to develop standards, including in international forums. Within the International Standards Organization (ISO), there are ongoing efforts to develop standards to support AI risk management, including the recent publication of ISO/IEC 42001, AI Management System (AIMS). Companion standards will also define controls and support assessments of their implementation. International standards help bring together global expertise in defining widely applicable practices that can serve as the basis for requirements in an interoperable ecosystem.

We have also partnered with the national security and innovation nonprofit MITRE to incorporate security guidance for generative applications into its ATLAS framework.[80] A recent update of the ATLAS framework includes the vulnerabilities of and adversarial attack tactics targeting generative AI and LLMs so organizations can better protect their applications.[81] The framework also highlights case studies of real-world incidents, including how AI red teams and security professionals mitigated identified issues. Finally, the ATLAS update integrates feedback and best practices from the wider community of government, industry, academia, and security experts. This resource provides an actionable framework so security professionals, AI developers, and AI operators can advance safety in generative applications.

We welcome these and other multi-stakeholder initiatives to advance responsible AI, knowing that these efforts produce results that address a broad range of concerns from a variety of stakeholders.

Phone/WhatsApp: +86 185 7567 8796 +1 (346) 638-0369

Supporting AI research initiatives

In addition to governmental and private sector investment in responsible AI, academic research and development can help realize the potential of this technology. Yet academic institutions do not always have the resources needed to research and train AI models.

The National AI Research Resource (NAIRR) seeks to address this challenge.[82] It intends to provide high-quality data, computational resources, and educational support to make cutting-edge AI research possible for more U.S. academic institutions. We would also welcome and support an extension of the NAIRR to provide access to academic institutions among partners globally. We believe that this comprehensive resource would enable the United States and like-minded nations to continue to lead in AI innovation and risk mitigation.

As currently proposed, a U.S.-focused NAIRR will support a national network of users in training the most resource-intensive models on a combination of supercomputer and commercial cloud-based infrastructure. This centralized resource would enable academics to pursue new lines of research and development without individual institutions needing to heavily invest in computing. Democratizing AI research and development is an essential step toward diversifying the field, leading to a greater breadth in background, viewpoints, and experience necessary to build AI applications that serve society as fully as possible. In short, NAIRR will enable the country to innovate at scale.

In 2023, we announced our support of the NAIRR pilot led by the National Science Foundation (NSF).[83] For this pilot, we committed $20 million worth of Azure compute credits and access to leading-edge models including those available in Azure OpenAI Service.

In the spirit of advancing AI research, we have developed the Accelerating Foundation Models Research (AFMR) program.[84] The AFMR program assembles an interdisciplinary research community to engage with some of the greatest technical and societal challenges of our time. Through the AFMR, we make leading foundation models hosted by Microsoft Azure more accessible to the academic research community. So far, we have extended access to Azure OpenAI Service to 212 AFMR principal investigators from 117 institutions across 17 countries. These projects focus on three goals:

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

In the next call for proposals, we will seek projects in the areas of AI cognition and the economy, AI for creativity, evaluation and measurement, and AI data engagement for natural and life sciences. We also launched an AFMR grant for AI projects advanced by Minority Serving Institutions, focused on Historically Black Colleges and Universities (HBCUs) and Hispanic-Serving Institutions (HSIs), with 10 inaugural grant recipients.[85]

$20 million

worth of Azure compute credits committed to the NAIRR pilot, in addition to access to leading-edge models.

In 2023, we announced the Microsoft Research AI & Society Fellows program to foster research collaboration between Microsoft Research and scholars at the intersection of AI and societal impact.[86] We recognize the value of bridging academic, industry, policy, and regulatory worlds and seek to ignite interdisciplinary collaboration that drives real-world impact. In the fall of 2023, Microsoft Research ran a global call for proposals to seek collaborators for a diverse set of thirteen research challenges. The 24 AI & Society Fellows were announced in early 2024. These fellows will join our researchers for a one-year collaboration with the goal of catalyzing research and contributing publications that advance scholarly discourse and benefit society more broadly.

Phone/WhatsApp: +86 185 7567 8796
+1 (346) 638-0369

Investing in research to advance the state of the art in responsible AI

Microsoft researchers are also advancing the state of the art in generative AI, frequently in partnership with experts outside of the company.

