Microsoft

# Governing AI:
# A Blueprint for Japan

# Table of Contents

# Foreword

By Brad Smith,
Vice Chair and President
of 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 with Carol Ann Browne 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 use or hear about the power of OpenAI's GPT-4 foundation model, they are often surprised or even astounded. Many are enthused or even excited. Some are 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.

All these questions and even more will be critical in Japan. Few countries have been more resilient and innovative than Japan the past half century. Yet the remainder of this decade and beyond will bring new opportunities and challenges that would put technology at the forefront of public needs and discussion.

In Japan, one of the questions that's being asked is how to manage and support a shrinking and aging workforce. Japan will need to harness the power of AI to simultaneously address population shifts and other societal changes while driving its economic growth. This paper offers some of our ideas and suggestions as a company, placed in the Japanese context.

To develop AI solutions that serve people globally and warrant their trust, we've defined, published, and implemented ethical principles to guide our work. And we are continually 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 to responsible AI have given us the capability and confidence to see ever-expanding ways for AI to improve people's lives. 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.

While this technology will benefit us in everyday tasks by helping us do things faster, easier, and better, AI's real potential is in its promise to unlock some of the world's most elusive problems. 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.

We are optimistic about the innovative solutions from Japan that are included in Part 3 of this paper. These solutions demonstrate how Japan's creativity and innovation can address some of the most pressing challenges in various domains such as education, aging, health, environment, and public services.

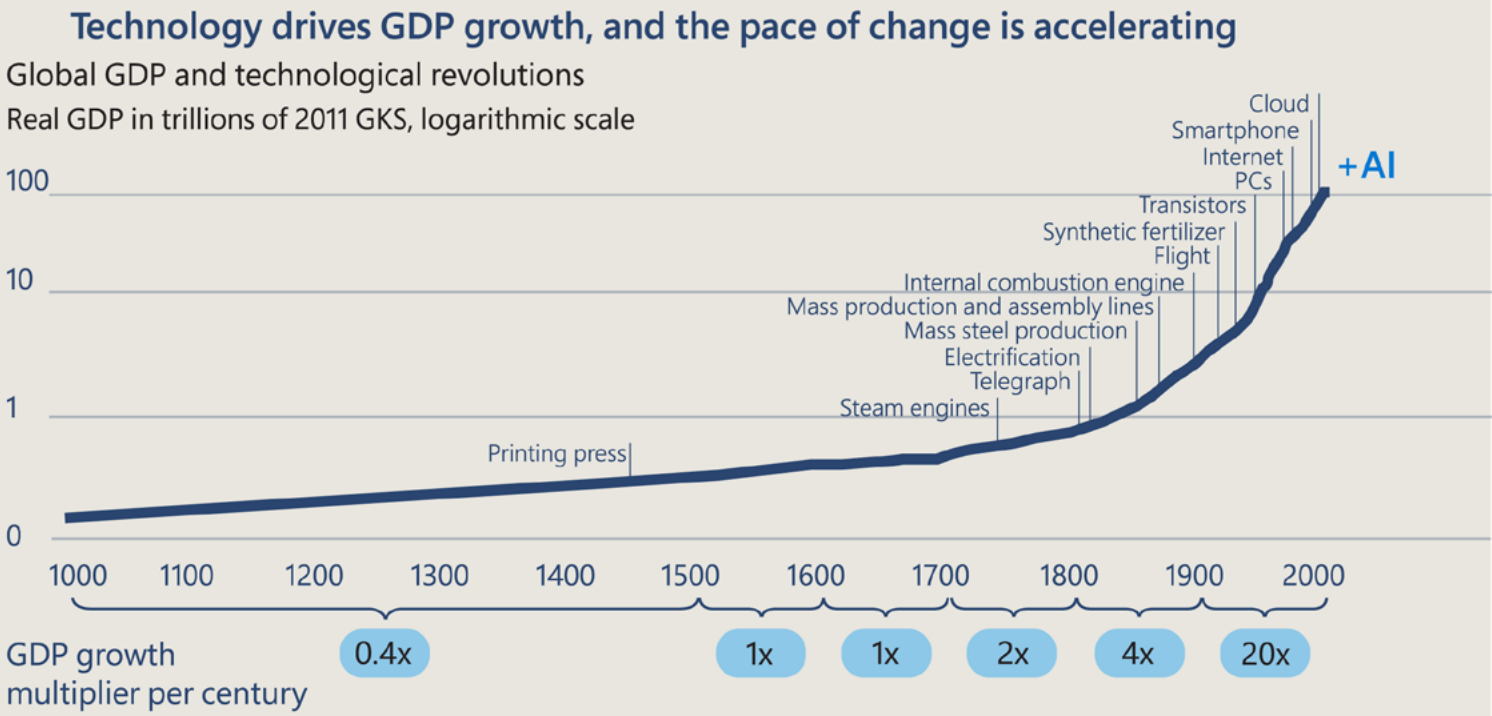
In so many ways, AI offers perhaps even more potential for the good of humanity than any invention that has preceded it. Since the invention of the printing press with movable type in the 1400s, human prosperity has been growing at an accelerating rate. Inventions like the steam engine, electricity, 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.

## 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.

We also believe that it is 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. The guardrails needed for AI require a broadly shared sense of responsibility and should not be left to technology companies




Source: Maddison Project Our World in Data[1]

alone. Our AI products and governance processes must be informed by diverse multistakeholder perspectives that help us develop and deploy our AI technologies in cultural and socioeconomic contexts that may be different than our own.

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 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?

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

## A five-point blueprint for the public governance of AI

Building on what we have learned from our responsible AI program at Microsoft, we released a blueprint in May that detailed our five-point approach to help advance AI governance. In this version, we present those policy ideas and suggestions in the context of Japan. We do so with the humble recognition that every part of this blueprint will benefit from broader discussion and require deeper development. But we hope this blueprint can contribute constructively to the work ahead. We offer 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.**

More broadly, to make the many different aspects of AI governance work on an international level, we will need a multilateral framework that connects various national rules and ensures that an AI system certified as safe in one jurisdiction can also qualify as safe in another. There are many effective precedents for this, such as common safety standards set by the International Civil Aviation Organization, which means an airplane does not need to be refitted midflight from Tokyo to New York.

