Microsoft logo white text version

Zzz

Une image contenant bleu, brouillard, capture d’écran, bleu vert

Description générée automatiquement

Responsible AI Workshop

Establishing your own Responsible AI journey for your (non-Generative vs. Generative) AI-powered solutions

A starter guide for data engineers, data scientists, ML developers, ML engineers, and other AI practitioners to help putting Responsible AI into practice

Version 1.1 – August 2022 (Updated: June 2024)

This page is intentionally left blank.

Table of contents

[Notice 4](#_Toc170393137)

[About this guide and the learning objectives 5](#_Toc170393138)

[Objectives of this guide 5](#_Toc170393139)

[Non-objectives of this guide 6](#_Toc170393140)

[Guide elements 6](#_Toc170393141)

[Guide prerequisites 6](#_Toc170393142)

[Module 1: Building Responsible AI solutions 7](#_Toc170393143)

[Understanding Microsoft’s Responsible AI journey 9](#_Toc170393144)

[Establishing a multi-stakeholder approach and a global dialogue 19](#_Toc170393145)

[Releasing responsible AI tools and practices 21](#_Toc170393146)

[Going forward 24](#_Toc170393147)

[Module 2: What about a Responsible AI Lifecycle? 26](#_Toc170393148)

[Assessing and preparing an AI system 27](#_Toc170393149)

[Designing, building, and documenting an AI system 29](#_Toc170393150)

[Validating and supporting an AI system 29](#_Toc170393151)

[Leveraging engineering systems and tools, patterns & practices 29](#_Toc170393152)

[Module 3: Initiating/Improving your own Responsible AI journey 37](#_Toc170393153)

[Understanding Microsoft’s AI commitments 37](#_Toc170393154)

[Defining your own AI principles 37](#_Toc170393155)

[Assessing your Responsible AI maturity 39](#_Toc170393156)

[Paving the road of your own journey 40](#_Toc170393157)

[As a conclusion 42](#_Toc170393158)

[To go beyond 43](#_Toc170393159)

# Notice

Information in this document, including URL and other Internet Web site references, is subject to change without notice. Unless otherwise noted, the example companies, organizations, products, domain names, e-mail addresses, logos, people, places, and events depicted herein are fictitious, and no association with any real company, organization, product, domain name, e-mail address, logo, person, place or event is intended or should be inferred. Complying with all applicable copyright laws is the responsibility of the user.

A drawing of a face

Description automatically generated

Attribution 4.0 International (CC BY 4.0)

Microsoft and any contributors grant you a license to this document under the [Creative Commons Attribution 4.0 International Public License](https://creativecommons.org/licenses/by/4.0/legalcode), see the [LICENSE](https://github.com/microsoft/responsible-ai-workshop/blob/main/LICENSE) file, and grant you a license to any code in the repository under the [MIT License](https://opensource.org/licenses/MIT), see the [LICENSE-CODE](https://github.com/microsoft/responsible-ai-workshop/blob/main/LICENSE-CODE) file.

Microsoft, Windows, Microsoft Azure and/or other Microsoft products and services referenced in the document may be either trademarks or registered trademarks of Microsoft in the United States and/or other countries. The license for this document does not grant you rights to use any Microsoft names, logos, or trademarks. Microsoft's general trademark guidelines can be found at <http://go.microsoft.com/fwlink/?LinkID=254653>.

Privacy information can be found at <https://privacy.microsoft.com/en-us/>

Microsoft and any contributors reserve all other rights, whether under their respective copyrights, patents, or trademarks, whether by implication, estoppel or otherwise.

# About this guide and the learning objectives

Welcome to this guide Establishing your own Responsible AI journey for your (non-Generative vs. Generative) AI-powered solutions.

The tech industry is being called upon to develop and deploy Artificial Intelligence (AI) technologies and Machine Learning (ML)-powered systems (products, applications, or services) and/or features more responsibly – we will further refer to these as (non-Generative vs. Generative) AI systems -. Yet many organizations implementing such AI systems report being unprepared to address AI risks and failures.

To meet these challenges, Microsoft is striving to adopt a human-centered approach to AI, designing and building technologies that benefit people and society while also mitigating potential harms. This includes understanding human needs and using these insights to drive development decisions from beginning to end.

This guide consists of a series of modules and tutorials for data engineers, data scientists, AI developers and other AI practitioners, as well as potentially anyone interested considering the wide range of socio-technical aspects involved in the subject.

## Objectives of this guide

As its name indicates, this guide is part of the Responsible AI Workshop. It more particularly introduces the RAI considerations throughout all phases of the assessment, development, and deployment processes, helping the above-mentioned roles to build responsible AI-powered solutions.

As such, it is more particularly intended to:

1. Share and comment from the intended audience's perspective our [ongoing journey towards Responsible AI (RAI)](https://aka.ms/RAI), starting from core principles that constitute our “North Star”, and from which this effort is deeply anchored to the practices we adopt, enforce, and evolve company-wide in terms of end-to-end lifecycle for the design, the development, the deployment, and the monitoring of these AI systems,
2. Ultimately give you insights to establish your own journey towards Responsible AI.

By the end of the guide, you will be able, as part of the workshop and in terms of learning objective to more specifically have an overview of the why, the what and the how regarding the adoption of a Responsible AI strategy and approach for your (non-Generative vs. Generative) AI-powered solutions, illustrated through the lenses of Microsoft’s journey, from Microsoft’s core Responsible AI principles to the way these principles translate into a framework of requirements, guidance, and governance through the Responsible AI standard and related practices.

## Non-objectives of this guide

As such, this guide isn’t aimed at giving a complete overview of RAI tooling aimed at helping to protect and control non-Generative vs. Generative AI systems.

For tutorials of the most prominent tools we open-sourced for non-Generative AI, please refer to the guide [Responsible AI Workshop - Leveraging Responsible AI Tooling for your non-Generative AI-powered solutions](https://github.com/microsoft/responsible-ai-workshop/blob/main/nongen-ai-tooling-tutorials/docs/leveraging-responsible-ai-tooling.docx).

For guidance and tutorials for your Generative AI-powered solutions, please refer to the guide [Responsible AI Workshop - Building and using Generative AI responsibly with Azure and beyond](https://github.com/microsoft/responsible-ai-workshop/blob/main/nongen-ai-tooling-tutorials/docs/leveraging-responsible-ai-tooling.docx).

These two guides are also part of this Responsible AI Workshop, which is available on GitHub at <https://github.com/microsoft/responsible-ai-workshop>.

**Note** For a complete overview of Microsoft’s resources designed to help you responsibly implement (non-Generative vs. Generative) AI systems, please refer to the [Microsoft Responsible AI resources page](https://aka.ms/rairesources).

## Guide elements

The document is organized as follows.

Module 1 introduces what Responsible AI means. One should note that the definition of Responsible AI varies from organization to organization. Here we will focus on [Microsoft's definition of Responsible AI](https://www.microsoft.com/en-us/ai/responsible-ai-resources), and the related approach, and provide you for that purpose with an overview of Microsoft’s Responsible AI journey starting with our core RAI principles, the standards in place, the related activities and practices in terms of AI systems development lifecycle, along with the tooling to sustain the implementation of the requirements and to help fulfill the various RAI objectives.

Module 2 illustrates how to put int action such learnings and the provided purpose-built tooling for non-Generative AI systems throughout their end-to-end lifecycle sustained by [MLOps practices](https://learn.microsoft.com/en-us/azure/machine-learning/concept-model-management-and-deployment?view=azureml-api-2).

With this “foundation”, Module 3 more specifically discusses your own Responsible AI journey.

## Guide prerequisites

No specific prerequisites are required to get the most out of this guide.

# Module 1: Building Responsible AI solutions

Responsible innovation is top of mind. Advancements in AI are indeed different than other technologies because of the pace of innovation – there has been hundreds of research papers published every year in the past few years -, but also because of its proximity to human intelligence. These advancements are:

* Impacting us at a personal and societal level. We refer to this as the sociotechnical impact of AI, which has given rise to an industry-wide debate about how the world should/shouldn’t use these new capabilities. It is not because you can do something that you should necessarily do it.
* And also changing the way companies approach AI. The topic of Responsible AI is increasingly becoming an important theme as more companies struggle with challenges in terms of governance, security and compliance – you have hereafter some considerations and related percentages capturing interest of companies when investing in AI and ML technologies from a study conducted by the Capgemini Research Institute back in 2019 (See [Organizations must address ethics in AI to gain public’s trust and loyalty](https://www.capgemini.com/news/ethics-in-ai/)).

Une image contenant texte

Description générée automatiquement

Figure 1. Most important considerations when investing in AI and ML technologies

The ability of AI systems to ensure data security and privacy, and the level of transparency of how systems work and are trained, are now the two most important requirements when investing in ML technologies. It’s a marked shift from previous years, when scalability, performance, and ease-of-use topped investment priorities, in that order.

At the same time, studies have shown that explainability of algorithms are one of the key factors holding businesses back from implementing AI-powered solutions. Trust in an AI system is tied to its transparency, and without trust, people do not want to adopt AI or any other technology for that matter.

There are hundreds of ML algorithms and related techniques that data scientists can already access, and new ones emerge every day through various research papers, and all the AI innovation resulting from that is pushing the boundaries of science and technology – for example, the [AI at Scale](https://www.microsoft.com/en-US/ai/ai-at-scale), [on Microsoft Research (MSR)](https://www.microsoft.com/en-us/research/project/ai-at-scale/) constitutes a new approach to AI that is fueling the next generation of AI innovation at scale with massive AI models trained with innovative tools and advanced infrastructure -.

