



Governing AI: A Blueprint for the Future

May 25, 2023

Table of contents

.....
By Microsoft Vice Chair and President Brad Smith

.....
.....

.....
.....

.....
.....

.....
.....

.....

.....

.....

.....

.....

Foreword: How Do We Best Govern AI?



Brad Smith, Vice Chair
and President, Microsoft

“Don’t ask what computers can do, ask what they should do.”

That is the title of the chapter on AI and ethics in a book I coauthored in 2019. At the time, we wrote that “this may be one of the defining questions of our generation.” Four years later, the question has seized center stage not just in the world’s capitals, but around many dinner tables.

As people have used or heard about the power of OpenAI’s GPT-4 foundation model, they have often been surprised or even astounded. Many have been enthused or even excited. Some have been concerned or even frightened. What has become clear to almost everyone is something we noted four years ago—we are the first generation in the history of humanity to create machines that can make decisions that previously could only be made by people.

Countries around the world are asking common questions. How can we use this new technology to solve our problems? How do we avoid or manage new problems it might create? How do we control technology that is so powerful?

These questions call not only for broad and thoughtful conversation, but decisive and effective action. This paper offers some of our ideas and suggestions as a company.

These suggestions build on the lessons we’ve been learning based on the work we’ve been doing for several years. Microsoft CEO Satya Nadella set us on a clear course when he [wrote in 2016](#) that “perhaps the most productive debate we can have isn’t one of good versus evil: The debate should be about the values instilled in the people and institutions creating this technology.”

Since that time, we’ve defined, published, and implemented ethical principles to guide our work. And we’ve built out constantly improving engineering and governance systems

to put these principles into practice. Today we have nearly 350 people working on responsible AI at Microsoft, helping us implement best practices for building safe, secure, and transparent AI systems designed to benefit society.

New opportunities to improve the human condition

The resulting advances in our approach have given us the capability and confidence to see ever-expanding ways for AI to improve people’s lives. We’ve seen AI help save individuals’ eyesight, make progress on new cures for cancer, generate new insights about proteins, and provide predictions to protect people from hazardous weather. Other innovations are fending off cyberattacks and helping to protect fundamental human rights, even in nations afflicted by foreign invasion or civil war.

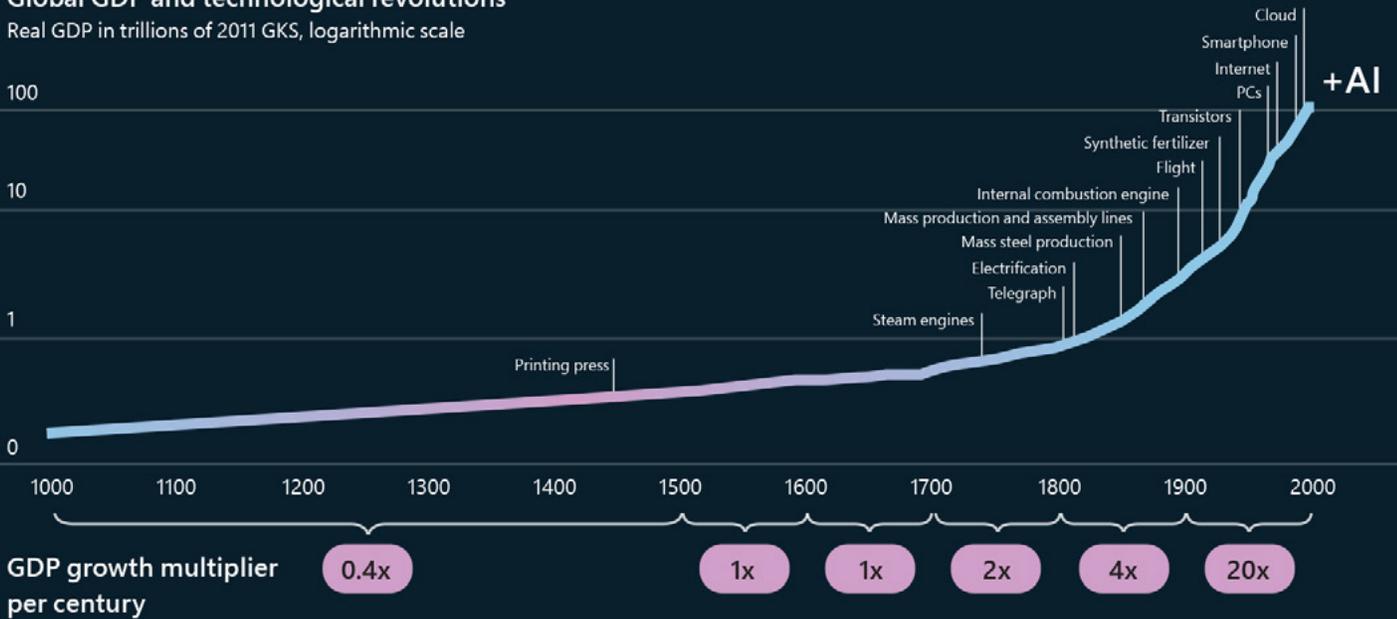
Everyday activities will benefit as well. By acting as a copilot in people’s lives, the power of foundation models like GPT-4 is turning search into a more powerful tool for research and improving productivity for people at work. And for any parent who has struggled to remember how to help their 13-year-old child through an algebra homework assignment, AI-based assistance is a helpful tutor.

In so many ways, AI offers perhaps even more potential for the good of humanity than any invention that has preceded it. Since the invention of the printing press with movable type in the 1400s, human prosperity has been growing at an accelerating rate. Inventions like the steam engine, electricity, the automobile, the airplane, computing, and the internet have provided many of the building blocks for modern civilization. And like the printing press itself, AI offers a new tool to genuinely help advance human learning and thought.

Technology drives GDP growth, and the pace of change is accelerating

Global GDP and technological revolutions

Real GDP in trillions of 2011 GKS, logarithmic scale



Source: Maddison Project Our World in Data¹

Guardrails for the future

Another conclusion is equally important: it's not enough to focus only on the many opportunities to use AI to improve people's lives. This is perhaps one of the most important lessons from the role of social media. Little more than a decade ago, technologists and political commentators alike gushed about the role of social media in spreading democracy during the Arab Spring. Yet five years after that, we learned that social media, like so many other technologies before it, would become both a weapon and a tool—in this case aimed at democracy itself.

Today, we are 10 years older and wiser, and we need to put that wisdom to work. We need to think early on and in a clear-eyed way about the problems that could lie ahead. As technology moves forward, it's just as important to ensure proper control over AI as it is to pursue its benefits. We are committed and determined as a company to develop and deploy AI in a safe and responsible way. We also recognize,

however, that the guardrails needed for AI require a broadly shared sense of responsibility and should not be left to technology companies alone.

When we at Microsoft adopted our six ethical principles for AI in 2018, we noted that one principle was the bedrock for everything else—accountability. This is the fundamental need: to ensure that machines remain subject to effective oversight by people and the people who design and operate machines remain accountable to everyone else. *In short, we must always ensure that AI remains under human control.* This must be a first-order priority for technology companies and governments alike.

This connects directly with another essential concept. In a democratic society, one of our foundational principles is that no person is above the law. No government is above the law. No company is above the law, and no product or technology should be above the law. This leads to a critical conclusion: people who design and operate AI systems

¹ Visit the [Our World in Data](#) site to download a CSV file of the full dataset used in the chart.

cannot be accountable unless their decisions and actions are subject to the rule of law.

In many ways, this is at the heart of the unfolding AI policy and regulatory debate. How do governments best ensure that AI is subject to the rule of law? In short, what form should new law, regulation, and policy take?

A five-point blueprint for the public governance of AI

Part 1 of this paper offers a five-point blueprint to address several current and emerging AI issues through public policy, law, and regulation. We offer this recognizing that every part of this blueprint will benefit from broader discussion and require deeper development. But we hope this can contribute constructively to the work ahead.

First, implement and build upon new government-led AI safety frameworks. The best way to succeed is often to build on the successes and good ideas of others. Especially when one wants to move quickly. In this instance, there is an important opportunity to build on work completed

just four months ago by the U.S. National Institute of Standards and Technology, or NIST. Part of the Department of Commerce, NIST has completed and launched a new AI Risk Management Framework.

We offer four concrete suggestions to implement and build upon this framework, including commitments Microsoft is making in response to a recent White House meeting with leading AI companies. We also believe the Administration and other governments can accelerate momentum through procurement rules based on this framework.

Second, require effective safety brakes for AI systems that control critical infrastructure. In some quarters, thoughtful individuals increasingly are asking whether we can satisfactorily control AI as it becomes more powerful. Concerns are sometimes posed regarding AI control of critical infrastructure like the electrical grid, water system, and city traffic flows.

This is the right time to discuss this question. This blueprint proposes new safety requirements that in effect would create safety brakes for AI systems that control the operation of designated critical

A five-point blueprint for governing AI

- 1 Implement and build upon new government-led AI safety frameworks
- 2 Require effective safety brakes for AI systems that control critical infrastructure
- 3 Develop a broader legal and regulatory framework based on the technology architecture for AI
- 4 Promote transparency and ensure academic and public access to AI
- 5 Pursue new public-private partnerships to use AI as an effective tool to address the inevitable societal challenges that come with new technology

infrastructure. These fail-safe systems would be part of a comprehensive approach to system safety that would keep effective human oversight, resilience, and robustness top of mind. In spirit, they would be similar to the braking systems engineers have long built into other technologies such as elevators, school buses, and high-speed trains, to safely manage not just everyday scenarios, but emergencies as well.

In this approach, the government would define the class of high-risk AI systems that control critical infrastructure and warrant such safety measures as part of a comprehensive approach to system management. New laws would require operators of these systems to build safety brakes into high-risk AI systems by design. The government would then ensure that operators test high-risk systems regularly to make certain that the system safety measures are effective. And AI systems that control the operation of designated critical infrastructure would be deployed only in licensed AI datacenters that would ensure a second layer of protection through the ability to apply these safety brakes, thereby ensuring effective human control.

Third, develop a broad legal and regulatory framework based on the technology architecture for AI.

We believe there will need to be a legal and regulatory architecture for AI that reflects the technology architecture for AI itself. In short, the law will need to place various regulatory responsibilities upon different actors based upon their role in managing different aspects of AI technology.

For this reason, this blueprint includes information about some of the critical pieces that go into building and using new generative AI models. Using this as context, it proposes that different laws place specific regulatory responsibilities on the organizations exercising certain responsibilities at three layers of the technology stack: the applications layer, the model layer, and the infrastructure layer.

This should first apply existing legal protections at the applications layer to the use of AI. This is the layer where the safety and rights of people will most be impacted, especially because the impact of AI can vary markedly in different technology scenarios. In many areas, we don't need new laws and regulations. We instead need to apply and enforce existing laws and regulations, helping agencies and courts develop the expertise needed to adapt to new AI scenarios.

KY3C:

Applying to AI services the "Know Your Customer" concept developed for financial services

Know your Cloud

Know your Customer

Know your Content

There will then be a need to develop new law and regulations for highly capable AI foundation models, best implemented by a new government agency. This will impact two layers of the technology stack. The first will require new regulations and licensing for these models themselves. And the second will involve obligations for the AI infrastructure operators on which these models are developed and deployed. The blueprint that follows offers suggested goals and approaches for each of these layers.

In doing so, this blueprint builds in part on a principle developed in recent decades in banking to protect against money laundering and criminal or terrorist use of financial services. The “Know Your Customer”—or KYC—principle requires that financial institutions verify customer identities, establish risk profiles, and monitor transactions to help detect suspicious activity. It would make sense to take this principle and apply a KY3C approach that creates in the AI context certain obligations to know one’s *cloud*, one’s *customers*, and one’s *content*.

In the first instance, the developers of designated, powerful AI models first “know the cloud” on which their models are developed and deployed. In addition, such as for scenarios that involve sensitive uses, the company that has a direct relationship with a customer—whether it be the model developer, application provider, or cloud operator on which the model is operating—should “know the customers” that are accessing it.

Also, the public should be empowered to “know the content” that AI is creating through the use of a label or other mark informing people when something like a video or audio file has been produced by an AI model rather than a human being. This labeling obligation should also protect the public from the alteration of original content and the creation of “deep fakes.” This will require the development of new laws, and there will be many important questions and details to address. But the health of democracy and future of civic discourse will benefit from thoughtful measures to deter the use of new technology to deceive or defraud the public.

Fourth, promote transparency and ensure academic and nonprofit access to AI. We believe a critical public goal is to advance transparency and broaden access to

AI resources. While there are some important tensions between transparency and the need for security, there exist many opportunities to make AI systems more transparent in a responsible way. That’s why Microsoft is committing to an annual AI transparency report and other steps to expand transparency for our AI services.

We also believe it is critical to expand access to AI resources for academic research and the nonprofit community. Basic research, especially at universities, has been of fundamental importance to the economic and strategic success of the United States since the 1940s. But unless academic researchers can obtain access to substantially more computing resources, there is a real risk that scientific and technological inquiry will suffer, including relating to AI itself. Our blueprint calls for new steps, including steps we will take across Microsoft, to address these priorities.