As the current holder of the G7 Presidency, Japan has demonstrated impressive leadership in launching and driving the Hiroshima AI Process (HAP) and is well positioned to help advance global discussions on AI issues and a multilateral framework. Through the HAP, G7 leaders and multi-stakeholder contributors are strengthening coordinated approaches to AI governance and promoting the development of trustworthy AI systems that champion human rights and democratic values. Efforts to develop global principles are also being extended beyond G7 countries, including with organizations like the Organization for Economic Cooperation and Development (OECD) and the Global Partnership on AI.

The G7 Digital and Technology Ministerial Statement released in September 2023 recognized the need to develop international guiding principles for all AI actors, including developers and deployers of AI systems. It also endorsed a code of conduct for organizations developing advanced AI systems. Given Japan's commitment to this work and its strategic position in these dialogues, many countries will look to Japan's leadership and example on AI regulation.

Working towards an internationally interoperable and agile approach to responsible AI, as demonstrated by Japan, is critical to maximizing the benefits of AI globally. Recognizing that AI governance is a journey, not a destination, we look forward to supporting these efforts in the months and years to come.

## 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.

At Microsoft, we engage stakeholders from around the world as we develop our responsible AI program—it's important to us that our program is informed by the best thinking from people working on these issues globally and to advance a representative discussion on AI governance. It is for this reason that we're excited to participate in upcoming multistakeholder convenings in Japan.

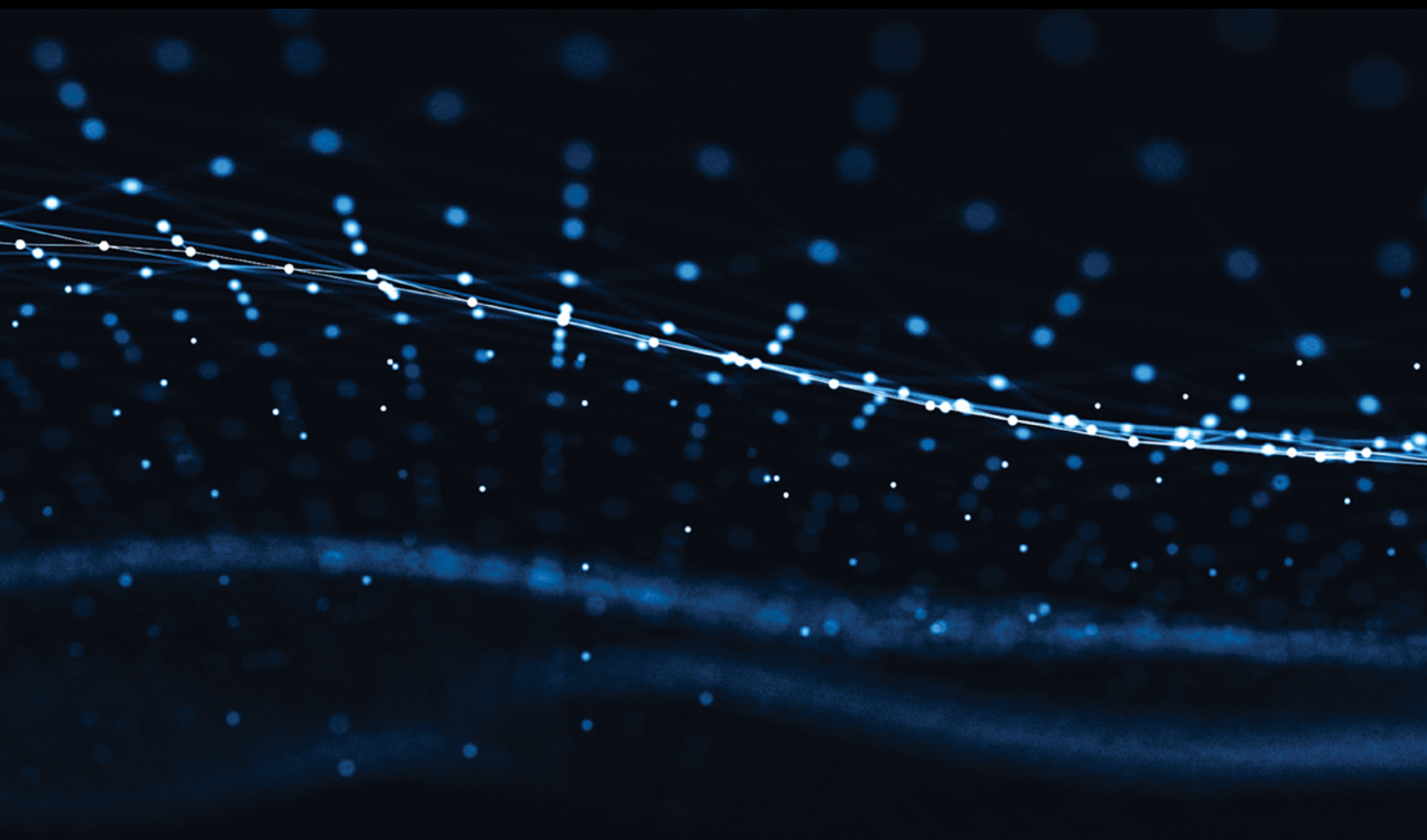
This October, the Japanese government will host the Internet Governance Forum 2023 (IGF) centered on the theme "The Internet We Want - Empowering All People". The IGF will include critical multistakeholder meetings to advance international guiding principles and other AI governance topics. We're looking forward to these and other meetings in Japan to learn from others and offer our experiences developing and deploying advanced AI systems, so that we can make progress toward shared rules of the road.