Researchers affiliated with Microsoft Research[87] and Aether[88] published extensive research in 2023 to advance our practices for mapping, measuring, and managing AI risks,[89] some of which we summarize here.

Identifying risks in LLMs and their applications

One of our approaches for identifying risks is the Responsible AI Impact Assessment, which includes envisioning the benefits and harms for stakeholders of an AI application. To address the challenge of identifying potential risks before AI application development or deployment, researchers introduced AHA! (anticipating harms of AI),[90] a human-AI collaboration for systematic impact assessment.

Our researchers also contributed greatly to advancing red teaming knowledge through the production of tools, like AdaTest++,[91] that augment existing red teaming practices. Our researchers uncovered and shared novel privacy and security vulnerabilities, such as privacy-inferencing techniques,[92] or attack vectors when integrating LLMs for AI-assisted coding.[93] These researchers play a key role in shaping the emerging practice of responsible AI-focused red teaming and in producing resources to share this practice more broadly.[94]

Research to advance our practices for measuring risks

After we've identified potential risks, we can measure how often risks occur and how effectively they're mitigated.

For scaling measurement practices, our researchers developed a framework that uses two LLMs.[95] One LLM simulates a user's interaction with a generative application, and one LLM evaluates the application's outputs against a grading scheme developed by experts. Another area explored by our researchers is measurement validity. Thinking beyond measuring model accuracy, researchers are advancing metrics that align more appropriately with user needs—for example, when capturing productivity gains.[96]

Our researchers have also made advancements in the emerging field of synthetic data for training and evaluating generative AI models. These include an English-language dataset for evaluating stereotyping and demeaning harms related to gender and sexuality[97] and a framework for increasing diversity in LLM-generated evaluation data.[98]

Managing AI risks through transparency

Responsible use of AI applications is a shared responsibility between application developers and application users. As application developers, it's important that we design mitigations that enable appropriate use of our generative applications—in other words, to minimize users' risk of overreliance [99] on AI-generated outputs.[100]

Our researchers have developed a number of tools and prototypes to assess AI-generated outputs and improve our products. These include an Excel add-in prototype that helps users assess AI-generated code,[101] a case study of how enterprise end users interact with explanations of AI-generated outputs,[102] and research on when code suggestions are most helpful for programmers.[103]

In setting out a roadmap for transparency in the age of LLMs,[104] our researchers argue that human-centered transparency is key to creating better mitigations and controls for AI applications. Their contributions to further a human-centered approach include research on how users interact with AI transparency features, such as explanations of AI-generated outputs,[105] and how to communicate model uncertainty when interacting with AI-generated code completions.[106]

Here, we've just scratched the surface of the contributions our researchers are making to advance our understanding and practice of responsible AI.

Tuning in to global perspectives

Like many emerging technologies, if not managed deliberately, AI may either widen or narrow social and economic divides between communities at both a local and global scale. Currently, the development of AI applications is primarily influenced by the values of a small subset of the global population located in advanced economies. Meanwhile, the far-reaching impact of AI in developing economies is not well understood.

When AI applications conceived in advanced economies are used in developing ones, there is considerable risk that these applications either will not work or will cause harm. This is particularly the case if their development does not carefully consider the nuanced social, economic, and environmental contexts in which they are deployed. Both real and perceived AI-related harms are primary drivers behind increasing calls for AI regulation. Yet developing countries are often left out of multistakeholder regulatory discussions related to AI, even though they are pursuing AI regulation themselves.

We affirm that to be responsible by design, AI must represent, include, and benefit everyone. We are committed to advancing responsible AI norms globally and adapting to the latest regulations. As our AI footprint continues to grow in developing countries, we must ensure our AI products and governance processes reflect diverse perspectives from underrepresented regions. In 2023, we worked with more than 50 internal and external groups to better understand how AI innovation may impact regulators and individuals in developing countries. Groups included the United Nations Conference on Trade and Development (UNCTAD), the U.S. Agency for International Development (USAID), the U.S. Department of State, and the Microsoft Africa Research Institute. What we learned informed two goals:

1. **Promote globally inclusive policy-making and regulation.** AI regulation is still in its infancy, especially in developing countries. We must recruit and welcome diverse perspectives, such as representatives from the Global South, in the global AI policy-making process.