If these AI systems are opaque and unable to explain how or why certain results are presented, this lack of transparency will undermine trust in the system and in any results they produce. The stance is clear: people don’t use what they don’t trust.

While it is true that sometimes the current state of the art of some technologies does not allow to achieve an acceptable level of transparency today, e.g., an intimate understanding of how a Deep Learning model comes up with a specific prediction, it doesn’t mean we will remain in this situation forever; breakthroughs in innovation keep reminding us of how they possibly enable an alternative perspective and a new horizon for the future in various dimensions, and in the meantime we can already approximate the global vs. local behavior of such a models.

“The more powerful the tool, the greater the benefit or damage it can cause… Technology innovation is not going to slow down.  The work to manage it needs to speed up.”

*- Brad Smith, Vice Chair and President, Microsoft Corporation*

At the same time, one should also acknowledge that responsible innovation and consequently Responsible AI are an important yet underrated subject among data engineers, data scientists, AI developers, and other AI practitioners in general. This is a current reality that undermines our collective efforts to gain control over AI systems and tackle all the concerns they induce, ranging from privacy concerns and information transparency to raising questions of who should be held accountable when AI systems behave unexpectedly or induce undesired side effects.

Addressing these ethical, transparency, and accountability concerns is no easy feat and requires coordinated multi-disciplinary efforts between not only data scientists, technical experts, but also ethics experts, law makers, etc. These efforts range from defining AI principles and outlining requirements to follow and goals to pursue in enacted policies and standards, to defining practices for solving some of these issues along with adapted guidelines and tooling.

This must then translate into the development lifecycle of these AI systems to ensure that *what we (decide to) build* benefits people and society, and that *how we build* it begins and ends with people in mind.

Microsoft is heavily committed in developing and deploying AI systems that are ethical, safe, secure, and trustworthy. We know that this will lead to more trustworthy AI systems that benefit not only our customers, but the whole of society, allowing to unlock the AI’s positive impact across the globe.

Here are a few examples of where we’re seeing a positive impact with AI.

Une image contenant texte, habits, homme, capture d’écran

Description générée automatiquement

Importantly people must be able to trust the technologies they use and the organizations behind them.

Responsible AI (RAI) is an holistic approach to assessing, developing, and deploying AI systems guided by ethical values and enduring principles. RAI is Human-Centered. See [The A to Z of responsible AI](https://www.linkedin.com/pulse/z-responsible-ai-microsoft-on-the-issues/?trackingId=7KzdsSvIsfL536ulDrYPpA%3D%3D&culture=en-us&country=us).

So, with all of that in mind, let’s start by our own journey in this space to illustrate this before dedicating some time to explore the already available Responsible AI resources (that continuously evolve).

## [The book Tools and Weapons by Brad Smith and Carol Ann Browne.](https://news.microsoft.com/on-the-issues/tools-and-weapons/)Understanding Microsoft’s Responsible AI journey

“When your technology changes the world, you bear a responsibility to help address the world you have helped create.”

*- Brad Smith, Vice Chair and President, Microsoft Corporation*

[*Tools and Weapons: The Promise and the Peril of the Digital Age*](https://news.microsoft.com/on-the-issues/tools-and-weapons/)

As technology like AI becomes more embedded in our daily lives, we must build it responsibly - from design to deployment. The question isn't what technology can do, but what it should do.

The goal of this section is threefold:

1. Share our principles and how they translate into practices company-wide for our services and products.
2. Create awareness of the governance framework we have in place and the way we establish and evolve our standard to translate our responsibility in design, development, and deployment of (non-Generative vs. Generative) AI systems and models.
3. Further present and discuss related practices and the already available tooling to sustain the implementation of the requirements outlined in the standard, and the related processes and activities as part of our ongoing Responsible AI journey.

Our journey towards Responsible AI begins in 2016, with Satya Nadella penning an article in the Slate magazine titled [The partnership of the future](https://slate.com/technology/2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societys-challenges.html) where our CEO explores how humans and AI can work together to solve society’s greatest challenges. This article introduced concepts of transparency, efficiency but not at the expense of the dignity of people, intelligent privacy, algorithmic accountability, and protection against bias.

Une image contenant texte, capture d’écran, diagramme, Police

Description générée automatiquement

Figure 2. Microsoft’s Responsible AI Journey

At the Microsoft Build conference in May 2017, Satya built on this with references to George Orwell and Aldus Huxley, and shortly after that, in July 2017, Microsoft formed our AI Ethics committee – AETHER for *AI, Ethics, and Effects in Engineering and Research*), an executive committee only, formed as a cross-company advisory group on AI ethics and effects in engineering and research for the Microsoft Senior Leadership Teams (SLT).

**"*AETHER will ensure our AI platform and experience efforts are deeply grounded within Microsoft’s core values and principles and benefit the broader society. Among other steps, we are investing in strategies and tools for detecting and addressing bias in AI systems and implementing new requirements established by the GDPR. While there is great opportunity, ensuring we always act responsibly for our customers and partners will continue to be a hallmark of our work.*"**

***- Satya Nadella,*** *Microsoft CEO and Chairman****, Microsoft Corporation***

AETHER spent time listening to our customers and internal experts, and then partnered with Legal Affairs to publish in January 2018 the eBook [The Future Computed](https://news.microsoft.com/futurecomputed/), which articulate [six core principles](https://blogs.microsoft.com/on-the-issues/2018/12/17/six-principles-to-guide-microsofts-facial-recognition-work/) that should guide our work and investments around AI.

As outlined above, Microsoft [calls for Facial Recognition technology regulation](https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/) in July 2018, and published later by the end of the year in December [Facial Recognition principles](https://blogs.microsoft.com/on-the-issues/2018/12/06/facial-recognition-its-time-for-action/).

Microsoft formed the Office of Responsible AI (ORA) in early 2019 and in the fall of this year, we published internally the first version of our Microsoft Responsible AI Standard (RAIS), i.e., a set of rules for how we enact our responsible AI principles underpinned by Microsoft’s corporate policy. We publicly released the second version RAIS in June 2022.

We have moved from initial ideas and research to a point where we are now operationalizing responsible AI at scale, which is now at the heart of implementing our solutions involving (non-Generative vs. Generative) AI use. While we are proud of the steps we have taken, we expect to take many more in the future as we learn from each other on our path to responsible AI, see section Going forward below.

In the past year, Microsoft has made numerous announcements, from launching new responsible AI tools to making new external commitments on how we implement or will implement responsible AI, to launching new organizations to advance the safety and security of AI worldwide. Some of these are listed in the above Figure. For more information, see [Meeting the AI moment: advancing the future through responsible AI](https://blogs.microsoft.com/on-the-issues/2023/02/02/responsible-ai-chatgpt-artificial-intelligence/).

With that timeline in mind, lets now further consider the abovementioned core principles that pave our journey.

### Introducing Microsoft's AI principles

AI can be a powerful tool with the potential to transform how we work and learn—when it’s designed with people in mind.

To this end, as Microsoft, we believe that the development and deployment of AI must be guided by the creation of an ethical framework. We set out our view back in 2018 in the eBook The Future Computed with six core principles.

These [Microsoft’s AI principles](https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6) are the foundation for a responsible and trustworthy approach to AI at Microsoft. They act as a mental tool or framework in which to organize thinking about ethics at Microsoft.

Une image contenant texte, capture d’écran, Système d’exploitation, nombre

Description générée automatiquement

Figure 3. Microsoft Responsible AI Principles

They indeed call out the aspirations of designing our systems in accordance with goals of fairness, reliability and safety, privacy and security, inclusiveness, transparency and accountability. A brief explanation of each of these principles is given below for your convenience:

1. Fairness. AI systems must be developed to treat everyone fairly and avoid affecting similarly situated groups of people in different ways. This principle acknowledges that defining and mitigating fairness issues for an AI system depends on understanding the system’s purpose and context of use, and that a system’s fairness reflects decision-making during both development and deployment.

For example, when AI systems provide guidance on medical treatment, loan applications, or employment, they should make the same recommendations to everyone with similar symptoms, financial circumstances, or professional qualifications, regardless of other sensitive features like sex and race for example.

1. Reliability & Safety. To build trust, it’s also important that AI systems operate reliably, safely, and consistently under normal circumstances and in unexpected conditions. How they behave and the variety of conditions they can handle reliably and safely largely reflects the range of situations and circumstances that developers can anticipate during design, build, prototyping, and testing.

This principle encompasses consideration of the harms that might come from a technology, and ways employees can strive to minimize those risks, so technologies can give the greatest benefits to their users.

1. Privacy and Security. It’s also crucial to develop AI systems that can protect private information and resist attacks. As AI becomes more prevalent, protecting privacy and securing important personal and business information is becoming more critical and complex. Privacy and data security issues require special close attention for AI because access to data is essential for AI systems to make accurate and informed predictions and decisions about people. Data must be secure at all stages, and to this end, actions must be taken to institutionalize privacy and security.
2. Inclusiveness. For the 1 billion people with disabilities around the world, AI technologies can be a game-changer. AI can improve access to education, government services, employment, information, and a wide range of other opportunities.