As another example of our multistakeholder engagement, earlier in 2023, Microsoft's Office of Responsible AI partnered with the Stimson Center's Strategic foresight hub to launch our Global Perspectives Responsible AI Fellowship. The purpose of the fellowship is to convene diverse stakeholders from civil society, academia, and the private sector in Global South countries for substantive discussions on AI, its impact on society, and ways that we can all better incorporate the nuanced social, economic, and environmental contexts in which these systems are deployed. A comprehensive global search led us to select fellows from Africa (Nigeria, Egypt, and Kenya), Latin America (Mexico, Chile, Dominican Republic, and Peru), Asia (Indonesia, Sri Lanka, India, Kyrgyzstan, and Tajikistan) and Eastern Europe (Turkey). Later this year, we will share outputs of our conversations and video contributions to shine light on the issues at hand, present proposals to harness the benefits of AI applications, and share key insights about the responsible development and use of AI in the Global South.

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 that keep people at the center of AI systems globally, we believe it can.

Brad Smith
Vice Chair and President, Microsoft



# Part 1
# Governing AI in Japan

## A five-point blueprint for the public governance of AI

Around the world, governments are looking for or developing frameworks to govern AI. Of course, there is no single or right approach. We offer here a five-point approach to help advance AI governance 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 initiatives, 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.**

**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

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

One of the most effective ways to accelerate government action is to build on existing or emerging governmental frameworks to advance AI safety. A key element to ensuring the safer use of this technology is a risk-based approach, with defined processes around risk identification and mitigation as well as testing systems before deployment.

For example, the AI Risk Management Framework developed by the U.S. National Institute of Standards and Technology (NIST) and launched earlier in 2023 provides a strong template for advancing AI governance. 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. Microsoft has committed to implementing the NIST AI Risk Management Framework, and we're not alone in our high regard for NIST's approach, as numerous governments, international organizations, and leading businesses have also validated the value of the new AI Risk Management Framework. Similarly, new international standards are in development, including the forthcoming ISO/IEC 42001 on AI Management Systems, which is expected to be published in the coming months. This will provide another important framework for companies and governments alike to put to work in advancing responsible AI.

We believe there is an opportunity for governments to help accelerate progress around these frameworks, using both carrots and sticks. In many countries, government procurement mechanisms have repeatedly demonstrated their value in improving the quality of products and advancing industry practice more generally. Governments could explore inserting requirements related to the AI Risk Management Framework or other relevant international standards into their procurement processes for AI systems, with an initial focus on critical decision systems that have the potential to meaningfully impact the public's rights, opportunities, or access to critical resources or services.

Many governments have either published or are in the process of publishing national strategies on AI, and Japan is no exception. The Japanese government is in the process of revisiting three existing AI guidelines (the AI Research and Development Guidelines adopted in 2017, the AI Utilization Principles Guidelines adopted in 2019, and the Governance Guidelines for Implementation of AI Principles adopted in 2022). It also plans to compile new AI guidelines by the end of 2023, with the goal of building trustworthy AI aligned with democratic values. According to the plan released in September

2023, the new AI guidelines will adopt a risk-based approach, call for agile governance, and set a baseline governance approach for all AI actors, including AI developers, AI deployers, and AI users.

In April 2023, Japan's Liberal Democratic Party released a national AI strategy document, the AI White Paper, covering recommendations to regulate as well as facilitate the adoption of AI technologies in Japan. The white paper outlines strategies and policies for AI regulation that consider emerging AI technologies, techniques, and trends, including large language models, infrastructure development, and risk assessments. Furthermore, it proposes ideas to increase AI adoption in public sector organizations and AI utilization in the public and private sectors. The review also considers novel approaches to AI regulation in high-risk sectors, including education, and promotes the implementation of an agile governance approach to AI regulation.

Beyond a national AI strategy, in May 2022, Japan passed the Economic Security Promotion Act (ESPA), a comprehensive economic security law that aims to enhance Japan's national security by regulating economic activity that could have security implications. It identified and brought into scope fourteen critical infrastructure sectors. The screening system under ESPA may apply to AI tools if they are critical to those sectors' operations.

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. As part of this, Microsoft recently joined with OpenAI, Google, and Anthropic to launch the Frontier Model Forum to continue to advance safety best practices around highly capable models, including how to evaluate and address the risks these models may present. We look forward to continuing to work with others and further advance frameworks and practices for responsible AI.

## 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

## 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

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.

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.** 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 of 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 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.

## The technology stack for AI foundation models

| | |
|---|---|
| **Applications** | Software programs where the output of an AI model is put to work |
| **API Services** | APIs (Application Program Interfaces), or endpoints, through which applications access pre-trained models |
| **Powerful Pre-Trained AI Models** | Pre-trained models like GPT-4 that can be used to solve similar problems without starting from scratch |
| **Machine Learning Acceleration Software** | Software that speeds up the process of developing and deploying large AI models |
| **AI Datacenter Infrastructure** | Advanced supercomputing infrastructure, including clusters of advanced GPUs (Graphics Processing Units) with high bandwidth network connections |

### 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 built and today operates exclusively for them. As Microsoft announced upon its opening in March 2020, this datacenter contains a single supercomputing system that then ranked in the top five supercomputers in the world. 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.

By 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 over 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 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.