2. **Develop globally relevant technology.** We must work to ensure the responsible AI by design approach works for all the world's citizens and communities by actively collaborating with stakeholders in developing countries.

As our AI presence expands globally, we will continue to make our AI services, products, and responsible AI program more inclusive and relevant to all. We are pursuing this commitment via several avenues.

- **UNESCO AI Business Council:** Microsoft and Telefonica co-chair the UNESCO AI Business Council. This public-private partnership promotes the implementation of UNESCO's Recommendation on the Ethics of AI, which has been adopted by 193 countries so far. For example, we showcase resources and processes to align with responsible AI standards in webinars and UNESCO events. We expect this effort to bring more companies and countries under a cooperative, globally relevant regulatory framework for responsible AI.

- **Global Perspectives Responsible AI Fellowships:** The Strategic Foresight Hub at Stimson Center and the Office of Responsible AI established a fellowship to investigate the impacts of AI on developing countries.[107] The fellowship convenes experts from Africa, Asia, Latin America, and Eastern Europe working to advance AI responsibly. They represent views across academia, civil society, and private and public sectors, offering insights on the responsible use and development of AI in the Global South.

We recognize that we do not have all the answers to responsible AI. We have prioritized collaboration by partnering with a diverse range of private companies, governmental groups, civil society organizations, regulators, and international bodies. This dynamic mix of perspectives, lived experiences, technical expertise, and concerns pushes us to continue to do better.

50+

groups engaged to better understand how AI innovation may impact regulators and individuals in developing countries.

Phone/WhatsApp: +86 185 7567 8796
+1 (346) 638-0369

Looking ahead

The progress we've shared in this report would not be possible without the passion and commitment of our employees across the company and around the world. Everyone at Microsoft has a role to play in developing AI applications responsibly. Through innovation, collaboration, and a willingness to learn and evolve our approach, we will continue to drive progress on our goals.

We'll continue to invest in four key areas to enable the scaling of responsible AI across the industry:

1. **Innovating new approaches to responsible AI development in our own products.**
2. **Creating tools for our customers to responsibly develop their own AI applications.**
3. **Sharing our learnings and best practices with the responsible AI ecosystem at large.**
4. **Supporting the development of laws, norms, and standards via broad and inclusive multistakeholder processes.**

We continue to develop policies, tools, and solutions to mitigate risks in our own AI products. For example, as we identify new potential risks in generative AI, we improve our existing mitigations or build new mitigations to try and stay ahead of emerging threats. In addition to using innovative tools to protect our users, we also make them available to customers, through tools like Azure AI Content Safety and Azure AI Studio. This approach supports customers' responsible deployment of their own AI applications.

To expand and improve our collective playbook of responsible AI best practices, we'll continue to share our learnings in deploying AI in a safe, secure, and trustworthy manner. We'll provide updates through our own channels like the Microsoft On the Issues blog. We'll also participate in multistakeholder organizations where we can both share our learnings and learn from experts outside the company.

Finally, we recognize that we cannot build AI applications that empower everyone on the planet to achieve more if we do not include everyone's voices in the way these technologies are built and governed. As we move into a new era of global AI governance, we'll do our part to ensure that laws, norms, and standards—and the policies we apply internally—are developed via broad and inclusive multistakeholder processes.

2024

promises to be another pivotal year in responsible AI as we work to harness recent momentum to further accelerate our progress toward safe, secure, and trustworthy AI—in our own products, the work of our customers, and throughout the global technology landscape.