Inclusive design practices can help system developers understand and address potential barriers in a product environment that could unintentionally exclude people. By addressing these barriers, we create opportunities to innovate and design better experiences that benefit everyone.

1. Transparency. When AI systems are used to help inform decisions that have tremendous impacts on people’s lives, it’s critical that people understand how those decisions were made. A crucial part of transparency is what we refer to as intelligibility or the useful explanation of the behavior of AI systems and their components.

Improving intelligibility requires that stakeholders comprehend how and why they function so that they can identify potential performance issues, safety and privacy concerns, biases, exclusionary practices, or unintended outcomes. We also believe that those who use AI systems should be honest and forthcoming about when, why, and how they choose to deploy them.

1. Accountability. We believe the people who design and deploy AI systems must be held accountable for how their systems operate and for the impact of that operation on society. This includes considering the structures that can be implemented to ensure accountability at multiple levels, including design, development, sales, marketing, and use, as well as advocacy for the regulation of technologies when warranted.

These six principles are our guiding star, meaning they articulate the values we must uphold when developing or deploying AI systems. However, we recognize that principles alone are not sufficient.

They do not apply themselves. Just because you wish and believe an AI system should be fair or inclusive does not make it so. Principles are also open to interpretation. Principles also do not answer the question of how.

The hard and essential work begins when you endeavor to turn those principles to work, i.e., into practices. Merely having principles indeed does not imply a change in a company’s culture unless those principles are made concrete through standards, practices and tools that help full time employees (FTEs) work through how to think responsibly when designing, developing, deploying, and monitoring AI systems.

### Translating AI principles into practice

**"*We have made huge investments in AI because we’re optimistic about what it can do to help people, industry and society, and because we’re committed to bringing technology and people together to realize the promises of AI responsibly.*"**

***-*** [*What is Microsoft's Approach to AI?*](https://news.microsoft.com/source/features/ai/microsoft-approach-to-ai/)

Our mission is to empower every person and every organization on the planet to achieve more, and this mission statement translates into our approach to AI. We are optimistic about the potential of AI and, in this era of AI, we remain committed to responsible AI, innovation, and a responsible future.

So, to advance these principles and make sure they are enforced and implemented into the company’s workflows, Microsoft developed several responsible AI practices and tools for incorporating applied ethics in technology. They all serve an ethical end; some are more procedural and are explored here, while others are more technical in nature and are explored in the other guides and tutorials of this Responsible AI Workshop.

At Microsoft, we are moving from principles to practice since 2019 with the objectives to also empower our customers to do the same, as we did before for example with our [Microsoft Security Development Lifecyle (SDL)](https://www.microsoft.com/en-us/securityengineering/sdl/).

**We believe that we need to work towards ensuring AI systems are responsible by design.**

We (continue to) [establish building blocks to enact our principles](https://blogs.microsoft.com/on-the-issues/2021/01/19/microsoft-responsible-ai-program/). To this end, a multifaceted approach is necessary to operationalize the abovementioned principles on a large scale. As a basis of our Responsible AI program at Microsoft, such an approach focuses on four key areas:

1. A governance structure to make progress and be accountable at the highest level of the organization.
2. Rules that governthe design, development, deployment, and use of AI technologies and systems. They frames our internal policies and, in this respect, our corporate standard for responsible AI, i.e., the so-called Microsoft Responsible AI standard, in terms of goals, requirements, and practices as well as goals as an opportunity to foster innovation and build better AI systems.
3. Training, tools and testing to support [secure DevOps](https://www.microsoft.com/en-us/securityengineering/sdl/) teams, data scientists and all other employees contributing to AI systems and models promoting a human-centered mindset, as well as providing appropriate measurement and testing capabilities.
4. And oversight for these AI systems and models, a (central) inventory and a set of monitoring, tracking, reporting, auditing capabilities to structure this effort and further demonstrate compliance with established processes, and controls in place.

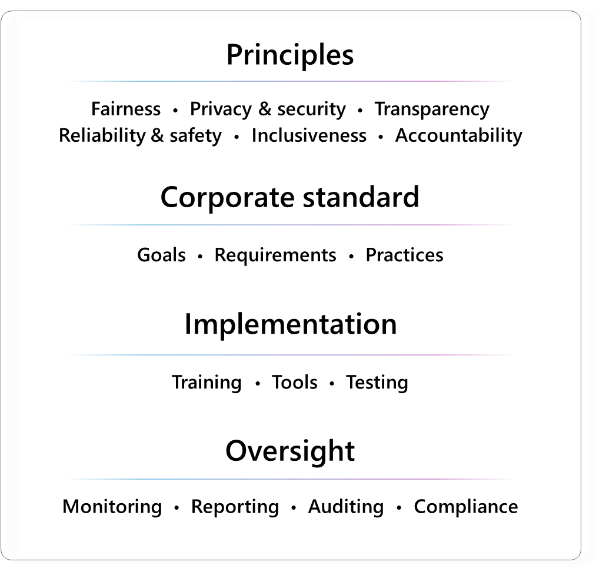


Figure 4. Microsoft’s Building blocks to enact principles and put them into practice

Critical to the Responsible AI discussion is the need for a governance framework that really starts from the very beginning in the business case and continues throughout the design, development, and ongoing management of AI. If there is much that is new and unexplored in the practice of responsible AI, there is also much to learn from adjacent fields, and this is particularly true when it comes to making governance a solid foundation.

Our governing practices help to ensure and foster Responsible AI both within our company and beyond.

Our Responsible AI [governance approach](https://www.microsoft.com/en-us/ai/our-approach?activetab=pivot1:primaryr5) borrows from the “hub-and-spokes” model that has worked successfully to integrate security, privacy and accessibility into our products and service.

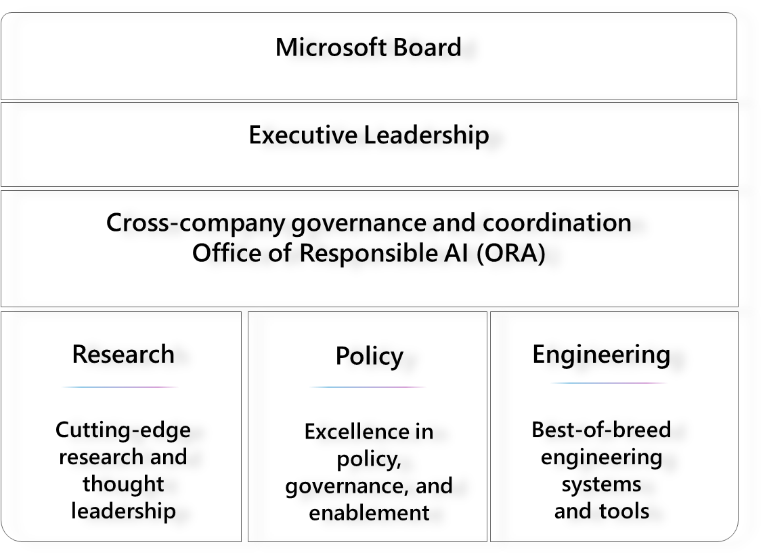


Figure 5. Microsoft’s Responsible AI governance framework following the “hub and spokes” model

Our current model relies upon centralized and decentralized functions to put our Responsible AI principles into practice. This “hub-and-spokes” model provides the accountability and authority to drive initiatives while also enabling Responsible AI policies to be implemented at scale.

Our “hub” includes:

1. The AI Ethics in Engineering and Research (AETHER) committee, whose working groups leverage top scientific and engineering talent to provide subject-matter expertise on the state-of-the-art and emerging trends regarding the enactment of Microsoft’s responsible AI principles.
2. The Office of Responsible AI (ORA), which defines, enables, governs, and coordinates the company’s approach to responsible AI. In other words, it sets our policies and governance processes, and as such, it enacts the already mentioned Responsible AI Standard.
3. And the Responsible AI Strategy in Engineering (RAISE) group, which enables our engineering groups to implement our responsible AI processes through systems and tools. This is both an engineering team and a strategic engineering initiative to enable the implementation of the above Standard across Microsoft’s engineering groups: Cloud and AI, Experiences and Devices, Technology and Research, Gaming, and LinkedIn.

The three groups work together to set a consistent bar for Responsible AI across the company and they empower our “spokes” to drive initiatives and be accountable for them.

People remain at the center of our responsible AI progress. The spokes of our governance include our Responsible AI Champs community. The Champs are appointed by company leadership and sit in engineering and sales teams across the company. They raise awareness about Microsoft’s approach to responsible AI and the tools and processes available, they spot issues and help teams assess ethical and societal considerations, and they cultivate a culture of responsible innovation in their teams. As of this writing, over 400 people at Microsoft are currently working on responsible AI, and more than half focus on responsible AI full-time, helping us implement practices at scale to create and operate trustworthy, safe and transparent AI systems designed to benefit society as a whole.

A governance system should be agile to the changing nature of technology and the business. In that sense, our governance system continues to evolve to this day, as it should. As illustrated above, at the center of our governance framework “hub” lays ORA for cross-company governance and coordination.

Microsoft's leadership in responsible development and implementation of AI technologies is based on an ecosystem of teams and functions within the company. Our ecosystem includes several levels, including the Microsoft Board of Directors, senior management, including sponsorship of this work by Brad Smith, President, and Kevin Scott, Microsoft's Chief Technology Officer.