Fifth, pursue new public-private partnerships to use AI as an effective tool to address the inevitable societal challenges that come with new technology. One

lesson from recent years is what democratic societies can accomplish when they harness the power of technology and bring the public and private sectors together. It’s a lesson we need to build upon to address the impact of AI on society.

We will all benefit from a strong dose of clear-eyed optimism. AI is an extraordinary tool. But like other technologies, it too can become a powerful weapon, and there will be some around the world who will seek to use it that way. But we should take some heart from the cyber front and the last year and a half in the war in Ukraine. What we found is that when the public and private sectors work together, when like-minded allies come together, and when we develop technology and use it as a shield, it’s more powerful than any sword on the planet.

Important work is needed now to use AI to protect democracy and fundamental rights, provide broad access to the AI skills that will promote inclusive growth, and use the power of AI to advance the planet’s sustainability needs. Perhaps more than anything, a wave of new AI technology provides an occasion for thinking big and acting boldly. In each area, the key to success will be to develop concrete initiatives and bring governments, respected companies,

and energetic NGOs together to advance them. We offer some initial ideas in this report, and we look forward to doing much more in the months and years ahead.

Governing AI within Microsoft

Ultimately, every organization that creates or uses advanced AI systems will need to develop and implement its own governance systems. Part 2 of this paper describes the AI governance system within Microsoft—where we began, where we are today, and how we are moving into the future.

As this section recognizes, the development of a new governance system for new technology is a journey in and of itself. A decade ago, this field barely existed. Today Microsoft has almost 350 employees specializing in it, and we are investing in our next fiscal year to grow this further.

As described in this section, over the past six years we have built out a more comprehensive AI governance structure and system across Microsoft. We didn't start from scratch, borrowing instead from best practices for the protection of cybersecurity, privacy, and digital safety. This is all part of the company's comprehensive Enterprise Risk Management (ERM) system, which has become a critical part of the management of corporations and many other organizations in the world today.

When it comes to AI, we first developed ethical principles and then had to translate these into more specific corporate policies. We're now on version 2 of the corporate standard that embodies these principles and defines more precise practices for our engineering teams to follow. We've implemented the standard through training, tooling,

and testing systems that continue to mature rapidly. This is supported by additional governance processes that include monitoring, auditing, and compliance measures.

As with everything in life, one learns from experience. When it comes to AI governance, some of our most important learning has come from the detailed work required to review specific sensitive AI use cases. In 2019, we founded a sensitive use review program to subject our most sensitive and novel AI use cases to rigorous, specialized review that results in tailored guidance. Since that time, we have completed roughly 600 sensitive use case reviews. The pace of this activity has quickened to match the pace of AI advances, with almost 150 such reviews taking place in the last 11 months.

All of this builds on the work we have done and will continue to do to advance responsible AI through company culture. That means hiring new and diverse talent to grow our responsible AI ecosystem and investing in the talent we already have at Microsoft to develop skills and empower them to think broadly about the potential impact of AI systems on individuals and society. It also means that much more than in the past, the frontier of technology requires a multidisciplinary approach that combines great engineers with talented professionals from across the liberal arts.

All this is offered in this paper in the spirit that we're on a collective journey to forge a responsible future for artificial intelligence. We can all learn from each other. And no matter how good we may think something is today, we will all need to keep getting better.

As technology change accelerates, the work to govern AI responsibly must keep pace with it. With the right commitments and investments, we believe it can.



Brad Smith
Vice Chair and President, Microsoft

01

Governing AI: A Legal and Regulatory Blueprint for the Future

Governing AI: A Legal and Regulatory Blueprint for the Future

Around the world, governments are looking for or developing what in effect are new blueprints to govern artificial intelligence. There, of course, is no single or right approach. We offer here a five-point approach to help governance advance more quickly, based on the questions and issues that are pressing to many. Every part of this blueprint will benefit from broader discussion and require deeper development. But we hope this can contribute constructively to the work ahead.

This blueprint recognizes the many opportunities to use AI to improve people's lives while also quickly developing new controls, based on both governmental and private initiative, including broader international collaboration. It offers specific steps to:

- **Implement and build upon new government-led AI safety frameworks.**
- **Require effective safety brakes for AI systems that control critical infrastructure.**
- **Develop a broader legal and regulatory framework based on the technology architecture for AI.**
- **Promote transparency and ensure academic and public access to AI.**
- **Pursue new public-private partnerships to use AI as an effective tool to address the inevitable societal challenges that come with new technology.**

This plan responds in part to the White House's recent call for commitments from AI companies to ensure AI safety and security, and it includes several specific commitments that Microsoft is offering in response.

1. Implement and build upon new government-led AI safety frameworks.

One of the most effective ways to move quickly is to build on recent advances in governmental work that advance AI safety. This makes far more sense than starting from scratch, especially when there is a recent and strong footing on which to start.

As events have it, just four months ago, the National Institute of Standards and Technology in the United States, or NIST, completed a year and a half of intensive work and launched an important new AI safety initiative. This new AI Risk Management Framework builds on NIST's years of experience in the cybersecurity domain, where similar frameworks and standards have played a critical role.

We believe the new AI Risk Management Framework provides a strong foundation that companies and governments alike can immediately put into action to ensure the safer use of artificial intelligence. While no single such effort can answer every question, the immediate adoption of this framework will accelerate AI safety momentum around the world. And we can all build upon it in the months ahead.

Part of the U.S. Department of Commerce, NIST developed its new framework based on direction by Congress in the National Artificial Intelligence Initiative Act of 2020. The framework is designed to enable organizations to help manage AI risks and promote the trustworthy and responsible development and use of AI systems. It was developed through a consensus-driven and transparent process involving work by government agencies, civil society organizations, and several technology leaders, including Microsoft.

NIST brings years of experience to the AI risk management space from its years of work developing critical tools to address cybersecurity risks. Microsoft has long experience working with NIST on the cybersecurity front, and it's encouraging to see NIST apply this expertise to help organizations govern, map, measure, and manage the risks associated with AI. We're not alone in our high regard for NIST's approach, as numerous governments, international organizations, and leading businesses have already validated the value of the new AI Risk Management Framework.

Now the question is how to build upon this recent progress so we can all move faster to address AI risks. We believe there are at least four immediate opportunities:

First, Microsoft is committing to the White House, in response to its recent meeting, that we will implement NIST's AI Risk Management Framework. Microsoft's internal [Responsible AI Standard](#) is closely aligned with the framework already, and we will now work over the summer to implement it so that all our AI services benefit from it.

Second, we are similarly committing that we will augment Microsoft's existing AI testing work with new steps to further strengthen our engineering practices relating to high-risk AI systems.

Under Microsoft's Responsible AI Standard, our AI engineering teams already work to identify potential harms, measure their propensity to occur, and build mitigations to address them. We have further developed red teaming techniques using multidisciplinary teams, which were originally developed to identify cybersecurity vulnerabilities, to stress test AI systems with a wide range of expertise, including privacy, security, and fairness.

For high-risk systems, Microsoft is committing that red teaming is conducted before deployment by qualified experts who are independent of the product teams building those systems, adopting a best practice from the financial services industry. We will rely upon these red teams, together with our product teams who are responsible for systematic evaluations of the products that they build, to help us identify, measure, and mitigate potential harms.

In addition to continually monitoring, tracking, and evaluating our AI systems, we will use metrics to measure and understand systemic issues specific to generative AI experiences, such as the extent to which a model's output is supported by information contained in input sources. (We are releasing the first of these metrics this week as part of our Azure OpenAI Service at Build, our annual developer conference.)

Third, we believe the Administration can accelerate momentum through an Executive Order that requires vendors of critical AI systems to the U.S. Government to self-attest that they are implementing NIST's AI Risk Management Framework.

It's important for governments to move faster, using both carrots and sticks. In the United States, federal procurement mechanisms have repeatedly demonstrated their value in improving the quality of products and advancing industry practice more generally. Building on similar approaches used for key technology priorities like cybersecurity, the U.S. Government could insert requirements related to the AI Risk Management Framework into the federal procurement process for AI systems.

As a starting point, we believe it makes sense to scope such procurement requirements to focus on critical decision systems, meaning AI systems that have the potential to meaningfully impact the public's rights, opportunities, or access to critical resources or services. This would align with the approach set out in the [Blueprint for an AI Bill of Rights](#), released last year by the White House's Office of Science and Technology Policy.

Finally, we are committed to working with other industry leaders and those in government to develop new and additional standards relating to highly capable foundation models. We recognize that the pace of AI advances raises new questions and issues related to safety and security, and we are committed to working with others to develop actionable standards to help evaluate and address them. Already, leaders at OpenAI, Google, Anthropic, and other AI companies have advanced important ideas that will help provide a foundation for future progress. We look forward to working with them and many others as these types of efforts move forward.

Using an Executive Order to implement the NIST AI Risk Management Framework

The following steps could be considered as part of a comprehensive approach to implementing the NIST AI Risk Management Framework using an Executive Order.

- **Require self-attestation by vendors of NIST AI RMF alignment.** Self-attestation is used by the government to advance cybersecurity standards amongst federal suppliers. A similar mechanism can be applied to the NIST AI RMF. The Office of Management and Budget (OMB) could issue guidance requiring federal agencies procuring AI services for use in critical decision systems to only do so from suppliers that have self-attested that they meet a minimum bar for implementation for the NIST AI RMF. The minimum bar could be set by the NIST AI RMF Program Office mentioned below.
- **Establish a NIST AI RMF Program Office to advance coordination and enablement.** We suggest the creation of a NIST AI RMF Program Office to provide ongoing guidance for the framework and promote adoption of it across agencies. This Program Office could also work with the new “Agency Equity Teams,” required by EO 14091 on Advancing Racial Equity, to include guidance that helps small- and medium-sized organizations.
- **Develop responsible procurement resources.** The General Services Administration (GSA) and OMB could

be directed to develop voluntary, standard contract language for agencies that are procuring critical decision systems, obligating a baseline set of actions in line with the framework’s recommendations.

Additionally, NIST’s important work to build out AI RMF “Profiles” (guides on how the NIST AI RMF applies to specific sectors and/or systems) could include the development of specific profiles for public sector uses of critical decision systems.

- **Advance training and education.** The NIST AI RMF Program Office, coupled with GSA and Agency Equity Teams, could deliver training on AI trustworthiness for individuals responsible for acquiring or procuring critical decision systems. This would support acquisition professionals in important roles that define the scope of contract solicitations, set contract requirements, or make vendor determinations. Training would cover the technology’s risks and benefits in order to help acquisition professionals determine whether the software under consideration meets standards for performance and does not unlawfully discriminate.
- **Augment baseline AI governance requirements for agencies.** Federal agencies could be required to implement the NIST AI RMF in their own AI development. In time, this could be supplemented with mandatory responsible AI controls for government systems.

2. Require effective safety brakes for AI systems that control critical infrastructure.

History offers an important and repeated lesson about the promise and peril of new technology. Since the advent of the printing press, governments have confronted the need to decide whether to accept or reject new inventions. Beginning in the latter half of the 1400s, Europe embraced the printing press, while the Ottoman Empire mostly

banned it. By 1500, citizens in the Netherlands were reading more books per capita than anyone else. It’s not a coincidence that the small nation soon found itself at the forefront of economic innovation.

Ever since, inventors and governments have typically concluded that the best path forward is to harness the power of new technology in part by taming it. The history of technology is replete with examples.

Modern cities would not be possible without tall buildings,

but tall buildings would not be possible without elevators. And in the 1800s, most people understandably were uncomfortable getting into what all of us today do without even thinking about—entering a metal box and being hoisted several stories into the sky by a cable. Elisha Otis, the American inventor of the elevator, found in the 1850s that the public was slow to accept his machines, deeming them too dangerous.

This changed in 1854 at the World’s Fair in New York, when Otis demonstrated a new safety brake for his elevator. He severed the cable holding his machine above the watching crowd, and the brake immediately caught the car, halting its fall. People were reassured, and in an important respect, the modern city was born.

This pattern has repeated itself for everything from electricity to railroads to school buses. Today houses and buildings have circuit breakers to protect against a surge in the electrical current. City codes require them. Similarly, hundreds of millions of people put what they hold most precious in the world—their children—on morning school buses, based in part on regulations that require buses to

have emergency brakes with bus drivers trained to use them. Planes today have ground proximity detectors and airborne collision avoidance systems that have helped to make commercial air travel incredibly safe, while empowering pilots—not machines—to make decisions in safety-critical scenarios.

As we look to a future with artificial intelligence, it’s worth remembering that the same fundamental approach has worked repeatedly in managing the potential dangers associated with new technology. Namely, identify when a new product could become the equivalent of a runaway train, and as for the locomotive itself, install an effective safety system that can act as a brake and ensure that the right people will use it quickly if it’s ever needed—whether to slow something down or even bring it to a halt.

Not every potential AI scenario poses significant risks, and in fact, most do not. But this becomes more relevant when one contemplates AI systems that manage or control infrastructure systems for electricity grids, the water system, emergency responses, and traffic flows in our cities. We need “safety brakes” to ensure these systems remain under human control.