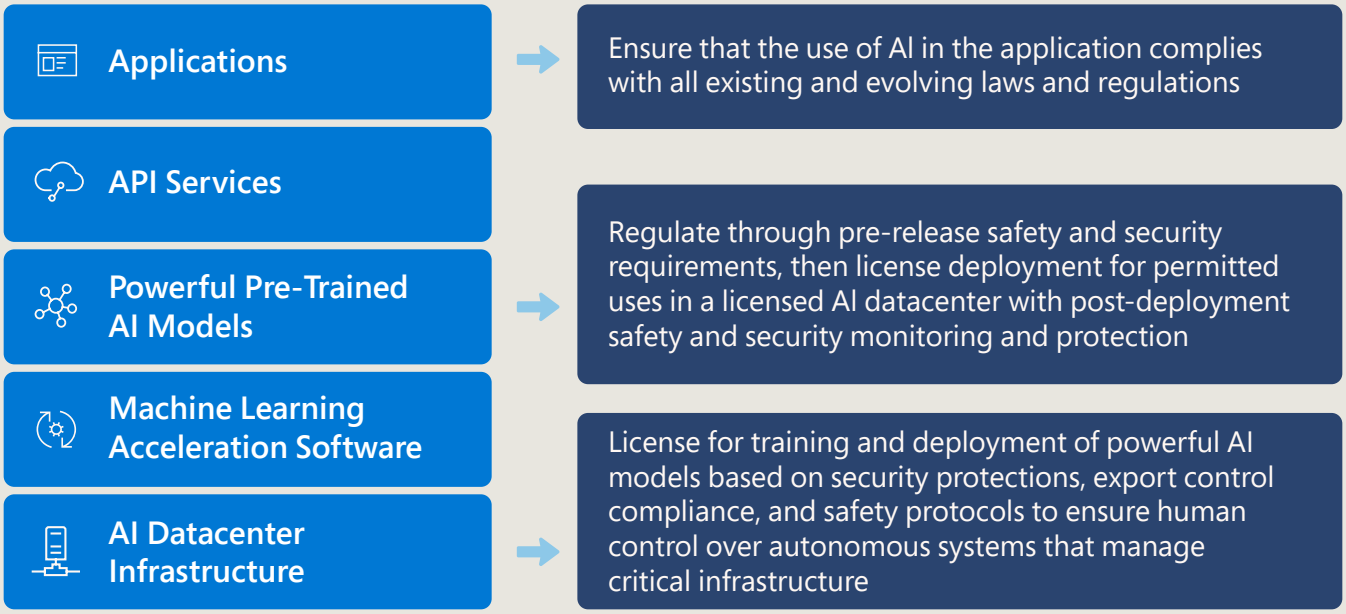
Such applications access the capabilities of an AI model through endpoints called APIs, or Application Program Interfaces. APIs have long been one of the most important methods of accessing core technology building blocks that our customers are not running themselves on their infrastructure.

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 longstanding 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

We believe it is sensible to design an AI regulatory architecture that roughly corresponds to the AI technology architecture described below. 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.

## A proposed AI regulatory architecture



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

For a great many individuals and organizations, the legal 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.

Japan has long had a wide variety of laws in place to protect the public, and Japan continues to explore how these laws can be adjusted to keep pace with change. Japan was swift in amending its copyrights act, back in 2018, to add an exception for the use of copyrighted works in machine learning. With the rise of generative AI in early 2023, the Agency for Cultural Affairs, as the agency responsible for the copyrights act, observed an opportunity to enhance societal understanding and conducted a public, online briefing on AI and copyright, which included a discussion of the machine learning exception mentioned above.

The good news is that in many areas relating to the impact of AI on society, 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 the discussion in Japan on this and the way in which 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. 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 who 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 implementing 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, a regulatory agency that oversees pharmaceuticals and medical devices will need more AI experts who can help evaluate the use of cutting-edge AI systems by companies in clinical trials for new drugs. Similarly, the Japan Civil Aviation Bureau (JCAB) will need additional AI experts to help evaluate new uses of AI by aircraft manufacturers building 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 recent white paper, 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.

Important public policy discussions continue around the world to advance ideas about how to apply existing law, upskill regulators, and address any remaining regulatory gaps. Within its April 2023 AI White Paper, Japan's Liberal Democratic Party called out the need for agile regulation, stating that agility is critical in order to "keep pace with the rapid rate of progress in AI technology". In other jurisdictions such as India, the Digital India Act is attempting to develop an integrated regulatory framework for AI that will take a comprehensive, risk-based approach and apply its most stringent requirements to AI systems that present the highest levels of risk to safety and security.

## 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

### 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 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 a famous historical figure 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 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.

Sam Altman, the CEO of OpenAI, testified before the United States 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 these frameworks will ultimately 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 transactions 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.

### Transparency as a critical ethical principle for 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.

When Microsoft adopted ethical guidelines for AI in 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.

**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 in transparency reports.

**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, one 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

## 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

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. This will require the development of new laws, and there will be many important questions and details to address. For example, 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.

Fortunately, there is an opportunity to use existing technical building blocks for AI transparency in addition to creating new transparency reporting initiatives. One of these is the Coalition for Content Provenance Authenticity, or C2PA, a global standards body with more than 60 members including Adobe, the BBC, Intel, Microsoft, Publicis Groupe Sony, and Truepic. The group is dedicated to bolstering trust and transparency of online information, including releasing the world's first technical specification for certifying digital content in 2022, which now includes support for generative AI. Leveraging the C2PA specification, Microsoft recently announced at our Build 2023 conference that we will deploy new state-of-the-art provenance tools to help the public identify AI-generated audio-visual content and understand its origin. 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. While this is an important step, it is only one step in what needs to be a broader effort to protect information integrity.

### Access to AI resources for academic research and the nonprofit community

Lastly, 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.

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 three focused commitments.

**First, Microsoft will support the establishment of the newly proposed National AI Research Resource (NAIRR) in the US to provide computing resources for academic research and would welcome and support an extension to accommodate access by academic institutions internationally.** The US is advancing the National AI Research Resource, "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." Microsoft supports the establishment of this type of research resource and believes it is important in advancing understanding around the opportunities and risks of AI.

We also would welcome and support an extension of the NAIRR to provide access to academic institutions internationally. We're already seeing similar and substantial interest in this type of resource among other countries around the world. This includes Japan, where the AI Strategy Council's Tentative Summary of AI Issues urges the government to develop computing resources and data-related infrastructure to help academics conduct AI research. This is consistent with the LDP's AI white paper, which suggests that government can fund access to computational resources so that any resource-intensive research can be conducted.

**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. 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 and engage with donors.