### Introducing the Microsoft Responsible AI Standard

Like other internal policies and standards previously rolled-out and in application within the company for security, privacy, and accessibility, and by using our AI principles as a North star, we've created the [Microsoft Responsible AI Standard (RAIS)](https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf).

The RAIS is our internal policy for developing innovative (non-Generative vs. Generative) AI systems aligned with our company’s core values, and thus assisting teams in implementing and deploying AI technologies in a responsible manner. As such:

* It records our practice of responsible AI by design, i.e., the proactive ways in which we guide the design, build, and testing of AI systems.
* It establishes a durable framework for the maturing practice of responsible AI and evolving regulatory requirements.
* It reflects our deeper explorationof what our six principles mean and the steps that must be taken to uphold them.

Defining the standard took us several years of effort and it was developed before the adoption of new detailed regulations on the subject, such as the [EU AI Act](https://data.consilium.europa.eu/doc/document/PE-24-2024-INIT/en/pdf), and while international standards in this field, like [ISO/IEC 42001](https://www.iso.org/standard/81230.html), were still under development.

The first version was published in the fall of 2019, with the understanding that we were just at the beginning of systematically moving from principles to practices. In the fall of 2021, we published a second version, which is a fundamental rewrite and reflects a better grasp of our principles.

The main lessons from the first version of the standard that were taken into account in the second version included feedback that engineering teams appreciated examples and struggled with open considerations. They asked for more concrete requirements along with a closer integration with engineering practices. It is on the basis of these considerations, and many others, that the second version of the standard was developed.

To learn more about the development of this standard, watch the video [Developing Microsoft's Responsible AI Standard](https://www.youtube.com/watch?v=lkIlsgrIMtU).

In June 2022, as already outlined, we made publicly available this second version of the [Microsoft’s Responsible AI Standard](https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf), i.e., a framework to guide how we build AI systems. For more information, see [Microsoft's framework for building AI systems responsibly](https://blogs.microsoft.com/on-the-issues/2022/06/21/microsofts-framework-for-building-ai-systems-responsibly/).

It was an important step in our journey to develop better, more trustworthy AI. As such, we released our latest RAIS to date to share what we have learned, invite feedback from others, and contribute to the discussion about building better norms and practices around AI.

**The** **Responsible AI Standard sets out our best thinking on how we will build AI systems to uphold these values and earn society’s trust. It provides specific, actionable guidance for our teams that goes beyond the high-level principles that have dominated the AI landscape to date.**

#### Understanding the structure of the Microsoft Responsible AI Standard

In terms of the “anatomy” of the RAIS, it comprises different levels:

1. Our principles tell us where we want to go. As already outlined, they are our North Star; they state values that we must uphold when designing, developing, deploying or using AI systems. These are our enduring values that guide our responsible AI work.
2. We have broken them down into general goals that define what it means to uphold our principles. The goals tell us what specific outcomes we need to secure in the AI system to achieve this principle.

Adding goals was an innovation for us as this activated problem-solving thinking and helped to frame a better understanding of the requirements that were being asked. The outcomes are high-level and apply to many different AI systems and technologies.

1. At the next level, each of these goals is accompanied by a set of specific requirements. These are the concrete steps that we must to take to secure the goals.

Many of these requirements aim at mitigating risk & harms, and other potential damages of AI systems while respecting our values. These requirements may vary depending on the application or technology considered, but they will remain common in most situations. This is extremely important because if we did not have this structure and each system was completely different, we could not create reproducible tools and practices.

1. Some requirements are therefore accompanied by recommended tools and practices when they exist to help teams comply with them. They detail the aids to help us to meet the requirements, but many more tools and practices will be necessary as this nascent field evolves.

You can see hereafter how we took our six principles, which constitute our North Star, and broke them down into fourteen general goals.  Goals and associated requirements are divided into groups that must be completed within specified timeframes.

Several different colored boxes

Description automatically generated with medium confidence

Figure 6. Microsoft’s Responsible AI standard’s goals at a glance

If we take Accountability, a team will have achieved all the goals expressed for this principle:

1. (A1) When an impact assessment has been conducted for the AI system.
2. (A2) When the AI system use cases have been reviewed to determine if any of them meet the criteria for a "sensitive use". Sensitive uses such as higher education admission systems or predictive maintenance products for safety-critical environments, like healthcare or chemical processing, may impose additional requirements.
3. (A3) Another goal to achieve is purpose adequacy, meaning the AI system achieves its goals and service levels. This may mean that the models that are part of the product are defined and documented.
4. (A4) The data governance and management goal defines requirements for data documentation and assessment of data quality and its suitability for the AI system's purpose.
5. (A5) Finally, human oversight and control involve ensuring that feedback loops, such as monitoring or auditing, are in place to detect errors, anomalies, or other information about how the system behaves in practice.

Likewise, regarding Transparency, the goals are to clearly communicate the capabilities, limitations, and potential risks to all stakeholders. It is important to ensure that users understand how the AI system works, how decisions are made, and that stakeholders have the necessary information to make informed decisions on how to use, deploy, and operate it. If you are the data scientist who developed the model, it may be perfectly logical to you, but for other types of stakeholders - like a member of your team, a business, or a chemistry researcher relying on the system - they need to understand the system's capabilities and limitations.

As for Fairness, there are only three goals, but the directions related to this standard cover about a third of the document. There are many references and resources to help teams align with fairness.

The goals of Reliability & Safety aim to minimize the risk of error. For example, *how is the system tested? Do the tests simulate the real environment and use cases? How will the system react in case of failure or service degradation?*

As you can see, each requirement needs specific evaluations to be fulfilled. For Privacy & Security, as well as Inclusiveness, we refer to other standards or policies already in place at Microsoft. The intention is to build on these other policies and standards rather than duplicating efforts in the field of responsible AI. That being said, one should keep in mind that these policies and standards in place are also subject to evolve. As an example for Security, and while the core principles remain the same (*what are the inputs? Where does the output go? What are the dependencies on third-party components?* etc.), threat modeling and red teaming will be a bit different for AI compared to traditional software projects.

The RAIS asks every FTE to:

* Learn about our Responsible AI assets to understand and raise awareness about the subject company wide.
* Follow the RAIS goals, fulfill the related requirements and recognize the occasions where particular domains or technologies might be impacted.
* Report sensitive uses where each employee is invited to work with their teams to detect and assess sensitive use cases of AI and report these to ORA for assistance and guidance.
* Ask for help and reach out to their point of contact when they are uncertain or wish to report a sensitive use case. This role is devoted to the abovementioned Responsible AI Champs community.

#### Conducting Responsible AI Impact Assessments

As per above goal A1 of Accountability, compliance with the RAIS involves an impact assessment. As illustrated below, this goal is then broken down into three requirements:

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

Figure 7. Goal A1: Impact Assessment

The first requirement, A1.1, is about i) identifying the potential impact of an AI system on people, organizations, and society, ii) determining if a sensitive use assessment is necessary, and iii) eventually determine which objectives and requirements of the RAIS apply to the intended AI system.

Impact assessments have proven valuable at Microsoft to ensure that teams thoroughly explore the impact of their AI system - including stakeholders, expected benefits, and potential disadvantages - from the very earliest stages of design. Stakeholder analysis brings these to light and allows teams to examine the impact the system may have on them. Questions related to the objectives of the standard help identify potential harm.

As such, the impact assessment model is an evolving document, and the process of conducting an impact assessment is highly iterative. You can revisit the impact assessment as you deepen your understanding of the challenges of responsible AI within the project.

The [Microsoft Responsible AI Impact Assessment Guide](https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Guide.pdf) and the [Impact Assessment template](https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Template.pdf), also publicly released in June 2022, can be used side by side to help the impact assessment’s questions and case studies, see [Microsoft's framework for building AI systems responsibly](https://blogs.microsoft.com/on-the-issues/2022/06/21/microsofts-framework-for-building-ai-systems-responsibly/) for more information.

#### Managing Sensitive Uses (SUs)

Some AI systems require more attention and oversight.

Our Sensitive Uses (SUs) program, developed in 2017, provides ongoing review and oversight of novel or higher-risk uses of AI systems that may involve or result in one of the following:

1. Consequential impacts to a person’s legal status or access to essential services. The scenario involves the use of AI in a way that may directly result in the denial of consequential services or support to an individual (e.g., financial, housing, insurance, education, recruiting, or healthcare services or support).
2. Risk of physical or psychological injury. The scenario involves the use of AI in a way that may create a significant risk of physical or emotional harm to an individual (e.g., life-or-death decisions in military contexts, safety-critical manufacturing environments, clinical decision making in healthcare, or almost any case involving children or other vulnerable populations).
3. Threats to fundamental human rights. The scenario involves the use of AI in a way that may result in a significant infringement of one’s human rights.

This program for detecting SUs and potential for unintended consequences has helped us navigate the grey areas that are inevitably encountered and leads in some cases to new red lines. Outcomes of the process include giving up opportunities to build and deploy specific AI systems because we were not confident that we could do so in a way that upheld our principles.

The framework to guide Yes/No decisions is of no use if the answer is always ‘Yes’ - you have to say ‘No’ - Most of the time it will not be Yes/No, but instead, a compromise and an outline of how we reframe the technology to ensure we are responsible and in line with our principles.

## Establishing a multi-stakeholder approach and a global dialogue

Throughout our (ongoing) responsible AI journey, the continuous development of the RAIS, and our Sensitive Uses (SUs) process, we have learned a lot and continue to learn.