Four steps governments can take to secure effective safety brakes for AI systems controlling critical infrastructure

- 1 Define the class of high-risk AI systems being deployed
- 2 Require system developers to ensure that safety brakes are built by design into the use of AI systems for the control of infrastructure
- 3 Ensure operators test and monitor high-risk systems to ensure AI systems that power critical infrastructure remain within human control
- 4 Require AI systems that control operation of designated critical infrastructure to be deployed only in licensed AI infrastructure

We believe that the following steps would help address these issues:

First, the government should define the class of high-risk AI systems that are being deployed to control critical infrastructure and warrant safety brakes as part of a comprehensive approach to system safety.

In the United States, the Secretary of Homeland Security is responsible for identifying and prioritizing critical infrastructure in coordination with other government agencies. Most notably, this includes the Cybersecurity and Infrastructure Security Agency, or [CISA](#), which has identified 16 [critical infrastructure sectors](#), including the communications sector, the emergency services sector, and the energy sector, to name a few.

For the purposes of applying the safety brake concept to AI systems, we need to focus on the AI systems that are used to control the operation of critical infrastructure. There will be many AI systems used within critical infrastructure sectors that are low risk and that do not require the same depth of safety measures—employee productivity tools and customer service agents are two such examples.

Instead, one should focus on highly capable systems, increasingly autonomous systems, and systems that cross the digital-physical divide. For the purposes of spurring further discussion, one place to start might be to focus on AI systems that:

- Take decisions or actions affecting large-scale networked systems;
- Process or direct physical inputs and outputs;
- Operate autonomously or semi-autonomously; and
- Pose a significant potential risk of large-scale harm, including physical, economic, or environmental harm.

Second, the government should require system developers to ensure that safety brakes are built by design into the use of AI systems for the control of critical infrastructure.

System safety is a well-established discipline that we have put to work in the aviation, automotive, and nuclear sectors, among others, and it is one that we must bring to

bear to the engineering of AI systems that control critical infrastructure. We should establish a layered approach to AI safety, with the “safety brake” concept implemented at multiple levels.

While the implementation of “safety brakes” will vary across different systems, a core design principle in all cases is that the system should possess the ability to detect and avoid unintended consequences, and it must have the ability to disengage or deactivate in the event that it demonstrates unintended behavior. It should also embody best practice in human-computer interaction design.

Third, the government should ensure operators test and monitor high-risk systems to make certain that AI-systems that power critical infrastructure remain within human control.

Specific system testing will be needed in the context of a planned deployment for critical infrastructure. In other words, the use of an advanced AI model must be reviewed in the context of how it will be used in a specific product or service.

In accordance with system safety best practices, the system and each of its components should be tested, verified, and validated rigorously. It should be provable that the system operates in a way that allows humans to remain in control at all times. In practice, we anticipate that this will require close and regular coordination between a system operator, their AI infrastructure provider, and their regulatory oversight bodies.

Fourth, AI systems that control the operation of designated critical infrastructure should be deployed only in licensed AI infrastructure.

We believe it would be wise to require that AI systems that control the operations of higher-risk critical infrastructure systems be deployed on licensed AI infrastructure. This is not to suggest that the AI infrastructure needs to be a hyperscale cloud provider such as Microsoft. Critical infrastructure operators might build AI infrastructure and qualify for such a license in their own right. But to obtain such a license, the AI infrastructure operator should be required to design and operate their system to allow another intervention point—in effect, a second and

separate layer of protection—for ensuring human control in the event that application-level measures fail.

These proposals might leave some wondering how realistic or futureproof “safety brakes” are if we are on a path to developing AI systems that are more capable than humans. They might ask: couldn’t the AI system itself work around safety brakes and override them? Won’t the AI system know how humans will respond at every step of the way and simply work around those responses?

In posing those questions, it’s important to be clear about the facts as they stand today. Today’s cutting-edge AI systems like GPT-4 from OpenAI and Claude from Anthropic have been specifically tested—by qualified third-party experts from the [Alignment Research Center](#)—for dangerous capabilities, such as the ability to evade human oversight and become hard to shut down. Those tests [concluded](#) that GPT-4 and Claude do not have sufficient capabilities to do those things today.

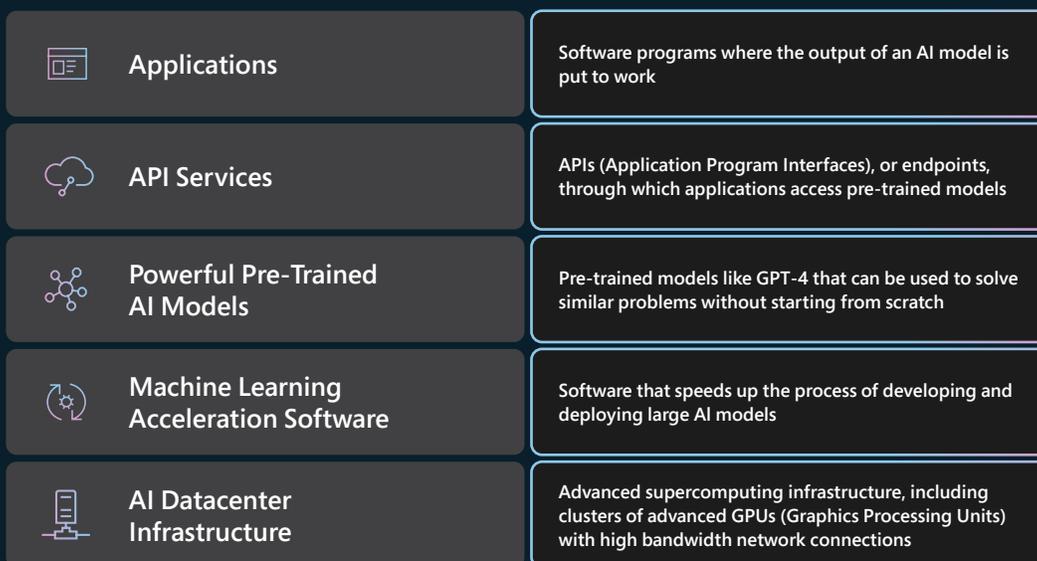
This rigorous testing and the conclusions drawn provide us with clarity as to the capabilities of today’s cutting-edge AI

models. But we should also heed the Alignment Research Center’s call for ongoing research on these topics and recognize the need for industry-wide commitment to AI capability evaluations. Put simply, we need to ensure that we have the right structures in place not only to understand the status quo, but to get ahead of the future. That is precisely why we need action with respect to the small but important class of highly capable AI models that are on the frontier—a topic that our next section addresses.

3. Develop a broad legal and regulatory framework based on the technology architecture for AI.

As we have given more thought to the various potential legal and regulatory issues relating to AI responsibilities, it has become more apparent that there will need to be a legal and regulatory architecture for AI that reflects the technology architecture for AI itself. In short, the law will need to place various regulatory responsibilities upon different actors based upon their role in managing different

The technology stack for AI foundation models





aspects of AI technology. For this reason, it’s helpful to consider some of the critical pieces that go into building and using new foundation AI models.

A grounding in the technology architecture for AI foundation models

Software companies like Microsoft build a “tech stack” with layers of technologies that are used to build and run the applications that organizations and the public rely upon every day. There’s no single right way to describe an AI tech stack, and there’s a good chance that any two developers will describe it differently. But for purposes of thinking about the future of AI regulation, a good way to start is to consider the chart on the previous page.

An advanced pretrained AI model like GPT-4 is shown on the third row above, in the middle of the stack. It’s created by developers and research scientists at a firm like OpenAI based on the two layers below it. In the case of GPT-4, OpenAI technical staff in San Francisco, California, did their model development work by harnessing the AI supercomputing infrastructure that Microsoft created and built exclusively for them in the datacenter complex shown

above, located just west of Des Moines, in Iowa.

As Microsoft [announced when it opened this datacenter](#) in March 2020, this datacenter contains a single supercomputing system that ranked upon opening in the top five supercomputers in the world. Built by Microsoft in collaboration with and exclusively for use by OpenAI to develop its GPT models, the supercomputing system has more than 285,000 Central Processing Unit (CPU) cores. (The CPU is perhaps the most fundamental component in any modern PC or laptop.) The system also has more than 10,000 of the most advanced Graphics Processing Units, or GPUs. Less advanced versions of such chips are contained in a gaming console or gaming laptop and can process a large number of mathematical equations simultaneously. Each GPU server in the datacenter has network connectivity that can process 400 gigabits of data per second.

As Microsoft Chief Technical Officer Kevin Scott said when we made this announcement in 2020, “the exciting thing about these [new GPT] models is the breadth of things they’re going to enable.” As OpenAI and Microsoft explained in 2020, machine learning experts had

historically built separate, smaller AI models with many labeled examples to learn a single task such as translating between languages.

But using this type of massive supercomputing infrastructure—and with the help of customized machine learning acceleration software—it became possible to create a single massive AI model that could learn by examining huge amounts of data, such as billions of pages of publicly available text. As Microsoft said in the 2020 announcement and as the world now recognizes in 2023, “this type of model can so deeply absorb the nuances of language, grammar, knowledge, concepts, and context that it can excel at multiple tasks: summarizing a lengthy speech, moderating content in live gaming chats, finding relevant passages across thousands of legal files or even generating code from scouring GitHub.”

As all this reflects, the core of what has struck some as the most surprising technological development of the decade was preannounced in plain and public view in just the third month as the decade began. The good news, at least from the perspective of Microsoft and OpenAI, is that we’ve been able to work the past several years to strengthen safety and security protocols to prepare for the more powerful AI models.

This brings one to how these large AI models are deployed for use. Given the very substantial computational resources required, these take place in multiple countries in advanced datacenters with large amounts of GPUs and advanced network connectivity, running in the case of GPT-4, on Microsoft’s Azure platform. This requires in its own right very substantial additional investments and deployment of the most advanced digital technology, but it does not require the same highly specialized infrastructure that is needed to build an advanced AI model in the first place.

The actual use of these models involves the top half of the technology stack. Users interact with a model like GPT-4 through an application, as shown at the top of the stack. ChatGPT, Bing Chat, and GitHub Copilot are all examples of such applications. Companies and organizations large and small will no doubt create new or modify existing applications to incorporate features and services that harness the power of generative AI models. Many will be

consumer applications, including those that are already household names. Many others will be created in-house by companies, governments, and nonprofits for their own internal use or by their customers. In short, a new wave of applications powered by generative AI will soon become part of daily life around the world.

Such applications access the capabilities of an AI model through endpoints called APIs, or Application Program Interfaces. APIs have long been one of the most important methods of accessing core technology building blocks that our customers are not running themselves on their infrastructure.

By way of illustration, Microsoft has created the Azure OpenAI Service to provide API access to OpenAI models like GPT-4. This API provides access to the model that is hosted on Microsoft’s infrastructure. In short, this means that our customers can harness the power of GPT-4 by building an application of their choosing and simply calling the API to submit prompts and receive outputs from GPT-4. There is no need for customers to maintain the sophisticated infrastructure that is needed to run an advanced model like GPT-4, and our customers benefit from Microsoft’s long-standing trust and compliance commitments, as well as the safety systems that we have built on top of the GPT-4 as part of the Azure OpenAI service.

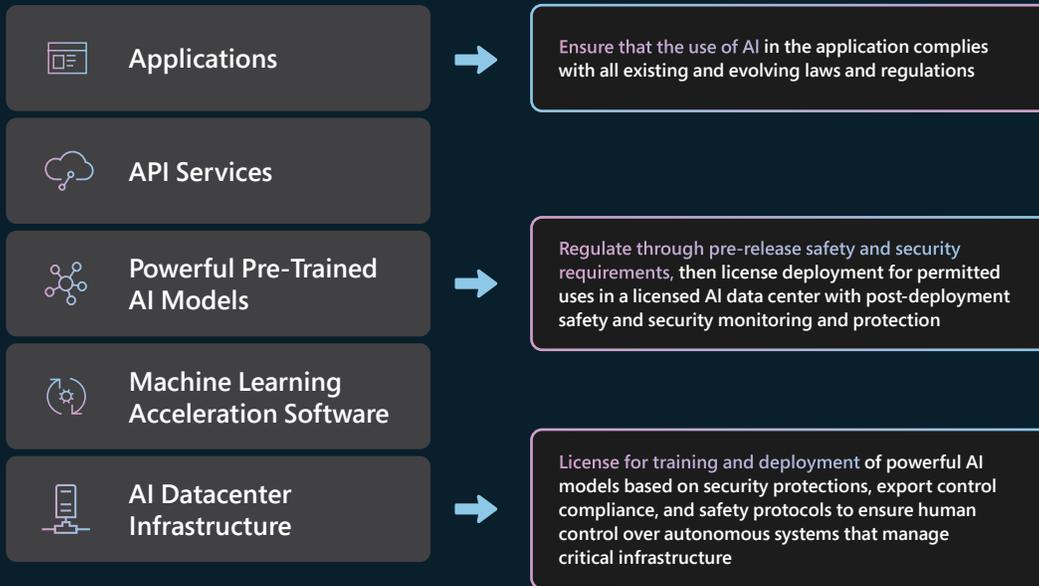
Creating a regulatory architecture that reflects AI’s technology architecture

It likely will make the most sense to design an AI regulatory architecture based on the AI technology architecture described below. At least as we’ve thought about these issues in recent months, we believe that law and regulation can probably have their most positive impact by focusing on three layers of the tech stack, with differing obligations at each level. The chart below illustrates this proposed approach, with further analysis and commitments we believe we can offer as a company to help advance these requirements.