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

One lesson from recent years is that democratic societies often can accomplish the most 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. AI is an extraordinary tool with incredible potential for good. Like other technologies, though, it can be used as a weapon, and there will be some around the world who will seek to use it that way. However, we can also leverage AI in the fight against those who abuse it and to address societal challenges. We must work together through public and private partnerships to do this.

Specifically, important work is needed now to use AI to strengthen 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, industry, and NGOs together to advance them. Microsoft will do its part in each area.

The Japanese government is on the forefront of this work and will host the Internet Governance Forum (IGF), to be held in Kyoto, Japan in October 2023. This important forum will provide a key opportunity to collect multistakeholder perspectives on AI policy at both the national and international levels.

# Part 2
# Responsible by Design: Microsoft's Approach to Building AI Systems that Benefit Society

## Microsoft's commitment to developing AI responsibly

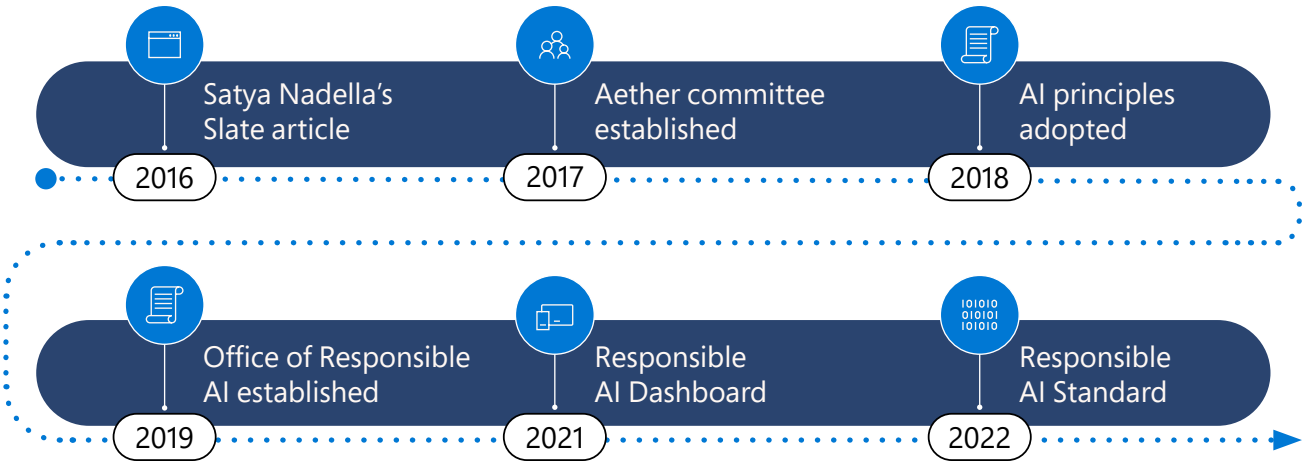
For the past seven years, we have worked to advance responsible AI—artificial intelligence that is grounded in strong ethical principles. We have approached our work with a humble recognition that trust is not given but earned, and our responsibility is not just to Microsoft but our community more broadly. 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.

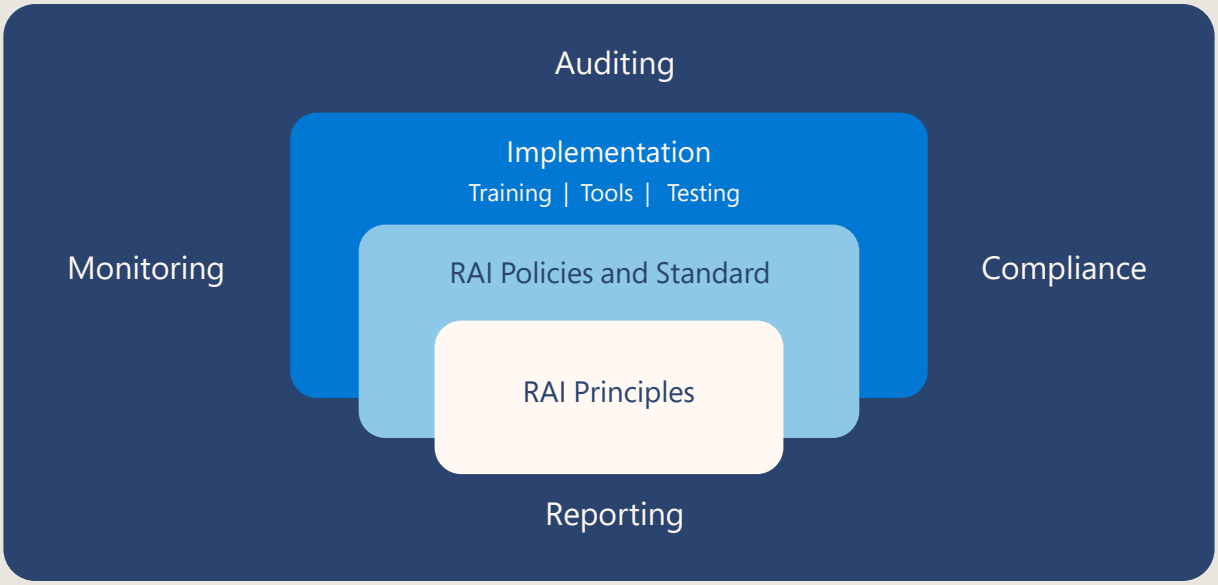
Building on this vision, we launched Microsoft's Aether Committee, comprised of researchers, engineers, and policy experts who provide subject matter expertise on the state-of-the-art and emerging trends with respect to our AI principles. This led to the creation and adoption of our AI principles in 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.

### Our Responsible AI Journey



## Responsible AI Governance Framework




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 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 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 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.