This process has helped us navigate the inevitably encountered gray areas and has led, in some cases, to the establishment of new red lines. Among the outcomes of this process, we have refused to build and deploy specific AI systems because we were not sure we could do so in accordance with our principles.

By working through the complexities of several other cases, we have also become aware of the importance of three key lessons:

1. First, by digging into the details of use cases, we’ve been able to understand and articulate their different risk profiles, such as the impact on failure and misuse on stakeholders, and the readiness of the technology for the particular use case.
2. Second, we’ve learned the important role that benchmarking and operational testing play, helping to ensure that AI systems serve their stakeholders well and meet quality bars not just in labs, but also in the real world*.*
3. And third, we’ve learned how we need to communicate with our customers to empower them to deploy their AI systems responsibly.

There is an important point we would like to reiterate: this is just the beginning. As (non-Generative vs. Generative) AI is a relatively new field, it is not surprising that the processes surrounding it are also evolving rapidly. In the future, we plan to refine our governance policies and standards as we invest increasingly more and more in AI, and we recommend other companies do the same.

A multi-stakeholder approach and a global dialogue is definitely needed on these issues. No single company, sector of the economy, or discipline will have all the answers on its own.

This multi-stakeholder approach is indeed essential. We collaborate with the best and brightest to advance the industry and society. Microsoft is notably a founding member of:

* The [Frontier Model Forum](https://www.frontiermodelforum.org/), an industry body ensuring the safe and responsible development of cutting-edge Generative AI models (see [Microsoft, Anthropic, Google, and OpenAI launch Frontier Model Forum](https://blogs.microsoft.com/on-the-issues/2023/07/26/anthropic-google-microsoft-openai-launch-frontier-model-forum/)).

Microsoft leverages the technical and operational expertise of member companies to benefit the entire AI ecosystem. This includes advancing technical evaluations and benchmarks, developing a public library of solutions to support industry best practices and standards, and ensuring fair and ethical access to these models.

* The [Partnership on AI](https://www.partnershiponai.org/), a non-profit partnership of academic, civil society, industry, and media organizations creating solutions so that AI advances positive outcomes for people and society.
* The [Rome Call for AI Ethics](https://www.romecall.org/), a document signed by the Pontifical Academy for Life, Microsoft and others to promote an ethical approach to artificial intelligence and a non-profit organization to promote anthropological and ethical reflection about the effects of new technologies on human life.
* The BID's [fAIrLAC+](https://fairlac.iadb.org/en) initiative, a partnership between the public and private sectors, civil society and academic institutions, designed to influence public policy and the entrepreneurial ecosystem in the promotion of the responsible and ethical use of AI.

As AI policies are considered and take shape globally, Microsoft is taking steps to address the highest-risk or particularly sensitive AI uses. More broadly, governments, civil society, and industry must come together to ensure that laws, standards, and norms are put in place to guide the responsible use of AI as it becomes an increasingly significant part of our lives. For more information, see section Going forward below.

If AI is advanced responsibly, it can create ripples of positive impact across the globe. AI has the potential to either narrow or widen existing social and economic divides - managing AI deliberately will help ensure that we maximize its potential benefits for the good of everyone. There are 700 million people in the world who lack access to reliable electricity and nearly 2.8 billion who lack access to the internet. To ensure that society can fully reap the benefits of AI, countries around the world will need to work together to close these gaps in accordance with the [UN’s Sustainable Development Goals](https://www.undp.org/sustainable-development-goals).

## Releasing responsible AI tools and practices

As we’ve been rolling out our responsible AI (RAI) program across the entire company, teams have often asked about the existence of engineering systems and tools to help them deliver on our commitments. Of course, tooling – particularly in its most technical sense – is not capable of the deep, human-centered thinking work that needs to be done in conceiving of (non-Generative vs. Generative) AI systems.

But we do think it is important to develop repeatable tools, patterns, and practices where possible, so that the creative thinking of our engineering teams can be directed toward the most novel and unique challenges, instead of reinventing the wheel to instrument their work. It is therefore necessary to provide teams with the results of cutting-edge research, practical advice, and innovative tools - from the beginning of development lifecycle so that others can be inspired by them and use them on a large scale, thus creating a virtuous circle.

In addition, integrated systems and tools contribute to consistency and ensure that responsible AI is part of the ‘coding path’ - it is part of the daily working method of our engineering teams -.To meet this need, we launched an initiative to build the ‘paved road’ of responsible AI at Microsoft.

We are building and setting up a set of tools, patterns, practices, and knowledge resources (e.g., RAI-specific knowledge, training, and education resources) that help teams easily integrate the requirements of responsible AI into their daily ideation practices, design practices, development practices and deployment practices.

### About tools

Our tools are a result of collaboration across disciplines to strengthen and accelerate responsible AI, spanning software engineering and development to social sciences, user research, law, and policy. These tools range from a variety of asset types.

At the highest level, we launched a series of research papers that set the context behind the responsible use of AI systems. To enable further collaboration, we started to make some of our early work available in the form of toolkits, libraries or datasets available in open source that others can in turn use to contribute and build upon. We will cover some of them as part of the next module: Module 2: What about a Responsible AI Lifecycle?.

The reasons for going with an open source approach are numerous:

* Freely available tools for RAI are an educational resource for learning and teaching the practice of RAI.
* More contributors, both internal to Microsoft and external, add quality, longevity, and excitement to the work and topic.
* The ability to integrate them into any platform or infrastructure encourages more widespread use.

We also democratized this tooling through our managed services offered through [Azure AI](https://azure.microsoft.com/en-us/solutions/ai/), i.e., our AI development platform to build AI solutions at enterprise scale. This tooling was and continues to be also (progressively) integrated into the Azure AI portfolio, which comprises:

* [Azure AI Studio](https://azure.microsoft.com/en-us/products/ai-studio/) to develop Generative AI solutions and custom copilots.
* [Azure AI Search](https://azure.microsoft.com/en-us/products/ai-services/ai-search/) to deliver accurate, hyper-personalized responses in context.
* [Azure OpenAI Service](https://azure.microsoft.com/en-us/products/ai-services/openai-service/) and the [Model Catalog](https://learn.microsoft.com/en-us/azure/machine-learning/concept-model-catalog?view=azureml-api-2) to quickly develop Generative AI experiences with a diverse set of prebuilt and curated (open source) models from OpenAI, Meta, NVIDIA, Mistral AI, and beyond.
* [Azure Machine Learning (Azure ML)](https://azure.microsoft.com/en-us/products/machine-learning) to leverage an enterprise-grade (non-Generative vs. Generative) AI service for the end-to-end ML lifecycle, MLOps practices, etc.
* Etc.

Azure AI serves as the foundation for this above-mentioned ‘paved road’, so that i) our customers will benefit from our development of engineering systems and tools too, and conversely ii) we can get feedback, continue to learn, and quickly improve such a foundation. This real-world experience and feedback is critical and can't be fully replicated in a lab. This establishes a virtuous cycle.

All of these tools continue to evolve. We’ve made good progress on understanding and mitigating technical and sociotechnical issues with deploying AI in the open world - but there’s much more to do.

Moving from principles to practices is difficult given the complexities, nuances, and dynamics of AI systems. There are no quick fixes and no silver bullet that address all risks with applications of AI technologies. But we can make headway by harnessing the best of research and engineering to create tools aimed at the responsible development and fielding of AI technologies.

### About Practices

Putting responsible AI into action starts with practices that are grounded in human-centric design that spans roles.

We’ve taken over 20 years of research and have applied it in the development of AI guidelines and standards that are meant to help others anticipate and address potential issues throughout the software development lifecycle and develop AI systems in a more responsible manner.

Une image contenant texte, capture d’écran, logiciel, Site web

Description générée automatiquement

Some of our guidelines include [Guidelines for Human-AI Interaction](https://aka.ms/aiguidelines), now as part of the newly released [Human-AI eXperience (HAX) Toolkit](https://www.microsoft.com/en-us/haxtoolkit/) (See [New toolkit aims to help teams create responsible human-AI experiences](https://blogs.microsoft.com/ai-for-business/hax-toolkit/)), [Conversational AI Guidelines](https://www.microsoft.com/en-us/research/publication/responsible-bots/), [Inclusive Design Guidelines](https://download.microsoft.com/download/b/0/d/b0d4bf87-09ce-4417-8f28-d60703d672ed/inclusive_toolkit_manual_final.pdf), an [AI Fairness Checklist](https://www.microsoft.com/en-us/research/project/ai-fairness-checklist/), and a [Datasheets for Datasets](https://www.microsoft.com/en-us/research/publication/datasheets-for-datasets/).

These learnings helped inform new practices at Microsoft. For example, we developed Transparency Notes, a new form of documentation that is intended to help our customers understand:

* How our AI technology works,
* What the capabilities and limitations are,
* The choices if any to possibly influence performance and behavior,
* And the importance of thinking about the whole system, including the technology, the people, and the environment,

so they have the knowledge necessary to make responsible choices. As such, Transparency Notes fill the gap between marketing and technical documentation, proactively communicating information that our customers need to know to deploy AI responsibly.

Our customers can use Transparency Notes when designing, developing or deploying their own (non-Generative vs. Generative) AI system(s): they can understand when and how to leverage and use our technologies. or share them with the people who will use or be affected by their system(s).