Applying existing legal protections at the applications layer to the use of AI

For a great many individuals and organizations, the legal

A proposed AI regulatory architecture



rubber will meet the road as applications use AI to deliver information and services to others. This is the layer where the safety and rights of people will most be impacted, especially because the impact of AI can vary markedly in different settings. As a result, we will need the laws and regulations that govern conduct and societal impact to apply to applications that use the output from AI models to deliver services to individuals and organizations.

We have long had a wide variety of laws in place to protect the public. In the United States, many of these laws are grounded in long-standing societal values that go back to our Constitution and the Bill of Rights. Repeatedly over the past two centuries, our courts and agencies have adapted to uphold values we regard as timeless amidst constant technological change. Rapid advances in AI mean they will need to do so again.

The good news is that in many areas, we don't need new laws and regulations. We instead need to apply and enforce existing laws and regulations, and it has been encouraging to see several regulators around the world indicate that

they will do just that. This will be especially relevant to the many applications that are being created to use new and more powerful AI. And this will be important for companies and other organizations in every economic sector and in every country.

For example, it's unlawful for a bank to discriminate today based on race or gender when deciding who to approve for a mortgage. If a bank now wants to use AI to help it make its lending decisions, it will need to ensure that this does not lead to unlawful discrimination. And what's true for banks and mortgages is true in every field. Existing laws will continue to apply to the decisions and actions of organizations and individuals alike. No one is proposing a new defense to illegal conduct that will enable people to stand up in court and proclaim, "but Your Honor, a machine made me do it."

While this conclusion is simple, its consequences are profound. It means that every organization that uses AI needs to master not only the technology itself but the ability to evaluate how the technology impacts its wide-

ranging legal responsibilities. And courts and agencies alike will need to develop new capabilities to analyze how AI was used in a particular system.

We believe that several steps can help achieve this, including those we can take as a company:

First, we will work with our customers to help them apply state-of-the-art best practices to deploy AI lawfully and responsibly. One of the critical characteristics of AI is that the real-world impact on specific groups and issues is defined not just by the developer of an AI model or system, but also in its implementation in a specific service or application. In fact, in many circumstances it is only at the application level that it's possible to specifically identify and test for these real-world impacts before AI is deployed. As a result, responsibilities are often shared or even distributed, with different organizations needing to play different roles.

This helps explain why it's so important for customers that use AI in their services to develop their own capabilities to do so responsibly. This also explains why it is so important for a leading tech company to share information and lend their expertise on state-of-the-art best practices and tooling for responsible AI deployment.

We have been doing this type of work for two decades on other issues involving digital technology, including to implement legal compliance systems, advance cybersecurity, and protect privacy. We began five years ago to do similar work relating to artificial intelligence, and we will expand this initiative to work more broadly and deeply with our customers in the year ahead.

Second, we believe that regulatory agencies will need to add new AI expertise and capabilities. Very quickly, this need will reach virtually every agency in most governments in the world. For example, an agency like the Food and Drug Administration will need more AI experts who can help evaluate the use of cutting-edge AI systems by companies in something like the clinical trials for new drugs. Similarly, agencies like the Federal Aviation Administration will need additional AI experts to help evaluate the new uses of AI by aircraft manufacturers in developing new planes.

Generative AI itself will be a powerful tool that will better

enable regulatory agencies to evaluate the use of AI. This is because models like GPT-4 and services like ChatGPT, GitHub Copilot, and Microsoft M365 Copilot make it far easier for people to harness the power of AI to access data and evaluate it more quickly. As Google rightly recommended in a [new white paper](#) just last week, it will be important for governments to "direct sectoral regulators to update existing oversight and enforcement regimes to apply to AI systems, including on how existing authorities apply to the use of AI." Agencies will need the funding, staff, and commitment to put these new tools to work.

Third, we will support broad educational initiatives to make information about AI technologies and responsible AI practices available to legislators, judges, and lawyers. Finally, rapid AI advances are creating new pressures on those who make or help enforce the law to learn about new AI technologies and how they work. We witnessed a similar need when the personal computer first became popular in the 1980s. For example, judges needed to decide cases that started to turn, in part, on evidence about or involving PC software and hardware.

Beginning in the 1990s, Microsoft supported broad initiatives to share information about how this new technology worked. We continue to do this today in selected areas such as electronic discovery. The accelerating use of AI means that new such efforts will be needed. We will support this work, including by supporting bar associations and other public interest and civic groups and activities.

Developing new laws and regulations for highly capable AI foundation models

While existing laws and regulations can be applied and built upon for the application layer of the tech stack, we believe that new approaches will be needed for the two additional layers beneath that reflect the new and more powerful AI models that are emerging. The first of these is for the development of the most powerful new AI models, and the second is for the deployment and use of these models in advanced datacenters.

From our work on the frontiers of AI, we have seen a new class of model emerge. Highly capable foundation models are trained on internet-scale datasets and are effective out-

Microsoft commitments to an AI licensing regime

Microsoft will share our specialized knowledge about advanced AI models to help governments define the regulatory threshold

Microsoft will support governments in their efforts to define the requirements that must be met in order to obtain a license to develop or deploy a highly capable foundation model

Microsoft will support government efforts to ensure the effective enforcement of a licensing regime

of-the-box at new tasks—a model like GPT-4 allows you to create a never-seen-before image using words in one prompt, and a speech in the style of Franklin Roosevelt in the very next.

At the cutting-edge, the capabilities of these foundation models are at once very impressive and can be harder to predict. As the models have been scaled up, we have seen anticipated advances in capabilities, as well as surprising ones that we and others did not predict ahead of time and could not observe on a smaller scale. Despite rigorous prerelease testing and engineering, we've sometimes only learned about the outer bounds of model capabilities through controlled releases with users. And the work needed to harness the power of these models and align them to the law and societal values is complex and evolving.

These characteristics of highly capable models present risk surfaces that need to be addressed. To date, we have benefited from the high safety standards self-imposed by the U.S. developers who have been working at the frontiers of AI model development. But we shouldn't leave these issues of societal importance to good judgment and self-restraint alone. We need regulatory frameworks that anticipate and get ahead of the risks. And we need to acknowledge the simple truth that not all actors are well-

intentioned or well-equipped to address the challenges that highly capable models present. Some actors will use AI as a weapon, not a tool, and others will underestimate the safety challenges that lie ahead.

Last week, Sam Altman, the CEO of OpenAI, testified before Congress and called for the establishment of a licensing regime for this small but important class of highly capable models at the frontiers of research and development. As Microsoft, we endorse that call and support the establishment of a new regulator to bring this licensing regime to life and oversee its implementation.

First, we and other leading AI developers will need to share our specialized knowledge about advanced AI models to help governments define the regulatory threshold.

One of the initial challenges will be to define which AI models should be subject to this level of regulation. The objective is not to regulate the rich ecosystem of AI models that exists today and should be supported into the future, but rather the small number of AI models that are very advanced in their capabilities and in some cases, redefining the frontier. We refer to this small subset of models as highly capable AI models in this white paper.

Defining the appropriate threshold for what constitutes a highly capable AI model will require substantial thought, discussion, and work in the months ahead. The amount of compute used to train a model is one tractable proxy for model capabilities, but we know today that it is imperfect in several ways and unlikely to be durable into the future, especially as algorithmic improvements lead to compute efficiencies or new architectures altogether.

A more durable but unquestionably more complex proposition would be to define the capabilities that are indicative of high ability in areas that are consequential to safety and security, or that represent new breakthroughs that we need to better understand before proceeding further. Further research and discussion are needed to set such a capability-based threshold, and early efforts to define such capabilities must continue apace. In the meantime, it may be that as with many complex problems in life, we start with the best option on offer today—a compute-based threshold—and commit to a program of work to evolve it into a capability-based threshold in short order.

Second, we will support governments in their efforts to define the requirements that must be met in order to obtain a license to develop or deploy a highly capable AI model.

A licensing regime for highly capable AI models should be designed to fulfill three key goals. First and foremost, it must ensure that safety and security objectives are achieved in the development and deployment of highly capable AI models. Second, it must establish a framework for close coordination and information flows between licensees and their regulator, to ensure that developments material to the achievement of safety and security objectives are shared and acted on in a timely fashion. Third, it must provide a footing for international cooperation between countries with shared safety and security goals, as domestic initiatives alone will not be sufficient to secure the beneficial uses of highly capable AI models and guard against their misuse. We need to proceed with an understanding that it is currently trivial to move model weights across borders, allowing those with access to the “crown jewels” of highly capable AI models to move those models from country to country with ease.

To achieve safety and security objectives, we envision licensing requirements such as advance notification of large training runs, comprehensive risk assessments focused on identifying dangerous or breakthrough capabilities, extensive prerelease testing by internal and external experts, and multiple checkpoints along the way. Deployments of models will need to be controlled based on the assessed level of risk and evaluations of how well-placed users, regulators, and other stakeholders are to manage residual risks. Ongoing monitoring post-release will be essential to ensuring that guardrails are functioning as intended and that deployed models remain under human control at all times.

In practice, we believe that the effective enforcement of such a regime will require us to go one layer deeper in the tech stack to the AI datacenters on which highly capable AI models are developed and deployed.

Third, we will support government efforts to ensure the effective enforcement of a licensing regime for highly capable AI models by also imposing licensing requirements on the operators of AI datacenters that are used for the testing or deployment of these models.

Today’s highly capable AI models are built on advanced AI datacenters. They require huge amounts of computing power, specialized AI chips, and sophisticated infrastructure engineering, like Microsoft’s facilities in Iowa, described above. Such AI datacenters are therefore critical enablers of today’s highly capable AI models and an effective control point in a comprehensive regulatory regime.

Much like the regulatory model for telecommunications network operators and critical infrastructure providers, we see a role for licensing providers of AI datacenters to ensure that they play their role responsibly and effectively to ensure the safe and secure development and deployment of highly capable AI models. To obtain a license, an AI datacenter operator would need to satisfy certain technical capabilities around cybersecurity, physical security, safety architecture, and potentially export control compliance.

In effect, this would start to apply for AI a principle developed for banking to protect against money laundering and criminal or terrorist use of financial services.

The “Know Your Customer”—or KYC—principle requires that financial institutions verify customer identities, establish risk profiles, and monitor transaction to help detect suspicious activity.

In a similar way, it would make sense for a similar KYC principle to require that the developers of powerful AI models first “know the cloud” on which their models are deployed. The use of authorized and licensed AI datacenters would ensure that those who develop advanced models would have several vendors from which to choose. And it would enable the developer of an advanced model to build or operate their own cloud infrastructure as well, based on meeting the requisite technical standards and obligations.

The licensed AI datacenter operator would then need to meet ongoing regulatory requirements, several of which are worth considering.

First, operators of AI datacenters have a special role to play in securing highly capable AI models to protect them from malicious attacks and adversarial actors. This likely involves not just technical and organizational measures, but also an ongoing exchange of threat intelligence between the operator of the AI datacenter, the model developer, and a regulator.

Second, in certain instances, such as for scenarios that involve sensitive uses, the cloud operator on which the model is operating should apply the second aspect of the KYC principle – knowing the customers who are accessing the model. More thought and discussion will be needed to work through the details, especially when it comes to determining who should be responsible for collecting and maintaining specific customer data in different scenarios.

The operators of AI datacenters that have implemented know-your-customer procedures can help regulators get comfortable that all appropriate licenses for model development and deployment have been obtained. One possible approach is that substantial uses of compute that are consistent with large training runs should be reported to a regulator for further investigation.

Third, as export control measures evolve, operators of AI datacenters could assist with the effective enforcement

of those measures, including those that attach at the infrastructure and model layers of the tech stack.

Fourth, as discussed above, the AI infrastructure operator will have a critical role and obligation in applying safety protocols and ensuring that effective AI safety brakes are in place for AI systems that manage or control critical infrastructure. It will be important for the infrastructure operator to have the capability to intervene as a second and separate layer of protection, ensuring the public that these AI systems remain under human control.

These early ideas naturally will all need to be developed further, and we know that our colleagues at OpenAI have important forthcoming contributions on these topics too. What is clear to us now is that this multitiered licensing regime will only become more important as AI models on the frontiers become more capable, more autonomous, and more likely to bridge the digital-physical divide. As we discussed earlier, we believe there is good reason to plan and implement an effective licensing regime that will, among other things, help to ensure that we maintain control over our electricity grid and other safety-critical infrastructure when highly capable AI models are playing a central role in their operation.