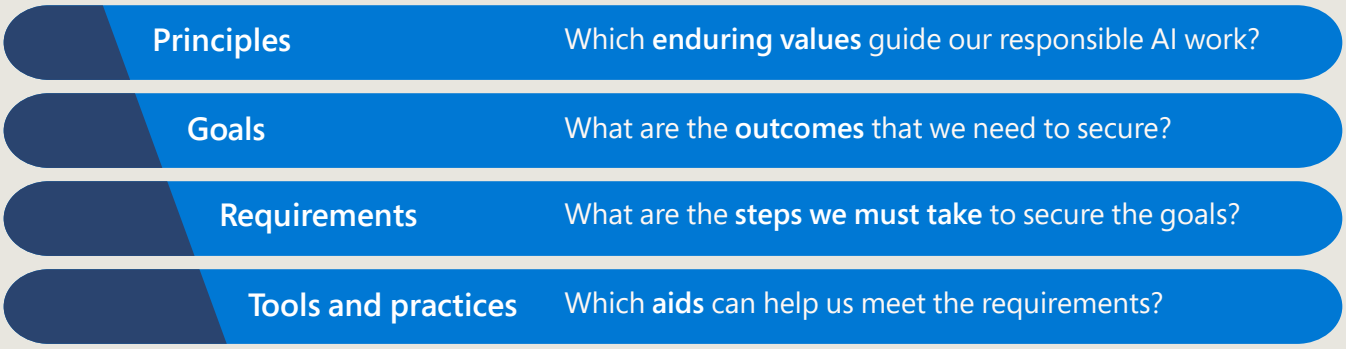
### The need for standardization

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.




- **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:

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.

## 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 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 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 help 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)
- 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 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.

## 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.

### 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.

## 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."

- 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 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.

## Empowering customers on their responsible AI journey

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 online.

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

# Part 3
# AI in Action in Japan

## How AI is addressing societal challenges

AI provides a huge opportunity to countries around the world to address major societal challenges. Below are some examples of how Japanese innovators are already utilizing AI to drive change in areas as varied as health, aging, environmental sustainability, education, and public services.

We are ready to accompany Japan's journey to unleash the possibility of generative AI. Microsoft, together with Kobe city and other partner companies, will open the Microsoft AI Co-Innovation Lab in Kobe city on October 11, 2023. Microsoft AI Co-Innovation Labs offer any interested company access and facilities to build, develop, prototype, and test solutions and Microsoft AI technology experts can be directly involved in this co-creation work. This Kobe City lab is the fifth one to be established by Microsoft worldwide, and will serve as the incubator of innovative solutions to social problems, global-level business, and future AI talent. We hope to see more use cases leveraging AI originating from this Co-Innovation Lab in the future.

## AI for a healthier future

### FRONTEO

The number of dementia patients in Japan is estimated to reach 7.3 million people (approximately 1 in 5 elderly individuals) by the year 2025. In Japan, a super-aging society, dementia care is an urgent national issue that needs to be addressed. But diagnosing dementia is challenging, as it requires specialized knowledge and experience, which can hinder early diagnosis and treatment.

To address this challenge, FRONTEO is developing an AI system for supporting dementia diagnosis on Microsoft Azure and AWS. The AI integrated model can identify signs of dementia from as little as 5-to-10 minutes of text data from daily conversations between patients and medical professionals by analyzing the speech content and word choice tendencies. Considering the rapid aging of Japan's population, early detection is crucial for ensuring timely treatment of dementia at specialized medical institutions. This innovative service not only reflects the industry's collective efforts to tackle a national issue exacerbated by an aging population but also signifies notable progress in digital healthcare in Japan.

The research and development project which served as the foundation of this dementia detection system was supported by the Japan Agency for Medical Research and Development (AMED) funded research program. FRONTEO obtained patent approval for its dementia diagnosis support AI system from the Japan Patent Office in 2020 and is now working to obtain manufacturing approval.

### Japan Agency for Medical Research and Development (AMED)

Japan is actively promoting the Full Genome Analysis Implementation Plan, initially established by the Ministry of Health, Labour, and Welfare in December 2019, to advance genomic medicine. The Japan Agency for Medical Research and

Development (AMED) has undertaken a significant endeavor, conducting full genome analysis for 12,000 cases between 2021 and 2022, with ambitions to scale up to over 10,000 cases annually. Given the substantial data volume involved, where a single patient's data can exceed 400GB, robust infrastructure is essential. In response, AMED is planning to transition the platform from an on-premises server to the cloud, harnessing Microsoft's AI capabilities for efficient data management and rapid scaling. AMED sees in Microsoft's AI technology promises for broader healthcare management and disease treatment efforts in Japan.

## AI for a more sustainable future

### AGRIST

AGRIST's journey in robot development began with the voice of farmers saying: 'We need harvesting robots to solve the shortage of human resources for harvesting'. The Japanese startup AGRIST's mission is to create sustainable business models to address agricultural challenges through technology and to achieve sustainable agriculture that will continue for the next 100 years. AGRIST develops automatic harvesting robots, through which they collect crop image data and transform it into agricultural big data. From there, they aim to establish a new agricultural system that analyzes conditions, such as changes in harvest quantity and the occurrence of pests and diseases, as big data.

Microsoft Azure acts as the foundation for data accumulation and analysis, leveraging the power of AI, and AGRIST is a valued member of the Microsoft for Startups program. In its next evolution, AGRIST is embarking on the development of an operating system called "agriss" to increase the yield of agricultural crops. By analyzing farm big data collected from robots using AI, they aim to achieve data-driven, replicable, and highly productive agriculture, including crop yield predictions. AGRIST's dream is to contribute to solving global food challenges beyond Japan, thereby enhancing the happiness and well-being of all humanity. AI might be able to make this big dream come true.

Automatic harvesting robot developed by AGRIST.