Phone/WhatsApp: +86 185 7567 8796
+1 (340) 938-0569

Sources and resources

[1] Overview of Responsible AI practices for Azure OpenAI models - Azure AI services | Microsoft Learn. https://learn.microsoft.com/en-us/legal/cognitive-services/openai/overview?context=%2Fazure%2Fai-services%2Fopenai%2Fcontext%2Fcontext

[2] Transparency Note for Azure OpenAI Service - Azure AI services | Microsoft Learn. https://learn.microsoft.com/en-us/legal/cognitive-services/openai/transparency-note?context=%2Fazure%2Fcognitive-services%2Fopenai%2Fcontext%2Fcontext&tabs=text

[3] HAX Design Library - Microsoft HAX Toolkit. https://www.microsoft.com/en-us/haxtoolkit/library/?taxonomy_application-type%5B%5D=96

[4] Introduction to red teaming large language models (LLMs) - Azure OpenAI Service | Microsoft Learn. https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/red-teaming

[5] Frontier Model Forum: What is Red Teaming? https://www.frontiermodelforum.org/uploads/2023/10/FMF-AI-Red-Teaming.pdf

[6] System message framework and template recommendations for Large Language Models (LLMs). https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/system-message

[7] Azure AI announces prompt shields for jailbreak and indirect prompt injection attacks (microsoft.com) https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/azure-ai-announces-prompt-shields-for-jailbreak-and-indirect/ba-p/4099140

[8] Introducing AI-assisted safety evaluations in Azure AI Studio (microsoft.com). https://techcommunity.microsoft.com/t5/ai-ai-platform-blog/introducing-ai-assisted-safety-evaluations-in-azure-ai-studio/ba-p/4098595

[9] Introducing risks & safety monitoring feature in Azure Open AI Service - Microsoft Community Hub. https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/introducing-risks-amp-safety-monitoring-feature-in-azure-openai/ba-p/4099218

[10] Microsoft Security Development Lifecycle Threat Modeling https://www.microsoft.com/en-us/securityengineering/sdl/threatmodeling

[11] Microsoft Security Development Lifecycle. https://www.microsoft.com/en-us/securityengineering/sdl/

[12] Cyberattacks against machine learning systems are more common than you think | Microsoft Security Blog. https://www.microsoft.com/en-us/security/blog/2020/10/22/cyberattacks-against-machine-learning-systems-are-more-common-than-you-think/

[13] Satya Nadella email to employees: Embracing our future: Intelligent Cloud and Intelligent Edge - Stories (microsoft.com). https://news.microsoft.com/2018/03/29/satya-nadella-email-to-employees-embracing-our-future-intelligent-cloud-and-intelligent-edge/

[14] Threat Modeling AI/ML Systems and Dependencies - Security documentation | Microsoft Learn. https://learn.microsoft.com/en-us/security/engineering/threat-modeling-aiml

[15] Microsoft Vulnerability Severity Classification for Artificial Intelligence and Machine Learning Systems https://www.microsoft.com/en-us/msrc/aibugbar?rtc=1

[16] Threat Modeling AI/ML Systems and Dependencies - Security documentation | Microsoft Learn. https://learn.microsoft.com/en-us/security/engineering/threat-modeling-aiml

[17] Responsible AI at Microsoft. https://aka.ms/rai

[18] Microsoft Enterprise Cloud Red Teaming. https://download.microsoft.com/download/C/1/9/C1990DBA-502F-4C2A-848D-392B93D9B9C3/Microsoft_Enterprise_Cloud_Red_Teaming.pdf

[19] Microsoft AI Red Team building future of safer AI | Microsoft Security Blog. https://www.microsoft.com/en-us/security/blog/2023/08/07/microsoft-ai-red-team-building-future-of-safer-ai/

[20] Microsoft's AI Red Team Has Already Made the Case for Itself. https://www.wired.com/story/microsoft-ai-red-team/

[21] Voluntary Commitments by Microsoft to Advance Responsible AI Innovation. https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2023/07/Microsoft-Voluntary-Commitments-July-21-2023.pdf

[22] Monitoring evaluation metrics descriptions and use cases (preview) - Azure Machine Learning | Microsoft Learn. https://learn.microsoft.com/en-us/azure/machine-learning/prompt-flow/concept-model-monitoring-generative-ai-evaluation-metrics?view=azureml-api-2#groundedness

[23] Copilot in Bing: Our approach to Responsible AI. http://aka.ms/responsibleAI-copilotinbing

[24] GitHub Copilot – Your AI Pair Programmer - GitHub. https://github.com/features/copilot

[25] Transparency Note for Azure OpenAI Service – Azure AI services | Microsoft Learn. https://learn.microsoft.com/en-us/legal/cognitive-services/openai/transparency-note?context=%2Fazure%2Fai-services%2Fopenai%2Fcontext%2Fcontext&tabs=text