4. Promote transparency and ensure academic and nonprofit access to AI.

One of the many AI policy issues that will require serious discussion in the coming months and years is the relationship and tension between security and transparency. There are some areas, such as AI model weights (which are components of a model that are core to a model’s capabilities), where many experts believe that secrecy will be essential for security. In some instances, this may even be needed to protect critical national security and public safety interests. At the same time, there are many other instances where transparency will be important, even to advance the understanding of security needs and best practices. In short, in some instances tension will exist and in other areas it will not.

Transparency as a critical ethical principle for AI

When Microsoft adopted ethical guidelines for AI in

Microsoft commitments to promote transparency for AI

Microsoft will release an annual transparency report to inform the public about its policies, systems, progress, and performance in managing AI responsibly and safely

Microsoft will support the development of a national registry of high-risk AI systems that is open for inspection so that members of the public can learn where and how those systems are in use

Microsoft will commit that it will continue to ensure that our AI systems are designed to inform the public when they are interacting with an AI system and that the system's capabilities and limitations are communicated clearly

We believe there is benefit in requiring AI generated content to be labeled in important scenarios so that the public "knows the content" it is receiving

2018, we made transparency one of our six foundational principles. As we've implemented that principle, we've learned that it's important to provide different types of transparency in different circumstances, including making sure that people are aware that they are interacting with an AI system. Generative AI makes this principle more important than in the past, and it's an area where ongoing research and innovation will be critical. To help spur new work in this area, Microsoft is making three commitments to the White House.

First, Microsoft will release an annual transparency report to inform the public about its policies, systems, progress, and performance in managing AI responsibly and safely.

Transparency reports have proven to be an effective measure to drive corporate accountability and help the public better understand the state-of-the-art and progress toward goals. Microsoft believes transparency reports

have a role to play in the responsible AI context too, and so we will release an annual transparency report to inform the public about our policies, systems, progress, and performance in managing AI responsibly and safely. If adopted across the industry, transparency reports would be a helpful mechanism for recording the maturing practice of responsible AI and charting cross-industry progress.

Second, Microsoft will support the development of a national registry of high-risk AI systems that is open for inspection so that members of the public can learn where and how those systems are in use.

Public trust in AI systems can be enhanced by demystifying where and how they are in use. For high-risk AI systems, Microsoft supports the development of a national registry that would allow members of the public to review an overview of the system as deployed and the measures taken to ensure the safe and rights-respecting performance of the system.

For this information to be useful to the public, it should be

expressed at the system level, providing details about the context of use, and be written for nontechnical audiences. To achieve this, the United States could implement the approach of several European cities in adopting the [Algorithmic Transparency Standard](#) and developing accessible explanations of how it uses AI (see, for example, the [City of Amsterdam’s Algorithm Register](#)).

Third, Microsoft will commit that it will continue to ensure that our AI systems are designed to inform the public when they are interacting with an AI system and that the system’s capabilities and limitations are communicated clearly.

We believe that transparency is important not only through broad reports and registries, but in specific scenarios and for the users of specific AI systems. Microsoft will continue to build AI systems designed to support informed decision making by the people who use them. We take a holistic approach to transparency, which includes not only user interface features that inform people that they are interacting with an AI system, but also educational materials, such as the [new Bing primer](#), and detailed documentation of a system’s capabilities and limitations, such as the [Azure OpenAI Service Transparency Note](#). This documentation and experience design elements are meant to help people understand an AI system’s intended uses and make informed decisions about their own use.

Fourth, we believe there is benefit in requiring AI-generated content to be labeled in important scenarios so that the public “knows the content” it is receiving.

This is the third part of the KY3C approach we believe is worth considering. As we are committing above for Microsoft’s services Bing Image Creator and Designer, we believe the public deserves to “know the content” that AI is creating, informing people when something like a video or audio has been originally produced by an AI model rather than a human being. This labeling obligation should also inform people when certain categories of original content have been altered using AI, helping protect against the development and distribution of “deep fakes.” This will require the development of new laws, and there will be many important questions and details to address. But the health of democracy and future of civic discourse will

benefit from thoughtful measures to deter the use of new technology to deceive or defraud the public.

Access to AI resources for academic research and the nonprofit community

We believe there is another element that adds to transparency and that deserves more prominent attention. This is the need to provide broad access to AI resources for academic research and the nonprofit community.

The high cost of computational resources for the training of large-scale AI models, as well as other AI projects, is understandably raising concerns in the higher education and nonprofit communities. We understand this issue well because Microsoft’s large technology investment in OpenAI in 2019 originated from precisely this need for OpenAI itself, due in part to its nonprofit status.

Basic research, perhaps especially at universities, has been of fundamental importance to the economic and strategic success of the United States since the 1940s. Much of the tech sector itself owes both its birth and ongoing innovation to critical basic research pursued in colleges and universities across the country. It’s a success story that has been studied and emulated in many other countries around the world. The past few decades have seen huge swaths of basic research in almost every field propelled by growing computing resources and data science. Unless academic researchers can obtain access to substantially more computing resources, there is a real risk that scientific inquiry and technological innovation will suffer.

Another dimension of this problem is also important. Academic researchers help ensure accountability to the public by advancing our understanding of AI. The public needs academics to pursue research in this area, including research that advances AI accountability by analyzing the behavior of the models the commercial sector is creating.

While new and smaller open-source AI models are emerging and clearly are important, other basic research projects involving AI will almost certainly require more computational power than in the past. And unless new funding sources come together to provide a more centralized resource for the academic community, academic research will be at risk. This has led us to offer two focused commitments:

First, Microsoft will support the establishment of the newly proposed National AI Research Resource (NAIRR) to provide computing resources for academic research and would welcome and support an extension to accommodate access by academic institutions in allied nations abroad, including Japan, the United Kingdom, the European Union, and other like-minded countries.

The National AI Research Resource has its origins in the National Initiative AI Act of 2020, passed by Congress. The Act called on the National Science Foundation, in consultation with the White House Office of Science and Technology Policy, to create a task force to create a roadmap for “a shared research infrastructure that would provide AI researchers and students with significantly expanded access to computational resources, high-quality data, educational tools, and user support.” This January, the Task Force completed its work, publishing a [final report](#) calling for the creation and funding of a federated mix of computational and data resources, testbeds, software, and testing tools, based on a platform that can reduce the barriers to participation in the AI research ecosystem and increase the diversity of AI researchers.

Microsoft supports the establishment of the National AI Research Resource and believes it to be of fundamental importance to the United States’ leadership in AI innovation and risk mitigation. We will collaborate with the National Science Foundation to explore participation in a pilot project to inform efforts to stand up the National AI Research Resource, including by facilitating independent academic research relating to the safety of AI systems.

We also would welcome and support an extension of the NAIRR to provide access by academic institutions in like-minded nations. Already we’re seeing similar and substantial interest in these other countries. For example, Japan’s recent “[National Strategy in the New Era of AI](#)” calls for work to expand the computing resources for public and private use. We believe that a multilateral AI research resource would accelerate existing efforts to establish global norms and interoperable approaches to risk mitigation, including those underway in the U.S.-EU Trade and Technology Council and the G7.

Second, we will increase investment in academic research programs to ensure researchers outside Microsoft can access the company’s foundation models and the Azure OpenAI Service to undertake research and validate findings.

This expanded commitment builds on the success of our Turing Academic Program and Accelerating Foundation Models Research Program. It is designed to help the academic community gain API-based access to cutting-edge foundation models from Microsoft, as well as OpenAI models via Microsoft’s Azure OpenAI Service. This will ensure that researchers can study frontier applications and the sociotechnical implications of these models. Microsoft will ensure that its program design accommodates API-based access by a diverse community of academic researchers, including researchers at Minority Serving Institutions across the United States.

An important complement to providing such access is the development of governance best practices for the academic community engaged in frontier research on applications and the safety and security implications of highly capable models. Microsoft would welcome the opportunity to develop such practices by supporting and collaborating with a multistakeholder group, including representatives across the academic community.

Third, Microsoft will create free and low-cost AI resources for use by the nonprofit community.

Finally, we deeply appreciate the critical role that nonprofit organizations play in addressing societal needs around the world. Given their role as great incubators of innovative solutions, we believe it is critical for nonprofits to have broad, easy, and inexpensive access to new AI models and features for their work. Microsoft Philanthropies, including its Tech for Social Responsibility arm, supports 350,000 nonprofits in the Microsoft Cloud. It provides more than \$4 billion annually in cash and technology donations and discounts to nonprofits worldwide, a figure comparable to one of the 10 largest government foreign aid budgets.

Last week we expanded this support by announcing AI solutions to Microsoft Cloud for Nonprofit. These AI solutions are designed to improve the ability of nonprofit organizations to optimize operations, engage with donors,

and manage campaigns. This is the first of several steps we will take to reduce technical and cost barriers and enable nonprofits to harness the latest advances in AI.

5. Pursue new public-private partnerships to use AI as an effective tool to address the inevitable societal challenges that come with new technology.

Finally, we believe there is enormous opportunity to bring the public and private sectors together to use AI as a tool to improve the world, including by countering the challenges that technological change inevitably creates. We are clear-eyed about the future and realize that some will seek to use AI as a weapon rather than a tool. And even when people of goodwill do their best, technological change inevitably creates unforeseen bumps in the road ahead.

But we've also learned from numerous efforts over the years what democratic societies can accomplish when they harness the power of technology and bring the public and private sectors together. Two examples are perhaps the most profound.

The first is the Christchurch Call, born from the tragic terrorist tragedy that took place in Christchurch, New Zealand, on March 15, 2019. The attack claimed the lives of 51 innocent Muslims at two mosques and was livestreamed worldwide. The internet provided a stage not only to broadcast the attack but perhaps provided an incentive to pursue the assault in the first place.

New Zealand Prime Minister Jacinda Ardern vowed that the world would learn from the attack and take steps to prevent technology from being used this way again. Partnering with French President Emmanuel Macron, she brought leading tech companies together to pursue concrete steps to prevent the livestreaming and internet distribution of similar violent attacks in the future. Exactly two months after the attack, on May 15, 2019, government and tech leaders met at the Elysée in Paris to sign the [Christchurch Call](#) and committed to collective action that has continued in the four years that have followed.

This work provided inspiration for the larger assault that

began when the Russian military unleashed waves of cyberattacks on Ukraine on February 23, 2022. As we [noted last year](#), this reflected an age-old lesson from history: countries wage wars using the latest technology, and the wars themselves accelerate technological change.

But the role of technology in the war in Ukraine has brought a new dimension to the defense not only of Ukraine, but of democracy itself. The war has required a new form of collective defense. It has pitted Russia, a major cyberpower, not just against an alliance of countries, but also against a coalition of tech companies and NGOs.

Across the tech sector, companies have stepped up to support Ukraine's remarkable tenacity and innovation. Individual and collective technology measures have sustained Ukraine's digital operations, defeated cyberattacks, documented war crimes, and enabled students to stay in school even when their schools are damaged or destroyed. Microsoft has now provided \$450 million of financial and technology assistance to Ukraine, an amount that is unprecedented in the company's history.

The lessons from the Christchurch Call and the war in Ukraine should guide us on the role of AI in the future. One key is to focus on specific problems that can benefit from new initiatives and concrete action. Another is to bring governments, companies, and NGOs together on an international basis not only to move faster, but to accomplish more than any single organization or even country can achieve on its own. Microsoft is committed to pursuing and supporting similar initiatives in the months ahead.

In recent years, there has been a growing focus on addressing the new risks to democracy and the public from the potential weaponization of AI to alter content and create "deep fakes," including videos. The concern about future technology is well-placed (although we are concerned that countries are doing too little to address foreign cyber influence operations that are prolific and impactful already). In short, we will all need to do more collectively to combat this type of threat.

As we do so, it will be important to start with important building blocks that exist already. One of the most important is the [Coalition for Content Provenance and](#)

[Authenticity](#), or C2PA. Co-founded by companies such as Adobe, the BBC, Intel, Microsoft, Sony, and Truepic, C2PA unifies the efforts of the Adobe-led [Content Authenticity Initiative](#) (CAI), which focuses on systems to provide context and history for digital media, and [Project Origin](#), a Microsoft- and BBC-led initiative that tackles disinformation in the digital news ecosystem.

As Microsoft's Chief Scientific Officer, Eric Horvitz, [said last year](#), success will require "a multi-pronged approach, including education aimed at media literacy, awareness, and vigilance, [with] investments in quality journalism." There will be opportunities in the coming months to take important steps together.

This week, Microsoft will deploy new state-of-the-art provenance tools to help the public identify AI-generated audio-visual content and understand its origin. At Build, our annual developer conference, we are announcing the development of a new media provenance service. The service will mark and sign AI-generated videos and images with metadata about their origin, enabling users to verify

that a piece of content was generated by AI. The service implements the [C2PA specification](#). Microsoft will initially support major image and video formats and release the service for use with two of Microsoft's new AI products, [Microsoft Designer](#) and [Bing Image Creator](#).

This is an important step, but just a single one. Fortunately, many others are moving forward with similar and critical measures. We will need the right combination of focused steps and broader initiatives.