### Panasonic Connect

Panasonic Connect, a subsidiary of the major Japanese electronics manufacturer Panasonic Group, introduced its own version of an AI assistant, "Connect AI," to all domestic employees, totaling 12,500 people in February 2023. It is intended for daily use for tasks such as email composition, information gathering, and computer code creation. Connect AI was built using the Microsoft Azure OpenAI Service, which enables enterprise-level data and privacy protection. Once adopted, the number of queries received by the AI assistant jumped from 2,000 per day to more than 5,000 per day within a few months. A senior representative at Panasonic Connect stated: "We believe that all our business professionals should use AI as part of their daily routine. Therefore, we considered not whether to use AI but when to start using it." Japan has the world's most rapidly aging population which looms over the prospect of sustainable economic growth. Panasonic Connect believes generative AI is one way to enhance employee productivity by enabling employees to focus on creative tasks that only humans can do.

Panasonic Connect CEO, Yasuyuki Higuchi, created a welcome speech for new employees using ConnectAI.

### PKSHA

Japanese startup PKSHA Technology's subsidiary, PKSHA Workplace, aims to realize the future of work by leveraging the power of generative AI. It released "AI Helpdesk for Microsoft Teams" in 2022 as a system that handles internal inquiries, provides automatic responses, and otherwise facilitates collaboration among team members. In the latest integration, PKSHA Workplace introduced a new feature to allow for the automatic generation of FAQs directly from AI Helpdesk conversation logs, using a large-scale language model via the Azure OpenAI Service. This feature enables cost reduction in knowledge sharing and creation, as well as an improvement in automatic response rates and self-resolution rates.

### JERA

JERA, a company focused on transforming the energy industry through digital innovation and addressing global energy challenges, has embarked on a journey to enhance the operational efficiency of power plants and reduce environmental impact. Microsoft has served as a key partner to JERA on this journey. Using Microsoft Azure Digital Twins and generative AI, JERA and Microsoft are co-developing new Operations and Maintenance (O&M) solutions, aimed at improving the performance of power plants and reducing downtime, thereby increasing operational efficiency. This project involves utilizing cloud and AI to remotely analyze real-time operational data from power plants, allowing for automated anticipation and detection of problems, quicker decision-making, and more efficient plant management. This will dramatically

transform traditional plan operation by freeing those involved in plant operations from time consuming data collection and analysis work, allowing for more focus on advanced plant operations and value creation. Through this partnership with Microsoft, JERA expects to reduce its CO2 emissions and lower its environmental footprint, contributing to the realization of JERA's Zero Emission 2050 goal.

## AI for education and empowerment

### Okinawa University

Since the enactment of the "Act on Promotion of Elimination of Discrimination against Persons with Disabilities" in 2016, the number of students with disabilities in higher education has been steadily increasing. A significant challenge in this context is providing support for students with hearing impairment, as lectures at universities often rely more on oral explanations with fewer written notes on the blackboard. Moreover, seminars involve discussions and presentations primarily conducted orally.

Okinawa University has been proactive in promoting the acceptance of students with disabilities, based on its founding principles: "equal educational opportunities". Previously, the university relied on volunteer students for notetaking, which required coordination with assisting students and resulted in variable quality in transcriptions. Now Okinawa University utilizes the "AI mimi" service provided by SeiCommu Co., Ltd. to support students with hearing impairments during lectures. "AI mimi" combines speech recognition using Azure Cognitive Services with correction work performed by professional human operators, creating a hybrid transcription service. This service has also been adopted by Microsoft's "AI for Accessibility" program. It achieves extremely high accuracy in converting speech into text with a time lag of just 1 to 2 seconds. Improved information accessibility is one important step forward to equal educational opportunities and empowerment for everyone in Japan.


Text correction and subtitle delivery by operators. Collaboration between AI and humans.

### Microsoft Base Ritsumeikan

In August 2023, Ritsumeikan University unveiled its plans to establish a next-generation research base, Microsoft Base Ritsumeikan, a pioneering initiative supported by Microsoft. Ritsumeikan University aims to align Japan's education system with the rapid technological advancements prevalent in society today and to better harness cutting-edge technologies like generative AI and the metaverse for innovative learning. In particular, Ritsumeikan plans to develop its proprietary generative AI model, using Microsoft Azure to manage and utilize its internal database for both operational and student-related purposes. Scheduled to open in April 2024, Microsoft Base Ritsumeikan will serve as a communal base for Ritsumeikan students and faculty members as well as those from other educational institutions, municipalities, and businesses, all coming together to collaboratively address the emerging societal challenges of the new era.

### Bengo4.com

In Japan, only about 20% of people involved in legal processes are able to reach a lawyer, and this is referred to as the "20% Judiciary." Bengo4.com is a startup founded by the lawyers in 2005. Building upon an existing service called "Everyone's Legal Consultation," where everyone can consult with a lawyer for free, Bengo4.com created "Chat Legal Consultation" in May 2023, an AI legal consultation chat service that utilizes questions and answers extracted from over 1.25 million consultations submitted to "Everyone's Legal Consultation" since 2007, while taking into account the requirements under the Attorney Act of Japan. This is powered by Microsoft Azure. This service will allow anyone to easily and freely seek legal advice 24/7, enabling a future where legal consultation resources are more accessible for everyone.

## AI for the future of public services

### Tokyo Metropolitan Government

Tokyo Metropolitan Government became a trailblazer in proactively adopting and leveraging among the most transformative technologies of our time—ChatGPT. In August 2023, the Tokyo Metropolitan Government started utilizing AI for its operations. Approximately 50,000 employees are using ChatGPT for tasks such as document summarization and as a reference in policy planning. Strong data and privacy protection offered by Microsoft Azure made this important decision possible. To support a high utilization rate, the government compiled guidelines based on the deliberations of a project team established back in April. Guidelines provide practical methods to leverage ChatGPT, something similar to prompt engineering, and specific examples.

They hope other local governments that are considering adopting text generation AI will also make use of these guidelines. Tokyo Metropolitan Government's decision is best thought of not as one specific use case, but as one big step forward for the future of public service throughout Japan.