Our [Face API Transparency Note](https://azure.microsoft.com/en-us/resources/transparency-note-azure-cognitive-services-face-api/) was our first attempt at this new practice, and we now have a growing number of Transparency Notes being prepared across our Azure AI platform offerings and already available outside - You can Bing to search and retrieve all of them : [Overview of Responsible use of Azure AI services](https://learn.microsoft.com/en-us/azure/ai-services/responsible-use-of-ai-overview), [Transparency Note for Azure OpenAI](https://learn.microsoft.com/en-us/legal/cognitive-services/openai/transparency-note?tabs=text), [Transparency Note for Azure AI Search](https://learn.microsoft.com/en-us/legal/search/transparency-note?tabs=enrichment), etc.- .

We also proactively defined a [Code of Conduct](https://docs.microsoft.com/en-us/legal/cognitive-services/speech-service/tts-code-of-conduct?context=%2Fazure%2Fcognitive-services%2Fspeech-service%2Fcontext%2Fcontext), [retired capabilities](https://azure.microsoft.com/en-us/blog/responsible-ai-investments-and-safeguards-for-facial-recognition/) for our above-mentioned Face API, etc. Through these and other controls, we helped protect against misuse, while maintaining beneficial uses of the technology.

We also see synergies between our controls es and other industry efforts such as [Model Cards](https://modelcards.withgoogle.com/about), [AI FactSheets](https://www.ibm.com/blogs/research/2018/08/factsheets-ai/), etc., and we’re pleased to be playing an active role in the [Partnership on AI](https://www.partnershiponai.org/)’ initiative to evolve the artifacts and processes for responsible AI industry-wide, see section Establishing a multi-stakeholder approach and a global dialogue above.

Our [AI Security Guidelines](https://blogs.microsoft.com/on-the-issues/2019/12/06/ai-machine-learning-security/) are another illustration of these practices. Established in collaboration with Harvard University (Berkman Klein Center for Internet and Society at Harvard University), they are a series of findings we share that can protect your AI systems with guidance materials for modeling, detecting, and mitigating security risks and ethics issues.

Same is true with the [AI security risk assessment framework](https://aka.ms/airiskassessment) to empowers our customers to reliably audit, track, and improve the security of the AI systems. This framework provides a comprehensive perspective to AI systems’ security. Each element of an AI system’s lifecycle is looked at in a production setting from data collection, data processing, to model deployment. Similarly, AI supply chains are accounted for as well as the controls and policies with respect to backup, recovery, and contingency planning related to AI systems.

For more information, see [Failure Modes in Machine Learning](https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning), as well as the guide [Framing a (more) Trustworthy AI Lifecycle for your AI-powered solutions](https://github.com/microsoft/responsible-ai-workshop/blob/main/nongen-ai-trustworthy-ai-lifecycle/docs/framing-trustworthy-ai-lifecycle.docx), also part of this Responsible AI Workshop.

## Going forward

**"***Establishing codes of conduct early in the development of this emerging technology will help ensure its overall safety, security, and trustworthiness. It will also allow us to better unlock AI’s positive impact for communities around the world****.*"**

***-*** *Brad Smith, Vice Chair and President, Microsoft Corporation*

We’ve been engaged in advancing AI responsibly for the past eight years. (We’ve been advancing other technology responsibly for nearly fifty years, experience that informs our thoughtful approach to AI.)

But there’s an important point we’d like to stress: This is just the beginning. AI is still a relatively new field, so it should come as no surprise that the processes around it are evolving rapidly as well. Going forward, we plan on refining our governance policies as we invest further in (non-Generative vs. Generative) AI, and we advise other businesses do the same.

As we look ahead:

1. First, we’ll focus on consistently and systematically enacting our principles through the continued rollout of our (still evolving and improving) Microsoft Responsible AI Standard (RAIS) and the related governance framework in place, see section Translating AI principles into practice above. We are acutely aware that, as the adoption of (non-Generative vs. Generative) AI technologies accelerates, new and complex ethical challenges will arise. There is no finish line for responsible AI.

While we recognize that we don’t have all the answers, the above-discussed building blocks of our approach to responsible AI at Microsoft are designed to help us stay ahead of these challenges and enact a deliberate and principled approach at scale. We’re also committed to sharing our learnings regularly. Our commitments to transparency in our AI efforts extends beyond the [White House’s Voluntary AI Commitments](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/), see [Our commitments to advance safe, secure, and trustworthy AI](https://blogs.microsoft.com/on-the-issues/2023/07/21/commitment-safe-secure-ai/) for more information. The commitments Microsoft has made are detailed in the document [Voluntary Commitments by Microsoft to Advance Responsible AI Innovation](https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2023/07/Microsoft-Voluntary-Commitments-July-21-2023.pdf).

These commitments also led to the release of this inaugural annual report: [Responsible AI Transparency Report May 2024: How we build, support our customers, and grow](https://nam06.safelinks.protection.outlook.com/?url=https%3A%2F%2Faka.ms%2FRAITransparencyReport2024&data=05%7C02%7Cbwesolowski%40microsoft.com%7Cadb2e8601c064fd3c22e08dc69029e90%7C72f988bf86f141af91ab2d7cd011db47%7C1%7C0%7C638500707207440864%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C0%7C%7C%7C&sdata=REtL7ICJQIiDCXY0Ou40wSk2K01n5fsVWIaRv5iVqhQ%3D&reserved=0). Beside the content of this guide and the rest of this workshop, we believe we have an obligation to share our practices with the public as they evolve.

This report is a look back on Microsoft’s approach to developing AI responsibly and part of fulfilling our broader transparency commitments around our responsible AI efforts as covered above. It enables us to record and share our maturing practices, reflect on what we have learned, chart our goals, hold ourselves accountable, and earn the public’s trust.

We will continue to share what we learn, and we welcome opportunities to learn with others, see section Establishing a multi-stakeholder approach and a global dialogue above.

1. Second, governments are regulating AI. Governments around the world are indeed increasingly incorporating AI into various sectors, and this rise in AI usage heralds an imminent wave of regulatory measures and policies.

* One of the main concerns is the impact of AI on national security. The potential of AI to introduce new security threats, such as advanced chemical, biological, radiological or nuclear (CBRN) weapons that may pose a hazard to populations, territories, and forces, or sophisticated cyberattacks, represents a significant challenge for existing security paradigms.
* Another critical area is the issue of algorithmic bias. The concern here is the unintentional harm AI could cause to vulnerable groups due to the inherent biases in its algorithms. This could lead to discrimination and exacerbate social inequalities, highlighting the need for careful scrutiny and regulation of AI systems.
* The workforce is also at a crossroads with the advent of AI. The replacement of human workers by AI and other automation technologies is more than a mere probability that could have significant socio-economic implications. This change requires a thoughtful approach to manage the transition and mitigate potential negative effects on employment.

As AI policies’ conversations expanded last year, we shared the report [Governing AI: A Blueprint for the Future](https://aka.ms/May25WhitePaper) that details five ways governments should consider policies, laws, and regulations around AI, see [How do we best govern AI? - Microsoft On the Issues](https://blogs.microsoft.com/on-the-issues/2023/05/25/how-do-we-best-govern-ai/) for more information.

Interestingly enough, these AI policies’ conversations started to be punctuated by repeated references to unexpected abbreviations. Not the usual short names for new AI models or ML jargon, but acronyms for the different international institutions that today govern civil aviation, nuclear power, global capital flows, etc.

History has long taught us that the way in which technology transforms our world is in part a product of how effectively it is governed, and that international governance is vital for technologies that know no borders. To this end, we’ve recently shared the eBook [Global Governance: Goals and Lessons for AI](https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW1lhQ0), a collection of external perspectives on international institutions from different domains, brought together with our own thoughts on goals and frameworks for global AI governance. Through case studies and analysis, experts chart the history and evolution of institutions such as the [International Civil Aviation Organization (ICAO)](https://www.icao.int/pages/default.aspx), the [International Atomic Energy Agency (IAEA)](https://www.iaea.org/), and the [Financial Stability Board (FSB)](https://www.fsb.org/) and share insights on their successes and challenges to inform the global governance of AI, see [Global Governance: Goals and Lessons for AI](https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW1lhQ0) for more information.

1. Third, there will be increasingly mature responsible AI toolsets for data engineers, data scientists, AI engineers, and other AI practitioners – beyond the ones discussed previously and/or that we will discuss in the other guides part of this Responsible AI Workshop to help you mitigate these AI risks.

These tools are or will be made available to customers, not only as far as we are concerned, but also as an industry effort.

Ultimately, we will continue to build a culture of Responsible AI across the company. We cannot stress enough that we are just at the beginning of this journey.

# Module 2: What about a Responsible AI Lifecycle?

As you’ve seen so far, we endeavor to put our AI principles into practice with governance systems that shape the design, the development, the deployment, and the monitoring of (non-Generative vs. Generative) AI systems throughout their end-to-end lifecycle sustained by [MLOps practices](https://learn.microsoft.com/en-us/azure/machine-learning/concept-model-management-and-deployment?view=azureml-api-2) for non-Generative AI vs. [LLMOps ones](https://techcommunity.microsoft.com/t5/ai-machine-learning-blog/an-introduction-to-llmops-operationalizing-and-managing-large/ba-p/3910996), i.e., MLOps for Large Language Models (LLMs) for Generative AI.