Perhaps more than anything, a wave of new AI technology provides an occasion for thinking big and acting boldly. Important work is needed to use AI to protect democracy and fundamental rights, provide broad access to the AI skills that will promote inclusive growth, and use the power of AI to advance the planet's sustainability needs. In each area, the key to success will be to bring governments, respected companies, and energetic NGOs together.

There will be no shortage of opportunities or challenges. We need to seize the moment.

02

Responsible by Design: Microsoft's Approach to Building AI Systems that Benefit Society

Responsible by Design: Microsoft's Approach to Building AI Systems that Benefit Society

Microsoft's commitment to developing AI responsibly

For the past seven years, we have worked to advance responsible AI—artificial intelligence that is grounded in strong ethical principles. We have approached our work with a humble recognition that trust is not given but earned through action, and a deep understanding of our responsibility not just to Microsoft but our community more broadly. This has led us to be focused both on meeting our own commitments, and helping our customers and partners do the same.

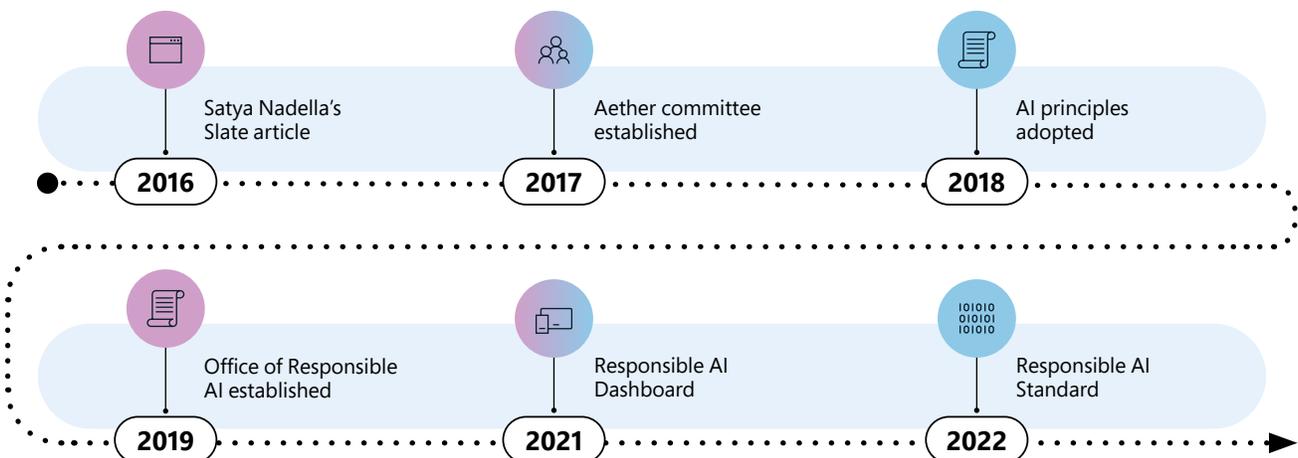
Our responsible AI journey began in 2016 with Satya Nadella, Microsoft's Chairman and CEO, sharing his [vision](#) of humanity empowered by AI. Satya expressed the beginnings of our core AI principles—values that endure today. Building on this vision, we launched [Microsoft's Aether Committee](#), comprised of researchers, engineers, and policy experts who provide subject matter expertise on the state-of-the-art and emerging trends with respect to our AI principles. This led to the creation and adoption of our AI principles in 2018.

We deepened our efforts in 2019 by establishing the Office of Responsible AI. This team coordinates the governance of our program, and collaborated across the company to write the first version of the Responsible AI Standard, a framework for translating high-level principles into actionable guidance for engineering teams building AI systems.

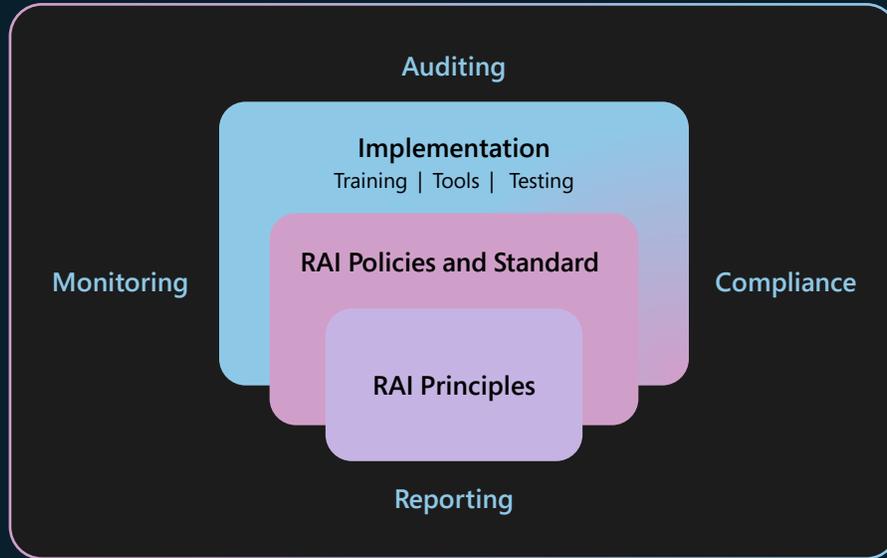
In 2021, we spoke publicly about the key building blocks that we had put in place to operationalize our program. We envisioned expanding training, processes, and tools to help us to implement and scale our responsible AI efforts. 2022 brought a new iteration of our Responsible AI Standard, evolving it into the version we use today, which we have also made publicly available. It sets out how Microsoft will build AI systems using practical methods to identify, measure, and mitigate potential risks ahead of time. This responsible-by-design approach establishes repeatable processes to minimize potential harms and magnify the benefits of AI from the outset.

We are proud of our progress over the last seven years. Those efforts have brought us to where we are today—deepening our commitment to embed safety and

Our Responsible AI Journey



Responsible AI Governance Framework



responsibility into the lifecycle of our AI systems. This is possible only when responsible AI principles and practices transcend traditional silos and multidisciplinary teams work together. With the opportunity and the potential risks at hand, we believe we must share what we have learned and help all organizations apply responsible AI practices to their work. That is precisely what we at Microsoft are doing, and we hope to lead by example.

Operationalizing Responsible AI at Microsoft

Setting foundational governance structures

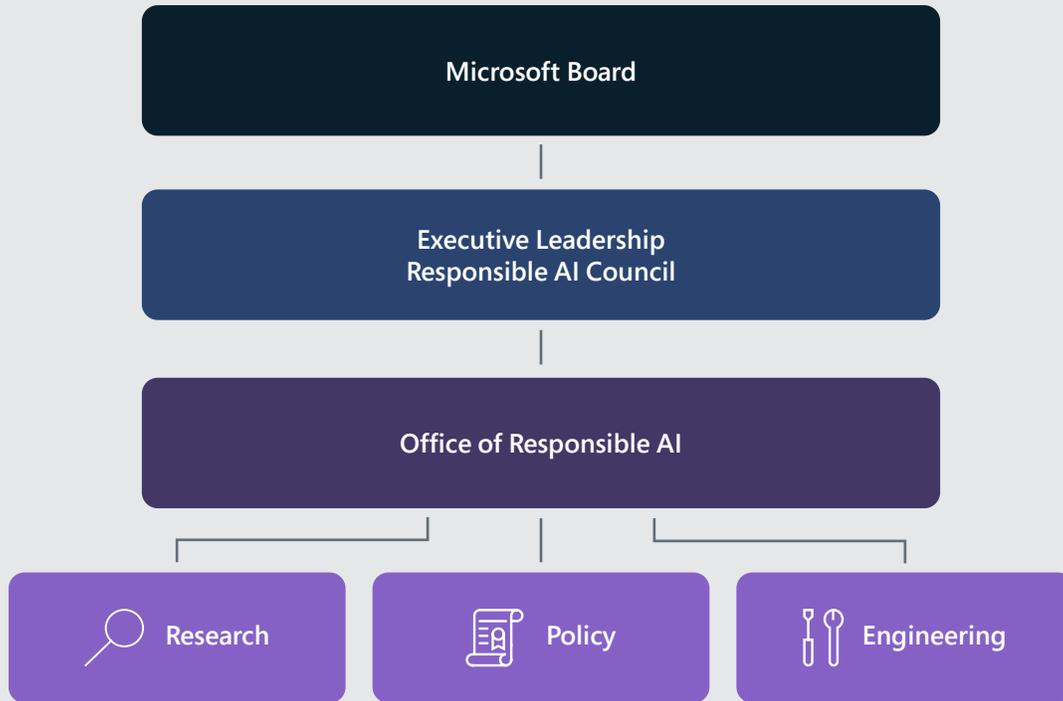
As the pace of AI continues to advance, we continue to evolve the governance structure we established to enable progress and accountability as a foundational piece of our responsible AI program. The creation of Microsoft’s governance structure—as well as the decision to scale responsible AI across the company—was driven by

leadership. Chairman and CEO Satya Nadella and the entire senior leadership team at Microsoft have made responsible AI a company-wide mandate.

Microsoft’s leadership recognized that a single team or discipline tasked with responsible AI would not be enough. Taking lessons from long-standing, cross-company commitments to privacy, security, and accessibility, we realized that responsible AI must be supported by the highest levels of leadership in the company and championed at every level across Microsoft.

To that end, Microsoft’s Office of Responsible AI developed a governance system that incorporates many diverse teams and functions across the company. At the working level, core teams within engineering, research, and policy play critical roles to advance responsible AI across the company, each bringing a set of unique skills. Responsible AI roles are also embedded within product, engineering, and sales teams by the appointment of “Responsible AI Champions” by leadership. Our Responsible AI Champions are tasked

Our ecosystem



with spearheading responsible AI practices within their respective teams, which means adopting the Responsible AI Standard, issue spotting and directly advising teams on potential mitigations, and cultivating a culture of responsible innovation. The Office of Responsible AI helps to orchestrate these teams across the company, drawing on their deep product knowledge and responsible AI expertise to develop a consistent approach across Microsoft.

At the next level, the Responsible AI Council is a forum for leadership alignment and accountability in implementing Microsoft’s responsible AI program. The Council is chaired by Microsoft’s Vice Chair and President, Brad Smith, and our Chief Technology Officer, Kevin Scott, who sets the company’s technology vision and oversees our Microsoft Research division. The Responsible AI Council convenes regularly, and brings together representatives of our core research, policy, and engineering teams dedicated to responsible AI, including the Aether Committee and the Office of Responsible AI, as

well as engineering leaders and senior business partners who are accountable for implementation.

At the highest level, the Environmental, Social, and Public Policy Committee of the Microsoft Board provides oversight of our responsible AI program. Our regular engagements with the Committee ensure the full rigor of Microsoft’s enterprise risk management framework is applied to our program.

The need for standardization

From crafting an AI system’s purpose to designing how people interact with it, we must keep people at the center of all AI decisions. While our responsible AI principles state the enduring values we seek to uphold, we needed more specific guidance on how to build and deploy AI systems responsibly. This is why we developed our [Responsible AI Standard](#), a more practical guide that memorializes a

set of rules of the road for our engineering teams so that upholding our AI principles is a daily practice.

The Responsible AI Standard provides engineering teams with actionable guidance on how to build AI systems responsibly. It was the result of a multi-year, cross-company effort that reflected a vast array of input from researchers, engineers, lawyers, designers, and policy experts. We consider it to be a significant step forward for our practice of responsible AI because it sets out much more concrete, practical guidance on how to identify, measure, and mitigate harms ahead of time. It also requires teams to adopt tools and controls to secure beneficial uses while guarding against potential misuses of their products.

There are two ways in which the Standard offers concrete direction to our engineering teams working across an AI product's lifecycle:

- **Articulating goals.** These define what it means to uphold the responsible AI principles. They break down a broad principle like accountability into definitive outcomes, such as ensuring AI systems are subject to impact assessments, data governance, and human oversight.

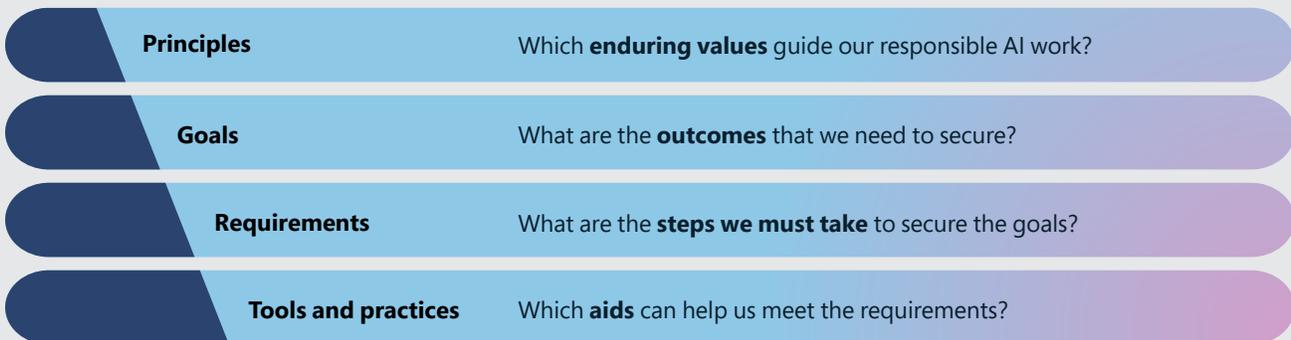
Red teaming AI systems

The term red teaming has historically described systematic adversarial attacks for testing security vulnerabilities. With the rise of large language models (LLMs), the term has extended beyond traditional cybersecurity and evolved in common usage to describe many kinds of probing, testing, and attacking of AI systems. With LLMs, both benign and adversarial usage can produce potentially harmful outputs, which can take many forms, including harmful content such as hate speech, incitement or glorification of violence, or sexual content.