[26] Azure OpenAI Service content filtering - Azure OpenAI | Microsoft Learn. https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/content-filter

[27] Azure AI Content Safety – AI Content Moderation | Microsoft Azure. https://azure.microsoft.com/en-us/products/ai-services/ai-content-safety/

[28] Azure AI announces prompt shields for jailbreak and indirect prompt injection attacks (microsoft.com). https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/azure-ai-announces-prompt-shields-for-jailbreak-and-indirect/ba-p/4099140

[29] Limited access to Azure OpenAI Service - Azure AI services | Microsoft Learn. https://learn.microsoft.com/en-us/legal/cognitive-services/openai/limited-access

[30] Code of conduct for Azure OpenAI Service. https://learn.microsoft.com/en-us/legal/cognitive-services/openai/code-of-conduct

[31] Project Origin - Microsoft Research. https://www.microsoft.com/en-us/research/project/project-origin/

[32] Meeting the Moment: combating AI deepfakes in elections through today's new tech accord. https://blogs.microsoft.com/on-the-issues/2024/02/16/ai-deepfakes-elections-munich-tech-accord/

[33] Microsoft-2024 Elections. https://www.microsoft.com/en-us/concern/2024elections

[34] Microsoft Content Integrity Check tool. https://contentintegrity.microsoft.com/

[35] See Chapter 21 of AI for Good: Applications in Sustainability, Humanitarian Action, and Health. https://www.microsoft.com/en-us/research/group/ai-for-good-research-lab/ai-for-good-book/

[36] Supporting journalists and newsrooms. https://www.microsoft.com/en-us/corporate-responsibility/journalism-hub

Phone/WhatsApp: +86 185 1867 8706
+1 (306) 630-0669

Sources and resources

[37] Microsoft announces new steps to help protect elections - Microsoft On the Issues. https://blogs.microsoft.com/on-the-issues/2023/11/07/microsoft-elections-2024-ai-voting-mtac/

[38] Content Credentials Verify Tool. https://contentcredentials.org/verify

[39] Microsoft Content Integrity Check tool. https://contentintegrity.microsoft.com/

[40] Content Credentials on Azure Open AI service. https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/content-credentials

[41] Microsoft and OpenAI extend partnership - The Official Microsoft Blog. https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/

[42] OpenAI's approach to alignment research (openai.com) https://openai.com/blog/our-approach-to-alignment-research

[43] Preparedness (openai.com). https://openai.com/safety/preparedness

[44] Microsoft Copilot Studio | Extend Copilots or Create Your Own. https://www.microsoft.com/en-us/microsoft-copilot/microsoft-copilot-studio

[45] GitHub Copilot · Your AI pair programmer. https://github.com/features/copilot

[46] GitHub Copilot Enterprise is now generally available - The GitHub Blog. https://github.blog/2024-02-27-github-copilot-enterprise-is-now-generally-available/#accenture-research-shows-the-productivity-impact-of-github-copilot-in-the-e-nterprise

[47] Research: Quantifying GitHub Copilot's impact on code quality - The GitHub Blog. https://github.blog/2023-10-10-research-quantifying-github-copilots-impact-on-code-quality/

[48] GitHub Copilot Trust Center. https://resources.github.com/copilot-trust-center/

[49] About GitHub Copilot Chat - GitHub Enterprise Cloud Docs. https://github.com/features/copilot#faq

[50] About GitHub Copilot Chat - GitHub Enterprise Cloud Docs. https://docs.github.com/en/enterprise-cloud@latest/copilot/github-copilot-chat/about-github-copilot-chat

[51] Microsoft Copilot for Security | Microsoft Security. https://www.microsoft.com/en-us/security/business/ai-machine-learning/microsoft-copilot-security

[52] Secure Future Initiative. https://blogs.microsoft.com/on-the-issues/2023/11/02/secure-future-initiative-sfi-cybersecurity-cyberattacks/

[53] Announcing Microsoft's AI Customer Commitments - The Official Microsoft Blog. https://blogs.microsoft.com/blog/2023/06/08/announcing-microsofts-ai-customer-commitments/