The set of needs in this field we’re addressing within Microsoft and as part of the larger tech industry is quite unique:

* Identifying the potential benefits, risks and harms to design a system responsibly requires attention to the specific context in which this system is deployed, rather than a “checklist” approach. There is no one, straightforward solution – we need to weigh the options of the best solution for a specific context.
* Many of the largest challenges come from intersections between two or more principles, for example preserving privacy while ensuring we know enough about sensitive attributes to evaluate and manage fairness or providing enough transparency about how systems function while making sure we’re not opening security risks.
* Focus on identifying opportunities and challenges in context and managing tensions between benefits and harms, principles, and stakeholders – This is work that requires thoughtful collaboration and deliberation, often returning to the same discussion several times to develop a well-prioritized plan for addressing the challenges in a project.
* There is no perfect solution – the challenge is balancing the benefits of a system along with the potential harms and managing for future impact as well as we can, understanding that we are often just one part of a larger ecosystem accountable for the system.
* Recognize that issues can come up at any stage of the product development cycle and we need to keep looking for them, the upside is that there are mitigation techniques for different development stages from ML research to user experience (UX) design as part of the Human-AI (H-AI) experiences to provide.
* We can use these techniques to create new products and/or features, and to improve existing ones.

Thus, in terms of an AI system’s lifecycle, Responsible AI is not just about data science, it’s about ALL disciplines working together and benefitting from each other: program manager, data engineers, data scientist, ML developers, software engineer, (UX) designer, user researcher, content writer, marketing, and customer service.

Responsible AI is a human-centered approach to developing, designing, and deploying AI technologies and ML-powered systems (products or services) and features. It’s a mindset and a toolset, and this applies both for new and existing services and products.

The previously discussed building blocks help (secure DevOps) teams to identify, evaluate, and mitigate possible harms in their technology stack and products throughout the entire product development lifecycle, internally called the Responsible AI Lifecycle (RAIL) organized according to the key phases of such a lifecycle while recognizing that AI product development often cycles through these phases iteratively:

* Assess and prepare: Evaluate the product’s benefits, the technology, the potential risks, and the team.
* Design, build, and document: Review the impacts, unique considerations, and the documentation practice.
* Validate and support: Select the testing procedures and the support to ensure products work as intended.

To be able to leverage on that, responsible AI practices need to fuel and be a part of every stage of the service or product and models design process. All the learnings that can be gathered from these stages constitute several opportunities to foresee how the considered AI systems should evolve to provide even better, more responsible solutions.

Putting our AI principles to work means operationalizing these stages with as a set of activities to conduct at each stage, the same way we’ve done with cybersecurity, privacy, and accessibility which are, as already abovementioned, systemically integrated into our services and products design, development and deployment practices. We endeavor to implement the same approach for our AI principles, in the form of governance systems that shape the design, the development, the deployment, and the monitoring of these systems.

We explore activities to be conducted at each stage of an AI system’s lifecycle below. For the sake of the illustration, we will consider a non-Generative AI system.

For more information, see [Building AI responsibly from research to practice](https://blogs.microsoft.com/ai-for-business/building-ai-responsibly-from-research-to-practice/), as well as the guide [Implementing a Responsible AI Lifecycle for MLOps processes](https://github.com/microsoft/responsible-ai-workshop/blob/main/nongen-ai-lifecycle-walkthrough/docs/implementing_responsible_ai_lifecycle.docx), also part of this Responsible AI Workshop, as an illustration for non-Generative AI.

For guidance for your Generative AI-powered solutions, please refer to the guide [Responsible AI Workshop - Building and using Generative AI responsibly with Azure and beyond](https://github.com/microsoft/responsible-ai-workshop/blob/main/nongen-ai-tooling-tutorials/docs/leveraging-responsible-ai-tooling.docx)., also part of this Responsible AI Workshop.

## Assessing and preparing an AI system

Today, we understand that it is critically important for our employees to think holistically about the AI systems we choose to build. As part of this, we all need to think deeply about and account for sociotechnical impacts of these AI systems. That’s why we’ve developed trainings designed to help our teams develop the muscle of asking ground-zero questions, such as, “*Why are we building this AI system?*” and, “*Is the AI technology at the core of this system mature enough?”.*

And these are exactly the types of questions you would ask during the first stage of the RAIL with impact assessment. Conducting such an assessment is a required first step in the development of any AI system at Microsoft. It includes identifying high priority areas within your AI development, building a way to track and review the process, and securing approvals.

For that purpose, teams must complete an extensive questionnaire, which considers the intended use cases of a product and its potential impacts on stakeholders, and a self-assessment of the potential risks. This process is facilitated and required by ORA and serves as an important tool to help ensure the responsible design, development, deployment, and monitoring of AI system.

This assessment is made easy with the use of the following resources:

* The [Harms Modeling](https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/). A framework for product teams, grounded from our core pillars of responsible innovation. It examines how people's lives can be negatively impacted by technology: injuries, denial of consequential services, infringement on human rights, and erosion of democratic & societal structures. Similar to the [Security Threat Modeling](https://strikecommunity.azurewebsites.net/articles/1941/course-threat-modeling-101.html), a foundational activity as part of Microsoft SDL, Harms Modeling enables product teams to anticipate potential real-world impacts of technology, which is a cornerstone of responsible development.

Hereafter is an example of the outcome of a qualitative assessment. Such an outcome is used to inform prioritization of responsible innovation mitigations depending on the encountered [types of harm](https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/type-of-harm).

Une image contenant texte, capture d’écran, nombre, logiciel

Description générée automatiquement

Figure 6. Harms Modeling qualitative assessment

* The [Judgment Call Game](https://microsofteur-my.sharepoint.com/personal/t-alazraq_microsoft_com/Documents/Documents/microsoft/responsible_ai/deliverables/whitepaper_responsible_ai_tools/Judgment%20Call). An award-winning responsible innovation game and team-based activity that puts Microsoft’s AI principles into action. The game provides an easy-to-use method for cultivating stakeholder empathy by making participants write product reviews from the perspective of a particular stakeholder, describing what kind of impact and harms the technology could produce from their point of view. To learn more about this game, you can download the [printable Judgment Call game kit](https://download.microsoft.com/download/3/3/d/33da5224-fb6e-4591-827d-9c2bd9ac47c2/JudgmentCall_printable.pdf).
* The [Community Jury](https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/community-jury/).  A technique that brings together diverse stakeholders impacted by a technology. It is an adaptation of the [citizen jury](https://jefferson-center.org/about-us/how-we-work/). The stakeholders are provided with an opportunity to learn from experts about a project, deliberate together, and give feedback on use cases and product design. This responsible innovation technique allows project teams to collaborate with researchers to identify stakeholder values and to understand the perceptions and concerns of impacted stakeholders. Wherever the topic to discuss is related to personal data privacy, the composition of the community jury should be balanced to include individuals with different [privacy indices](https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/community-jury/#privacy-index).

For more information, see [Responsible Innovation: A Best Practices Toolkit](https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/).

## Designing, building, and documenting an AI system

This is the second stage of both designing, building, and documenting Responsible AI systems, i.e., human-centered systems, with a number of activities for the ML model AND for the AI system or the feature itself, respectively.

While, as far as the former is concerned, the related activities range from data collection and handling to ensuring fairness in performance and transparency of the model(s), building a Responsible AI system is also about building it in a human-centered way.

So, besides and beyond the ML model(s), the latter implies to:

* Tie all technical decisions back to user needs,
* Involve diverse perspectives early in the design of the AI system (see above section) and throughout,
* Plan for failures so users can recover (or take control) when things go wrong or an unexpected situation is encountered,
* Etc.

With that said, it is worth mentioning that this second stage that covers the design and the development of AI systems can be conducted either with today’s agile practices - with several iterations and sprints - or not.

## Validating and supporting an AI system

Then comes the third stage which is about the validation, the deployment, and the support of Responsible AI systems. This stage includes reinforcing practices and empowering people to use AI responsibly through documentation, gating, scenario attestation, and more.

As we have been rolling out our Responsible AI program across the company, the existence of engineering systems and tools to help deliver on our Responsible AI commitments has been a priority for our teams.

Privacy, and the General Data Protection Regulation (GDPR) experience in particular, has taught us the importance of engineered systems and tools for enacting a new initiative at scale and ensuring that key considerations are baked in by design.

This leads us to the next section.

## Leveraging engineering systems and tools, patterns & practices

Although tooling – particularly in its most technical sense – is not capable of the deep, human-centered thinking work that needs to be undertaken when conceiving AI systems, we think it is important to develop repeatable tools, patterns, and practices where possible so the creative thought of our engineering teams can be directed toward the most novel and unique challenges, not reinventing the wheel. Integrated systems and tools also help drive consistency and ensure that Responsible AI is part of our engineering teams’ everyday work.

In recognition of this need, we are embarking on an initiative to build out a “paved road” for Responsible AI at Microsoft with a set of tools, and patterns & practices that help teams easily integrate responsible AI requirements into their everyday development practices. As outlined, Azure ML serves as the foundation for this paved road, leveraging the early integrations of our open-source tools.

We started our journey “paving that road” by the need to assess current processes and figure out what we wanted to change.

As you see in the figure hereafter, tooling is needed for the different stages of the RAIL to sustain the implementation of the requirements for each stage. Consequently, we have and continue to define:

1. *What needs to happen at each stage in terms of implementation (directions)?*
2. *Which tools are available, needed?* Etc.