### Osaka Prefecture

The Osaka Prefectural Government has forged a collaborative partnership with Microsoft to introduce AI into its operations, with the primary goal of enhancing residents' quality of life. In particular, Osaka Prefecture intends to leverage generative AI for smart city-related initiatives and incorporate AI into its human resource development efforts. And in order to further contribute to a resolution of societal challenges in Osaka, Microsoft will provide support for the development of a communication assistance service to support senior citizens, combat senior isolation, and encourage outdoor activities. "Speak with Dai-chan" features a dog character named "Dai-chan" with Kansai dialect text and voice, which enables conversations and provides event information. The service marks the first instance of a municipality employing generative AI to help address senior isolation, a key issue that Japanese society is grappling with due to its aging population.

### Hamamatsu City

Hamamatsu City, renowned for its diverse population with a substantial number of foreign residents, has taken proactive measures to cater to its multicultural community. One notable initiative involved the publication of English and Portuguese editions of its local newspaper, Hamamatsu PR, alongside the original Japanese edition. However, the increasing diversity of the population has presented continuous challenges in delivering adequate information due to language barriers.

To further address these challenges, the city government adopted PR Plus, a cloud-based public relations and public information support service by VOTE FOR, built on Microsoft Azure. PR Plus greatly extended the reach of Hamamatsu City's public communications by offering translation and text-to-speech capabilities, covering various languages. Using this service, the city was able to facilitate information access for foreign residents regardless of their language proficiency. It also helped to improve accessibility for the visually impaired. The transition to PR Plus has resulted in a significant increase in readership, with approximately 10,000 readers consistently accessing digital articles. Furthermore, AI-powered analytics tools are being considered for future use to better understand how residents engage with digital content. By understanding user behaviour, PR Plus and Hamamatsu City anticipate that AI can help optimize the timing and presentation of messages to make public relations even more effective.

# Bibliography

## Foreword

## Part 1 Governing AI in Japan

### Cited Sources


Japan's population drop since 2015
https://asia.nikkei.com/Economy/Japan-s-population-drops-to-126m-in-2020-census-down-0.7-vs.-2015
G7 Hiroshima Summit 2023
https://www.g7hiroshima.go.jp/en/
Hiroshima AI Process
https://www.bloomberg.com/news/articles/2023-05-20/g-7-leaders-agree-to-set-up-hiroshima-process-to-govern-ai
Tentative Summary of AI Issues
https://www8.cao.go.jp/cstp/ai/ronten_youshi_yaku.pdf
LDP AI Whitepaper
https://www.taira-m.jp/ldp%E2%80%99s ai whitepaper_etrans_2304.pdf
Economic Security Promotion Act
https://www.cfr.org/sites/default/files/pdf/economic security promotion act %28summary%29%28English%29.pdf
AI Governance in Japan v1.1 from Expert Group
https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20210709_8.pdf


### General sources


Governing AI: A blueprint for the future
https://aka.ms/GoverningAI-blueprint
Japan ESPA cutting edge technology development
https://japannews.yomiuri.co.jp/politics/politics-government/20220513-27693/
Japan Expert Group on Architecture for AI Principles to be Practiced
https://www.meti.go.jp/english/press/2022/0128_003.html


### References


Fortune business insights AI market report
https://www.fortunebusinessinsights.com/industry-reports/artificial-intelligence-market-100114


## Part 2 Responsible by Design: Microsoft's Approach to Building AI Systems that Benefit Society

### Cited Sources


NIST AI Risk Management Framework
https://www.nist.gov/itl/ai-risk-management-framework
Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies
https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/
OECD AI Principles
https://oecd.ai/en/ai-principles
OECD Framework for the Classification of AI systems
https://www.oecd.org/publications/oecd-framework-for-the-classification-of-ai-systems-cb6d9eca-en.htm
UNESCO Recommendation on the Ethics of Artificial Intelligence
https://www.unesco.org/en/artificial-intelligence/recommendation-ethics


### General sources


Governing AI: a blueprint for the future
https://aka.ms/GoverningAI-blueprint


### References


Critical Infrastructure Sectors | CISA
https://www.cisa.gov/topics/critical-infrastructure-security-and-resilience/critical-infrastructure-sectors

# Part 3 AI in Action in Japan

## General sources

FRONTEO for aging
https://www.fronteo.com/en/20201021
Japan Agency for Medical Research and Development
https://news.microsoft.com/ja-jp/2023/04/03/230403-whats-next-for-healthcare-the-latest-trends-in-dx-support-for-the-healthcare-industry/
AGRIST
https://www.microsoft.com/ja-jp/industry/blog/retail/2022/02/10/microsoft-for-startups-agrist/)
Panasonic Connect
https://news.microsoft.com/apac/features/not-if-but-when-why-japans-panasonic-connect-is-going-all-in-on-ai/
PKSHA
https://prtimes.jp/main/html/rd/p/000000088.000022705.html
JERA for energy
https://news.microsoft.com/ja-jp/2023/09/11/230911-jera-and-microsoft-partner-to-create-more-efficient-and-sustainable-power-plant-operations-jp/
Okinawa University
https://customers.microsoft.com/ja-jp/story/1426663873879628640-okinawa-higher-education-azure-jp-japan
Microsoft Base Ritsumeikan
https://news.microsoft.com/ja-jp/2023/08/21/230821-accelerating-learning-challenges/
Bengo4.com
https://www.bengo4.com/corporate/news/article/zztucptlqh6
Tokyo Metropolitan city
https://www.nikkei.com/article/DGXZQOCC156IM0V10C23A8000000/
Osaka smart city initiatives
https://news.microsoft.com/ja-jp/2023/09/08/230908-signed-a-business-collaboration-agreement-with-osaka-prefecture-to-promote-the-utilization-of-ai/
Hamamatsu city initiatives
https://customers.microsoft.com/ja-jp/story/1468104978479874435-hamamatsu-city-government-azure-ja-japan

## References

Our commitment to the UN Sustainable Development Goals - Microsoft
https://www.microsoft.com/en-us/corporate-responsibility/un-sustainable-development-goals