[54] Partner Innovation | Empowering Responsible AI Practices Through Partners (microsoft.com). https://partnerinnovation.microsoft.com/initiatives/empowering-responsible-ai-practices-through-partners/

[55] Microsoft announces new Copilot Copyright Commitment for customers - Microsoft On the Issues. https://blogs.microsoft.com/on-the-issues/2023/09/07/copilot-copyright-commitment-ai-legal-concerns/

[56] Azure OpenAI customers are also required to use guardrails made available by the service and follow responsible development practices. For the full list of required mitigations for Azure OpenAI customers, see: Customer Copyright Commitment Required Mitigations | Microsoft Learn. https://learn.microsoft.com/en-us/legal/cognitive-services/openai/customer-copyright-commitment

[57] Announcing Microsoft's open automation framework to red team generative AI Systems | Microsoft Security Blog. https://www.microsoft.com/en-us/security/blog/2024/02/22/announcing-microsofts-open-automation-framework-to-red-team-generative-ai-systems/

[58] Announcing Azure AI Studio preview (microsoft.com). https://techcommunity.microsoft.com/t5/ai-ai-platform-blog/unleashing-the-power-of-generative-ai-azure-ai-studio-leads-the/ba-p/3977692

[59] Introducing AI-assisted safety evaluations in Azure AI Studio (microsoft.com). https://techcommunity.microsoft.com/t5/ai-ai-platform-blog/introducing-ai-assisted-safety-evaluations-in-azure-ai-studio/ba-p/4098595

[60] Responsible AI Toolbox. https://github.com/microsoft/responsible-ai-toolbox

[61] System message framework and template recommendations for Large Language Models(LLMs) - Azure OpenAI Service | Microsoft Learn. https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/system-message

[62] Azure AI announces prompt shields for jailbreak and indirect prompt injection attacks (microsoft.com) https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/azure-ai-announces-prompt-shields-for-jailbreak-and-indirect/ba-p/4099140

[63] Detect and Mitigate Ungrounded Model Outputs - Microsoft Community Hub. https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/detect-and-mitigate-ungrounded-model-outputs/ba-p/4099261

[64] Introducing risks & safety monitoring feature in Azure Open AI Service - Microsoft Community Hub. https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/introducing-risks-amp-safety-monitoring-feature-in-azure-openai/ba-p/4099218

[65] Transparency Note for Azure AI Face service | Microsoft Learn. https://learn.microsoft.com/en-us/legal/cognitive-services/face/transparency-note?context=/azure/ai-services/computer-vision/context/context

[66] Transparency Note for Azure OpenAI - Azure AI services | Microsoft Learn. https://learn.microsoft.com/en-us/legal/cognitive-services/openai/transparency-note?context=%2Fazure%2Fai-services%2Fopenai%2Fcontext%2Fcontext&tabs=text

[67] Transparency note for Document Intelligence - Azure AI services | Microsoft Learn. https://learn.microsoft.com/en-us/legal/cognitive-services/document-intelligence/transparency-note

[68] Copilot in Bing: Our approach to Responsible AI. http://aka.ms/responsibleAI-copilotinbing

[69] Your Everyday AI Companion | Microsoft Bing. https://www.microsoft.com/en-us/bing?form=MG0AUO&OCID=MG0AUO#faq

[70] AI Ethics and Effects in Engineering and Research committee and working groups leverage top scientific and engineering talent to provide subject-matter expertise on the state-of-the-art and emerging trends regarding the enactment of Microsoft's responsible AI principles.

[71] Microsoft Research – Emerging Technology, Computer, and Software Research. https://www.microsoft.com/en-us/research/

[72] Isabel Gottleib, "AI's 'Unicorn Hunt' Gives Privacy Officers a Key Governance Role," Bloomberg Law, October 10, 2023. https://news.bloomberglaw.com/artificial-intelligence/ais-unicorn-hunt-gives-privacy-officers-a-key-governance-role

[73] Microsoft Global Diversity & Inclusion Report 2023. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW1e53b

Sources and resources

[74] PAI's Responsible Practices for Synthetic Media. https://syntheticmedia.partnershiponai.org/

[75] 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. https://partnershiponai.org/pai-model-deployment-guidance-press-release/