Red teaming is an essential practice in the responsible development of systems and features using LLMs. Red teamers help to uncover and identify harms and, in turn, enable measurement strategies to validate the effectiveness of mitigations.

Microsoft has conducted red teaming exercises and implemented safety systems for its Azure OpenAI Service models and applications of these models in consumer products, such as Bing chat.

The Anatomy of the Responsible AI Standard



- Outlining a playbook.** These specific procedures and steps are required of teams throughout an AI system's lifecycle in order to achieve the goals set in our Responsible AI Standard. The steps map to available resources, tools, and practices to equip teams to make these goals a reality. For example, one of our Responsible AI Standard goals is to minimize the time to remediate predictable or known failures of an AI system, and to secure that goal, we ask teams to identify potential harms through iterative red teaming. We then ask teams to measure the prevalence of those harms and mitigate them by testing and implementing various tools and established strategies. This cycle of identifying, measuring, and mitigating potential harms of an AI system underpins many of the requirements in the Responsible AI Standard.

Ensuring checks and balances

When building and updating the Responsible AI Standard, we recognized early on that it is impossible to reduce all the complex sociotechnical considerations—for many different use cases—into an exhaustive set of pre-defined rules. This led us to create a program and process for ongoing review and oversight of high-impact cases and rising issues and questions, which we call Sensitive Uses.

Our Sensitive Uses program provides an additional layer of oversight for teams working on higher-risk use cases of our AI systems. The program began under the Aether Committee in 2018 and has operated out of the Office of Responsible AI since that office's inception in 2019. From July 2019 to May 2023, we have processed over 600 Sensitive Use case reviews from across Microsoft, including almost 150 cases during the period July 2022 to May 2023.

Think of the Sensitive Uses program as a reporting, review, and guidance framework: it starts with a mandatory reporting requirement, which then begins a hands-on responsible AI project review and consulting process with the Office of Responsible AI's Sensitive Uses team. It culminates in project-specific guidance and requirements that are additional to the Responsible AI Standard's baseline requirements. The Sensitive Uses review process is triggered when Microsoft personnel are involved in developing or deploying an AI system and the foreseeable use or misuse of that AI system could:

Responsible AI Standard in action: fairness in speech-to-text

Speech-to-text technology can improve individuals' lives, from sending hands-free texts to helping people with hearing loss communicate. Yet an academic study revealed that this technology produced nearly double the error rates for members of some Black and African American communities than for white users.

These results spurred us to take a closer look at the AI systems that power Microsoft speech-to-text technologies, to ensure they had been sufficiently trained on the rich diversity of speech patterns. We turned to an expert sociolinguist and expanded our data collection efforts to narrow the performance gap in our solutions. During the process, we involved stakeholders from outside Microsoft, such as experts and people from diverse communities.

These lessons were invaluable, and we've since incorporated them into the Responsible AI Standard to further articulate specific steps teams must take to ensure Microsoft AI systems are designed to provide a similar quality of service for identified demographic groups, including marginalized groups, to help us and other organizations harness the benefits of these technologies and avoid potential harms in the future.

1. Have a consequential impact on a user's legal status or life opportunities;
2. Present the risk of significant physical or psychological injury; or
3. Restrict, infringe upon, or undermine the ability to realize an individual's human rights.

Once reported, the Office of Responsible AI's Sensitive Uses team engages to triage and begin the review process with members of the project team, their Responsible AI Champion, and other relevant stakeholders. To help structure the review and drill into issues, we use not only artifacts such as the team's Responsible AI Impact Assessment and product documentation, but also close, ongoing interactions with the project team itself. During the review process, we also

often call on subject matter experts from across Microsoft through focused consultations. For particularly high-impact or novel-use cases, we elevate the project for review and advice from our Sensitive Uses Panel, which is a group of Microsoft experts spanning engineering, research, human rights, policy, legal, and customer-facing organizations from around the world. Our Sensitive Uses team is also multidisciplinary by design—its members have backgrounds in social sciences, law, engineering, and policy, and prior professional experiences as data scientists, academic researchers, policy analysts, lawyers, international diplomats, and machine learning engineers.

At the conclusion of its review, the Sensitive Uses team issues its requirements for the project to move forward. Again, these are additional requirements that go beyond our Responsible AI Standard and are tailored to the specific project at hand. We have even declined opportunities to build and deploy specific AI applications as a result of a Sensitive Uses review because we concluded that the projects were not sufficiently aligned with our Responsible AI Standard and principles. For example, Microsoft Vice Chair and President Brad Smith has spoken publicly about how, through our Sensitive Uses review process, we determined that a local California police department's real-time use of facial recognition on body-worn cameras and dash cams in patrol scenarios was premature, and he shared the fact that we turned down the deal. In addition to navigating the technical challenges presented by facial recognition operating in an uncontrolled environment, our Sensitive Uses review process helped us to form the view that there needed to be a societal conversation around the use of facial recognition and that laws needed to be established.

Another important outcome of the Sensitive Uses process was our [Limited Access policy](#) for more sensitive AI platform services, which adds an extra layer of scrutiny on the use and deployment of those services. Under this policy, we not only implement technical controls to mitigate risks, but also require potential customers to submit an application for use, disclose their intended use so that it meets one of our predefined acceptable use cases, and acknowledge that they have reviewed and agree to the terms of service. Only applications for uses that align with our responsible AI principles are approved.

Sensitive Uses review in action: Azure Custom Neural Voice

Azure AI's Custom Neural Voice is an innovative Microsoft speech technology that enables the creation of a synthetic voice that sounds nearly identical to the original source. This technology has already been used by enterprise customers such as AT&T and Progressive; it also shows potential in education, accessibility, and entertainment. Yet one can imagine possible abuses, such as inappropriately impersonating speakers and deceiving listeners.

Consistent with our measured approach for higher-risk AI systems, Custom Neural Voice has undergone several Sensitive Use reviews as it has evolved and progressed to broader availability. The review led us to adopt a layered control framework for Custom Neural Voice. For example, we limited customer access to the service, ensured acceptable use cases were defined and communicated through an application form, implemented speaker consent mechanisms, created specific terms of use, published transparency documentation detailing risks and limitations, and established technical guardrails to help ensure the speaker's active participation when creating a synthetic voice.

Through these and other controls, we are helping protect against misuse while maintaining beneficial uses of the technology. One such beneficial use includes what is known as "voice banking." Custom Neural Voice allows people who are at risk of losing their voice to "bank" their voice for later use, or in other words, recreate their voice by training a synthetic voice model through Custom Neural Voice.

Case study: Applying our Responsible AI approach to the new Bing

In February 2023, Microsoft launched the new Bing, an AI-enhanced web search experience. It supports users by summarizing web search results and providing a chat experience. Users can also generate creative content, such as poems, jokes, letters, and, with Bing Image Creator, images. The new AI-enhanced Bing runs on a variety of advanced technologies from Microsoft and OpenAI, including GPT-4,

a cutting-edge large language model (LLM) from OpenAI. Responsible AI teams across Microsoft worked with GPT-4 for months prior to its public release by OpenAI to develop a customized set of capabilities and techniques to join this cutting-edge AI technology and web search in the new Bing.

In preparing for the launch, Microsoft harnessed the full power of our responsible AI ecosystem. The new Bing experience has been developed in line with Microsoft's AI Principles, Microsoft's Responsible AI Standard, and in partnership with responsible AI experts across the company, including Microsoft's Office of Responsible AI, our engineering teams, Microsoft Research, and our Aether Committee.

Guided by our AI Principles and our Responsible AI Standard, we sought to identify, measure, and mitigate potential harms and misuse of the new Bing while securing the transformative and beneficial uses that the new experience provides. In the sections below, we describe our approach.

Identify

At the model level, our work began with exploratory analyses of GPT-4 in the late summer of 2022. This included conducting extensive red teaming in collaboration with OpenAI. This testing was designed to assess how the latest technology would work without any additional safeguards applied to it. Our specific intention was to produce harmful responses (responses are outputs from the AI system—in this case, a large language model—and may also be referred to as “completions,” “generations,” and “answers”), to surface potential avenues for misuse, and to identify capabilities and limitations. Our combined learnings advanced OpenAI's model development, informed our understanding of risks, and contributed to early mitigation strategies for the new Bing.

In addition to model-level red teaming, a multidisciplinary team of experts conducted numerous rounds of application-level red teaming on the new Bing AI experiences before making them available in our limited release preview. This process helped us better understand how the system could be exploited by adversarial actors and improve our mitigations. Non-adversarial testers also extensively evaluated new Bing features for shortcomings and vulnerabilities.

Measure

Red teaming can surface instances of specific harms, but in production, users will have millions of different kinds of conversations with the new Bing. Moreover, conversations are multi-turn and contextual, and identifying harmful responses within a conversation is a complex task. To better understand and address the potential for harms in the new Bing AI experiences, we developed additional responsible AI metrics specific to those new AI experiences for measuring potential harms like jailbreaks, harmful content, and ungrounded content. We also enabled measurement at scale through partially automated measurement pipelines.

Our measurement pipelines enable us to rapidly perform measurement for potential harms at scale, testing each change before putting it into production. As we identify new issues through the preview period and beyond, as well as ongoing red teaming, we continue to expand the measurement sets to assess additional harms.

Mitigate

As we identified and measured potential harms and misuse, we developed additional mitigations to those used for traditional search. Some of those include:

- **Preview period, phased release.** Our incremental release strategy has been a core part of how we move our technology safely from the labs into the world, and we're committed to a deliberate, thoughtful process to secure the benefits of the new Bing. Limiting the number of people with access during the preview period allowed us to discover how people use the new Bing, including how people may misuse it, before broader release. We continue to make changes to the new Bing daily to improve product performance, improve existing mitigations, and implement new mitigations in response to our learnings.
- **AI-based classifiers and metaprompting to mitigate harms or misuse.** The use of LLMs may produce problematic content that could lead to harms or misuse. Classifiers and metaprompting are two examples of mitigations that have been implemented in the new Bing to help reduce the risk of these types

of content. **Classifiers** classify text to flag different types of potentially harmful content in search queries, chat prompts, or generated responses. Flags lead to potential mitigations, such as not returning generated content to the user, diverting the user to a different topic, or redirecting the user to traditional search.

Metaprompting involves giving instructions to the model to guide its behavior. For example, the metaprompt may include a line such as “communicate in the user’s language of choice.”

- **Grounding in search results.** The new Bing is designed to provide responses supported by the information in web search results when users are seeking information. For example, the system is provided with text from the top search results and instructions via the metaprompt to ground its response. However, in summarizing content from the web, the new Bing may include information in its response that is not present in its input sources. In other words, it may produce ungrounded results. We have taken several measures to mitigate the risk that users may over-rely on ungrounded generated content in summarization scenarios and chat experiences. For example, responses in the new Bing that are based on search results include references to the source websites for users to verify the response and learn more. Users are also provided with explicit notice that they are interacting with an AI system and are advised to check the web result source materials to help them use their best judgement.
- **Limiting conversational drift.** During the preview period, we learned that very long chat sessions can result in responses that are repetitive, unhelpful, or inconsistent with new Bing’s intended tone. To address this conversational drift, we limited the number of turns (exchanges which contain both a user question and a reply from Bing) per chat session, until we could update the system to better mitigate the issue.
- **AI disclosure.** The new Bing provides several touchpoints for meaningful AI disclosure, where users are notified that they are interacting with an AI system as well as about opportunities to learn more about the new Bing.

Our approach to identifying, measuring, and mitigating harms will continue to evolve as we learn more—and as we make improvements based on feedback gathered during the preview period and beyond.

We share more details about our responsible AI work for the new Bing, including our efforts on privacy, digital safety, and transparency, at <https://aka.ms/ResponsibleAI-NewBing>.

Advancing Responsible AI through company culture

Procedures and standards are a critical part of operationalizing responsible AI and helps us build a culture committed to the principles and actions of responsible AI. These complementary approaches help us turn our commitments into reality.

Our people are the core of Microsoft culture. Every individual contributes to our mission and goals. To deepen our culture of advancing responsible AI, we invest in talent focused on AI and embed ownership of responsible AI in every role.

Investing in talent

Over the years, we have invested significantly in people as part of our commitment to responsible AI. We now have nearly 350 employees working on responsible AI, with more than a third of those dedicated to it full-time. These staff work in policy, engineering, research, sales, and other core functions, weaving responsible AI into all aspects of our business.