[76] Creative Commons Global Summit on AI & the commons. https://summit.creativecommons.org/

[77] Defining AI in News - Center for News, Technology & Innovation (innovating.news). https://innovating.news/article/defining-ai-in-news/

[78] Announcing new options for webmasters to control usage of their content in Bing Chat | Bing Webmaster Blog. https://blogs.bing.com/webmaster/september-2023/Announcing-new-options-for-webmasters-to-control-usage-of-their-content-in-Bing-Chat

[79] NIST AISIC Working Groups. https://www.nist.gov/artificial-intelligence/artificial-intelligence-safety-institute/aisic-working-groups

[80] MITRE | ATLAS™. https://atlas.mitre.org/

[81] MITRE and Microsoft Collaborate to Address Generative AI Security Risks | MITRE. https://www.mitre.org/news-insights/news-release/mitre-and-microsoft-collaborate-address-generative-ai-security-risks

[82] National Artificial Intelligence Research Resource Pilot | NSF - National Science Foundation. https://new.nsf.gov/focus-areas/artificial-intelligence/nairr

[83] Broadening AI innovation: Microsoft's pledge to the National AI Research Resource pilot - Microsoft On the Issues. https://blogs.microsoft.com/on-the-issues/2024/01/24/national-ai-research-resource-nairr-artificial-intelligence/

[84] Accelerating Foundation Models Research program. https://www.microsoft.com/en-us/research/collaboration/accelerating-foundation-models-research/

[85] Announcing recipients of the AFMR Minority Serving Institutions grant - Microsoft Research. https://www.microsoft.com/en-us/research/blog/announcing-recipients-of-the-afmr-minority-serving-institutions-grant/

[86] Microsoft Research AI & Society Fellows. https://www.microsoft.com/en-us/research/academic-program/ai-society-fellows/overview/

[87] Microsoft Research – Emerging Technology, Computer, and Software Research. https://www.microsoft.com/en-us/research/

[88] AI Ethics and Effects in Engineering and Research committee and working groups leverage top scientific and engineering talent to provide subject-matter expertise on the state-of-the-art and emerging trends regarding the enactment of Microsoft's responsible AI principles.

[89] Advancing transparency: Updates on responsible AI research - Microsoft Research. https://www.microsoft.com/en-us/research/blog/advancing-transparency-updates-on-responsible-ai-research/

[90] AHA! Facilitating AI Impact Assessment by Generating Examples of Harms - Microsoft Research. https://www.microsoft.com/en-us/research/publication/aha-facilitating-ai-impact-assessment-by-generating-examples-of-harms/

[91] Supporting Human-AI Collaboration in Auditing LLMs with LLMs - Microsoft Research. https://www.microsoft.com/en-us/research/publication/supporting-human-ai-collaboration-in-auditing-llms-with-llms/

[92] Does Prompt-Tuning Language Model Ensure Privacy? - Microsoft Research. https://www.microsoft.com/en-us/research/publication/does-prompt-tuning-language-model-ensure-privacy/

[93] TROJANPUZZLE: Covertly Poisoning Code-Suggestion Models - Microsoft Research. https://www.microsoft.com/en-us/research/publication/trojanpuzzle-covertly-poisoning-code-suggestion-models/

[94] Planning red teaming for large language models (LLMs) and their applications - Azure OpenAI Service | Microsoft Learn. https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/red-teaming

[95] A Framework for Automated Measurement of Responsible AI Harms in Generative AI Applications - Microsoft Research. https://www.microsoft.com/en-us/research/publication/a-framework-for-automated-measurement-of-responsible-ai-harms-in-generative-ai-applications/

[96] Aligning Offline Metrics and Human Judgments of Value for Code Generation Models - Microsoft Research. https://www.microsoft.com/en-us/research/publication/aligning-offline-metrics-and-human-judgments-of-value-for-code-generation-models/

[97] FairPrism: Evaluating Fairness-Related Harms in Text Generation - Microsoft Research. https://www.microsoft.com/en-us/research/publication/fairprism-evaluating-fairness-related-harms-in-text-generation/

[98] Increasing Diversity While Maintaining Accuracy: Text Data Generation with Large Language Models and Human Interventions - Microsoft Research. https://www.microsoft.com/en-us/research/publication/increasing-diversity-while-maintaining-accuracy-text-data-generation-with-large-language-models-and-human-interventions/

[99] Overreliance on AI: Literature Review - Microsoft Research. https://www.microsoft.com/en-us/research/publication/overreliance-on-ai-literature-review/

[100] Appropriate reliance on Generative AI: Research Synthesis. https://www.microsoft.com/en-us/research/publication/appropriate-reliance-on-generative-ai-research-synthesis/

[101] CoIDeco: An End User Spreadsheet Inspection Tool for AI-Generated Code - Microsoft Research. https://www.microsoft.com/en-us/research/publication/coldeco-an-end-user-spreadsheet-inspection-tool-for-ai-generated-code/

[102] Surfacing AI Explainability in Enterprise Product Visual Design to Address User Tech Proficiency Differences - Microsoft Research. https://www.microsoft.com/en-us/research/publication/surfacing-ai-explainability-in-enterprise-product-visual-design-to-address-user-tech-proficiency-differences/

[103] When to Show a Suggestion? Integrating Human Feedback in AI-Assisted Programming - Microsoft Research. https://www.microsoft.com/en-us/research/publication/when-to-show-a-suggestion-integrating-human-feedback-in-ai-assisted-programming/

[104] AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. https://www.microsoft.com/en-us/research/publication/ai-transparency-in-the-age-of-llms-a-human-centered-research-roadmap/

[105] Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations - Microsoft Research. https://www.microsoft.com/en-us/research/publication/understanding-the-role-of-human-intuition-on-reliance-in-human-ai-decision-making-with-explanations/

[106] Generation Probabilities Are Not Enough: Exploring the Effectiveness of Uncertainty Highlighting in AI-Powered Code Completions - Microsoft Research. https://www.microsoft.com/en-us/research/publication/generation-probabilities-are-not-enough-exploring-the-effectiveness-of-uncertainty-highlighting-in-ai-powered-code-completions/

[107] Advancing AI responsibly – Microsoft Unlocked. https://unlocked.microsoft.com/responsible-ai/

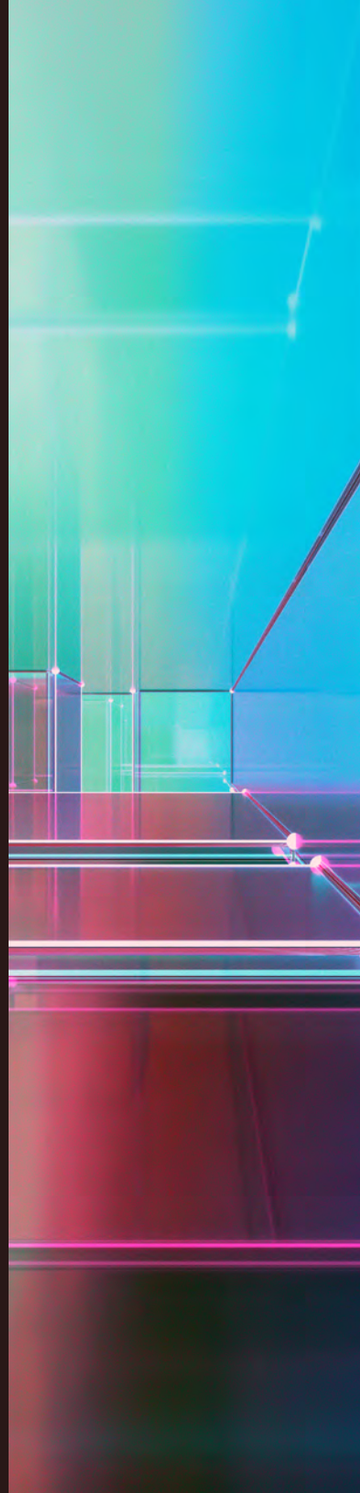
Phone/WhatsApp: +86 185 1667 0196
+1 (346) 223-0000

Microsoft

Responsible AI Transparency Report

Phone/WhatsApp: +86 185 7567 8796

+1 (346) 638-0369



©2024 Microsoft Corporation. All rights reserved.