We ask teams who develop and use AI systems to look at technology through a sociotechnical lens. This means we consider the complex cultural, political, and societal factors of AI as they show up in different deployment contexts—and it represents a fundamental shift in the conventional approach to computer science. While the training and practices we have developed help teams foresee the beneficial and potentially harmful impacts of AI at the individual, societal, and global levels, this is not enough. Teams developing AI systems and the leadership to whom they answer could still have blind spots. That is why diversity and inclusion are critical to our responsible AI commitment.

The case for investing in a diverse workforce and an inclusive culture is well established, yet it is hard to overstate the

importance of diversity and inclusion for responsible AI. That is why our ongoing and increasing investment in our responsible AI ecosystem includes hiring new and diverse talent. As our annual [Diversity and Inclusion Report](#) shows, Microsoft continues to make incremental progress on diversity and inclusion. Yet, as an industry, we still have a long way to go. The field of AI continues to be predominantly white and male: only about one-quarter of employees working on AI solutions identify as women or racial or ethnic minorities, according to McKinsey's [2022 Global Survey on AI](#).

We will continue to champion diversity and inclusion at all levels, especially within our responsible AI program. To build AI systems that serve society as broadly as possible, we must recruit and retain a diverse, dynamic, and engaged employee community.

Embedding ownership of responsible AI in every role

We believe that everyone at Microsoft has the opportunity and responsibility to contribute to AI systems that live up to our responsible AI commitments. All employees, in every role, bring something to this work through their diverse skills, perspectives, and passions. This shift in perspective—that no matter your job title or team, everyone can advance responsible AI—requires a shift in culture.

To support this cultural growth, we have invested in developing employee skills and fostering collaboration.

Developing knowledge and skills

We have developed training and practices to empower our teams to think broadly about the potential impact of AI systems on individuals and society.

For example, when teams are at the earliest stages of designing an AI system, our Impact Assessment guides them through:

- Articulating the intended use(s) of the AI system;
- Interrogating how the AI system will solve the problem it is intended to solve;
- Identifying impacted stakeholders (and not just Microsoft's immediate customer);

Working toward a global, inclusive future for AI

The creation of AI systems and regulatory discussions around AI tend to be centered in advanced economies. Yet the responsible development and use of AI must reflect a diversity of global perspectives, including voices from developing countries.

At Microsoft, we strive to include developing countries in our advocacy for a globally coherent AI policy framework and globally relevant responsible AI practices. We are eager to share two examples of this commitment.

- **UNESCO Ibero-American Business Council:** Microsoft and Telefónica are co-chairing the effort to promote the adoption of UNESCO's Recommendation on the Ethics of Artificial Intelligence in Ibero-America. This represents the first globally coherent policy framework signed by all 193 UNESCO member states.
- **Responsible AI fellowship program:** This program brings together representatives from civil society, academia, and private and public sectors from developing countries. Launched by Microsoft and Stimson Center's Strategic Foresight Hub, it aims to advance the responsible development and use of AI. Fellows will contribute to a discussion series covering emerging best practices and the multifaceted impacts of AI in developing countries.

- Articulating potential harms and benefits that may affect each stakeholder; and
- Describing preliminary mitigations for potential harms.

To help teams conduct their Impact Assessment, the Office of Responsible AI has developed on-demand training, in-person workshops, and supporting guidance documents with examples and prompt questions. As part of our commitment to share best practices, our Impact Assessment template and guidance document are publicly available.

In our broader responsible AI training courses available to all Microsoft employees, we orient employees to Microsoft's

Responsible AI built into Azure Machine Learning



Fairness

Assess fairness and mitigate fairness issues to build models for everyone.



Explainability

Understand model predictions by generating feature importance values for your model.



Counterfactuals

Observe feature perturbations and find the closest datapoints with different model predictions.



Prompt Flow

Create workflows for large language-based applications to simplify prompt building, evaluation, and tuning.



Causal analysis

Estimate the effect of a feature on real-world outcomes.



Error analysis

Identify dataset cohorts with high error rates and visualize error distribution in your model.



Responsible AI scorecard

Get a PDF summary of your Responsible AI insights to share with your technical and non-technical stakeholders to aid in compliance reviews.



Azure Content Safety

Detect hate, violent, sexual, and self-harm content across languages in both images and text.

approach to responsible AI, including deep dives on our responsible AI principles and governance processes, and we provide content specifically tailored for data scientists and machine learning engineers.

Teams also have access to a wide range of responsible AI experts across the Microsoft ecosystem. They provide real-time engagement and feedback throughout the product lifecycle. This community includes the Aether Committee, the Office of Responsible AI, and a large and growing community of Responsible AI Champions who drive adoption of the Responsible AI Standard.

Fostering collaboration

We recognized early in our responsible AI journey the critical roles that researchers, policy experts, and engineers at Microsoft play in building our responsible AI practice. Each group brings insights and expertise vital to our work, and we strive to enable collaboration between them.

- Researchers, with a range of expertise from machine learning to the humanities, help us envision the leading edge of AI systems. They offer best practices in the identification, measurement, and mitigation of potential

harms posed by AI systems as well as insights into the exciting opportunities for AI innovation.

- Policy experts define and operationalize governance for responsible AI, including crafting the rules to guide the responsible development of AI systems. Our governance framework outlines roles and responsibilities across the organization in a way that creates accountability and encourages collaboration.
- Engineers design and develop AI systems that adhere to the Responsible AI Standard. They automate and scale the steps needed to identify, measure, and mitigate potential harms posed by AI systems. They also create wholly new responsible AI solutions that are feasible based on learnings.

Frequent collaboration and reliance on each other's expertise—practices reinforced by leadership—have helped us create a culture that leads to more beneficial and responsible solutions. Through ongoing dialogue, teams consistently report that a human-centered and collaborative approach to AI results in not just a responsible product, but a better product overall.

Responsible AI Champions

Meet the Microsoft Responsible AI Champions

Microsoft has cultivated a network of Responsible AI Champions across the organization. These individuals are essential in advancing a responsible-by-design culture.

Mihaela Vorvoreanu, Research



"Responsible AI is not only a technical problem with technical solutions. It requires collaborating deeply and early with not only responsible AI experts, but also people experts."

Alejandro Gutierrez Munoz, Data Science



"Championing of responsible AI is essential for aligning AI systems with ethical principles, fostering trust, ensuring compliance, and promoting social responsibility."

Shweta Gupta, Customer Engineering



"I believe that applying responsible AI principles by bringing together a diverse set of stakeholders while developing AI solutions not only helps us identify and address potential risks, but also ensures that the system being developed holistically supports its objectives."

Ferdane Bekmezci, Data Science



"It takes time to inculcate a culture to an organization. I am passionate about championing its adoption across the company because it's important to ensure that AI is developed and used in a way that is ethically and socially trustworthy."

Lisa Mueller, Design



"AI is changing rapidly, so growing communities and company-wide adoption around AI principles is important to build, grow, and extend trust in AI systems. As part of this approach, it is also important to include many disciplines to contribute to this effort and really makes a difference."

Empowering customers on their Responsible AI journeys

One of our most important responsible AI commitments is to help our customers on their responsible AI journey by sharing our learnings with them. Our efforts alone are not enough to secure the societal gains we envision when responsible AI practices are adopted.

As part of this commitment, we provide transparency documentation for our platform AI services in the form of Transparency Notes to empower our customers to deploy their systems responsibly. Transparency Notes communicate in clear, everyday language the purposes, capabilities, and limitations of AI systems so

our customers can understand when and how to deploy our platform technologies. They also identify use cases that fall outside the solution's capabilities and the Responsible AI Standard. Transparency Notes fill the gap between marketing and technical documentation, proactively communicating information that our customers need to know to deploy AI responsibly. You can see an example of our Transparency Note for the Azure OpenAI Service [here](#).

Customers also need practical tools to operationalize responsible AI practices. Over the years, responsible AI research at Microsoft has led to the incubation

Northumbria Healthcare NHS: Personalizing surgery assessments using AI

Opting for an elective surgery, such as a knee replacement, is a big decision. Patients turn to their medical providers to weigh the pros and cons. Clinicians in the United Kingdom's National Health Service (NHS) system used traditional statistics to analyze data, leading to general risk assessments they could apply to patients. Northumbria Healthcare NHS Trust surgeons envisioned a way to create more personalized assessments using AI.

The team consulted with Microsoft experts to build a surgery risk assessment model using Microsoft Azure Machine Learning and the Responsible AI Dashboard. The model analyzes 220 data points from patients, including smoking history and age. It helped uncover that platelet count carried a significantly higher weight in determining risk than expected. Built-in tools within the dashboard help avoid bias

and empower clinicians to understand how the model arrived at its results. This transparency enables doctors to explain risks and make recommendations so patients can make informed decisions and even take steps to reduce risks before a procedure.

The model empowers clinicians to move away from one-size-fits-all risk evaluations to an individualized understanding of each patient. The personalized assessments guide patient-doctor conversations, help providers assign patients to the surgery centers where they can get the appropriate level of care, and identify patients with heightened risks—before problems occur. While the model is first being used to assess candidates for joint replacement procedures, clinicians are expanding its use to other surgeries, too.

of tools such as Fairlearn and [InterpretML](#). The collection of tools has grown in capability, spanning many facets of responsible AI practice including the ability to identify, diagnose, and mitigate potential errors and limitations of AI systems. Since their original conception within Microsoft, these tools continue to improve and evolve externally through the contributions of active open-source communities. The collection of tools can be found under the [Responsible AI Toolbox](#) GitHub repository. Our latest tool, which is in preview, is [Azure Content Safety](#) which helps businesses create safer online environments and communities through models that are designed to detect hate, violent, sexual, and self-harm content across languages in both images and text.

Building on the Responsible AI Toolbox, Microsoft's responsible AI program has invested in integrating some of the more mature responsible AI tools directly into Azure Machine Learning so our customers will also benefit from the development of engineering systems and tools. The collection of capabilities, known as [the Responsible AI Dashboard](#), offers a single pane of glass for machine learning practitioners and business stakeholders to debug models and make informed, responsible decisions as they build AI systems or customize existing ones. Some of our

latest features added in preview include support for text and image data that enables users to evaluate large models built with unstructured data during the model-building, training, and evaluation stages, and Prompt Flow, which provides a streamlined experience for prompting, evaluating, and tuning large language models, including on measurements such as groundedness.

We have and will continue to invest in translating research-led responsible AI innovations into practical tools that support our customers on their responsible AI journeys.

The community involved in developing, evaluating, and using AI expands beyond our direct customers. To serve this broad ecosystem, we publicly share key artifacts from our responsible AI program, including our Responsible AI Standard, Impact Assessment template, and collections of cutting-edge research. Our digital learning paths further empower leaders to craft an effective AI strategy, foster an AI-ready culture, innovate responsibly, and more. These resources can be found online at <https://aka.ms/rai>.

Conclusion: Advancing Responsible AI Together

We've long said that advancing AI responsibly is a journey, and our own years-long effort to build a responsible AI program at Microsoft has prepared us for this AI inflection point. As we continue to unlock greater benefits from the latest AI technologies, we remain clear-eyed about risks and mindful of the important role we play in advancing the state-of-the-art, not only for AI capabilities but for responsible AI governance, mitigations, and culture-building.

Our governance approach begins with how we structure and organize responsible AI at Microsoft, with coordination from the Office of Responsible AI and essential involvement across every part of the company—core responsible AI teams in engineering, research, and policy, embedded Responsible AI Champions throughout organizations, executive leadership and accountability as embodied in the Responsible AI Council, and oversight from Microsoft's Board. Governance extends to creating, maintaining, and implementing a shared set of rules and policies to operationalize responsible AI, which we do with our Responsible AI Standard. It also requires additional oversight and expert guidance for higher-risk or novel-use cases like the development of the new Bing, which is where our Sensitive Uses program of required reporting and deeply engaged, case-specific review is so critical.

Cutting across all our work is the imperative to build and sustain culture and community. In addition to investing in existing people, hiring new talent, and developing training and skills-building, we have and will continue to prioritize diversity, collaboration, and the capacity to see AI systems through a sociotechnical lens. Finally, Microsoft is committed to proactive, practical steps that institutionalize not just a culture of responsible AI within the company, but tangible tools and capabilities that make AI safer and more reliable for our customers and society.

We will continue to be transparent and share our learnings broadly. We know that our efforts will require adjustments and course corrections, especially as we learn from those outside the company. As societal conversations and government oversight of AI evolve, we will continue to share our commitments for the responsible development and deployment of AI. We will also share our thoughts and suggestions about policy, regulation, and the role that private-public sector dialogue and partnerships can play, as we have done in our blueprint for AI policy, law, and regulation.

The current AI moment calls for industry, governments, academia, and civil society to come together to better define the boundaries for AI in society. We welcome a robust, global, cross-sector discussion of how to build and deploy safe, secure, and transparent AI systems. We hope that by sharing more details on our responsible AI efforts, we are contributing useful information to this conversation.

Together, we can build a future where AI advances the best of humanity.