These resources can broadly be put into two categories:

1. Guidelines, guidance, patterns & practices, which also include trainings, workshops and assessment tools and games among others.
2. Technical tools, which are vehicles to (help) understand, assess, and mitigate the AI risks inherent of ML models and the datasets to train them and/or used for inference, and other AI issues which might arise when implementing AI systems along with their ML model(s).

For more information, see [New resources and tools to enable product leaders to implement AI responsibly](https://azure.microsoft.com/blog/new-resources-and-tools-to-enable-product-leaders-to-implement-ai-responsibly/).

### Guidelines, patterns & practices

Beyond the abovementioned resources to help in the initial (impact) assessment, some resources already provide a great, awaited, and welcome contribution (with more to come).

#### Responsible Human-AI partnership

As far as the above resources are concerned, this is notably the case for the [Human-AI eXperience (HAX) Toolkit](https://www.microsoft.com/en-us/haxtoolkit/). for creating (more) responsible Human-AI partnership.

The HAX Toolkit which is a flexible set of hands-on tools for building effective experiences in user-facing AI products. You can mix and match tools depending on your needs, use case, and where you are in your product’s life cycle. But if you need guidance, here are our suggestions for getting started:

1. **Learn**. Familiarize yourself with the [Guidelines for Human-AI Interaction](https://www.microsoft.com/en-us/haxtoolkit/ai-guidelines/). These eighteen Guidelines are evidence-based best practices for designing your AI user experiences. They prescribe how an AI system should behave.

Read an [overview](https://www.microsoft.com/en-us/haxtoolkit/ai-guidelines/)of the Guidelines and browse the [Design Library](https://www.microsoft.com/en-us/haxtoolkit/library/) to see Design Patterns for implementing them throughout the user experience: upon initial interaction, during interaction, when the AI system is wrong, and over time.

Une image contenant texte, capture d’écran, Police, Parallèle

Description générée automatiquement

Figure 7. Guidelines for Human-AI Interaction

1. **Plan.**Bring your team together and go through the [HAX Workbook](https://www.microsoft.com/en-us/haxtoolkit/workbook/). The HAX Workbook is a discussion and planning guide. Your HAX Workbook session will result in prioritized work items for which Guidelines to implement first, and how.
2. **Design**. Use [Design Patterns](https://www.microsoft.com/en-us/haxtoolkit/design-patterns/) and examples in the [HAX Design Library](https://www.microsoft.com/en-us/haxtoolkit/library/) to identify ways to apply the Guidelines.
3. **Prototype**. If you’re creating an AI system that uses natural language processing (NLP), use the [HAX Playbook](https://www.microsoft.com/en-us/haxtoolkit/playbook/) to identify how the system will fail so you can plan & prototype solutions for helping users recover from inevitable failures.

To learn how to use the HAX Toolkit to create human-centered AI, watch the video [Create human-centered AI with the Human-AI eXperience (HAX) Toolkit](https://www.youtube.com/watch?v=JppYmctp0a8).

#### Privacy and Security

Yet another dimension is regarding this time the Privacy and Security principles. We share in collaboration with Harvard University a series of findings that help protect AI systems with guidance materials for modeling, detecting, and mitigating security risks and ethics issues. For more information, see [Solving the challenge of securing AI and machine learning systems](https://blogs.microsoft.com/on-the-issues/2019/12/06/ai-machine-learning-security/).

Furthermore, we have also revisited our security considerations and in turn extended our practices for [Microsoft Security Development Lifecycle (SDL)](https://www.microsoft.com/en-us/securityengineering/sdl/). Microsoft SDL introduces security considerations throughout all phases of the development process, helping teams build highly secure software, addressing security compliance requirements, and reducing development costs. The guidance, best practices, tools, and processes in the Microsoft SDL are practices we use internally to build more secure products and services. Since first publicly shared back in 2008, we’ve regularly updated the SDL as a result of our growing experience with new scenarios, like here AI, and to keep it aligned with threats as they evolve.

A series of publications has been released to this end. A first publication “[Failure Modes in Machine Learning](https://docs.microsoft.com/en-us/security/failure-modes-in-machine-learning)” lays out the terminology we developed jointly with Harvard University’s [Berkman Klein Center](https://cyber.harvard.edu/). It includes vocabulary that can be used to describe intentional failure caused by an adversary attempting to alter results or steal an algorithm as well as vocabulary for unintentional failures like a system that produces results that might be unsafe.

This taxonomy informs two other publications and delivers new threat modeling, detection, mitigation and triage guidance in use today at Microsoft as part of our established security practices:

1. “[Threat Modeling AI/ML Systems and Dependencies](https://docs.microsoft.com/en-us/security/threat-modeling-aiml)” specifically for AI systems.
2. “[AI/ML Pivots to the Security Development Lifecycle Bug Bar](https://docs.microsoft.com/en-us/security/bug-bar-aiml)” to systematically triage attacks on AI systems.

We are committed to protecting our customers by providing security updates and guidance that address vulnerabilities when they are reported, and we are always looking for ways to provide clarity and transparency around how we assess the impact of vulnerabilities reported in our products and services.

The vulnerability types found in systems involving AI that Microsoft seeks to address with the continuously improving SDL are published publicly with severity ratings for AI systems. The resulting [Microsoft Vulnerability Severity Classification for AI Systems](https://www.microsoft.com/en-us/msrc/aibugbar) is an update to Microsoft’s existing vulnerability severity classification (i.e., our so-called “SDL bug bar”) to cover new vulnerability categories arising specifically from the use of AI in our products and services. It will continue to be updated as new risks emerge to better account for AI risks, including generative AI.

Finaly, the [Enterprise security and governance for AI](https://learn.microsoft.com/en-us/azure/machine-learning/concept-enterprise-security?view=azureml-api-2) for Azure ML customer provides as it title indicates guidance on enterprise security and governance. The features covered in the guidance might be useful for administrators, DevOps, and MLOps teams who want to create a secure configuration that is compliant with your policies in place.

### Technical tools

In terms of tooling, one of the challenges facing both experienced and inexperienced AI practitioners (data engineers, data scientists, etc.) is that harms can surface inadvertently in AI systems. Adverse behaviors and influences of AI systems can vary from trivial to deeply consequential. Flaws may remain concealed, or only be uncovered after applications are deployed, because they hide in algorithms, models, data and even assumptions. It can be a race against time to mitigate problems.

While growing numbers of tools and platforms have been available to help AI practitioners build AI systems, instruments to help ML engineers figure out what might go wrong are scarce. Several years ago, our aforementioned AETHER committee and Microsoft Research (MSR) recognized the need for a new class of tools and has coordinated and supported their creation. Frontier research has been important in studying the challenges with building and fielding AI systems, recognizing what needs to be done to build AI responsibly and developing methods to address potential failures.

To take the promising prototypes we built into practice, we relied on both research and engineering skills to develop robust, accessible tools that can be adopted by those who need them most. Our efforts have resulted in a set of open-source tools to help AI practitioners identify issues, diagnose causes, and mitigate problems before deploying AI systems, but also to accompany the effective deployment of the considered AI systems, and beyond.

All of the related available tools can be categorized into three broad categories each relating to our six principles:

1. Tools tounderstand the behavior of AI systems.
2. Tools to protect AI systems data.
3. Tools to establish control, monitoring, validation, and governance throughout AI systems development cycle (MLOps).

Let’s consider them in order. For the most prominent tools (we open-sourced) for non-Generative AI, the guide [Leveraging Responsible AI Tooling for your non-Generative AI-powered solutions](https://github.com/microsoft/responsible-ai-workshop/blob/main/nongen-ai-tooling-tutorials/docs/leveraging-responsible-ai-tooling.docx) further explores the first category of tools used to understand the behavior of AI systems, and will also tackle the tools to protect AI systems data.

#### Tools to understand the behavior of AI systems

These tools are used to make AI systems more fair, transparent, and inclusive. The [Responsible AI Toolbox](https://github.com/microsoft/responsible-ai-toolbox) falls “in this bucket.”

The Responsible AI Toolbox is an open-source framework with a suite of integrated and interoperable tools and functionalities providing data and model exploration and assessment, user interfaces and libraries that enable a better understanding of AI systems. These interfaces and libraries empower AI practitioners and stakeholders of AI systems to develop and monitor AI more responsibly, and take better data-driven actions.

In order to achieve these capabilities, the [Responsible AI Dashboard](https://nam06.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.microsoft.com%2Fen-us%2Fai%2Fai-lab-responsible-ai-dashboard&data=05%7C01%7CTanner.Briggs%40microsoft.com%7C32317c1fd8a3434c72c608dba35db25a%7C72f988bf86f141af91ab2d7cd011db47%7C1%7C0%7C638283395053584698%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=KGdgspbrCBFL6%2FXmP0%2Fji3QDqEIp5EXBAemMRlDt8Fw%3D&reserved=0) provides a single pane of glass, enabling you to easily flow through different stages of model debugging (identification, diagnostic, and mitigation), and decision-making.

It integrates together ideas and technologies from several open-source toolkits in the areas of:

* + Error Analysis powered by [Error Analysis](https://github.com/microsoft/responsible-ai-widgets/blob/main/docs/erroranalysis-dashboard-README.md), which identifies cohorts of data with higher error rate than the overall benchmark. These discrepancies might occur when the system or model underperforms for specific demographic groups or infrequently observed input conditions in the training data.

thumbnail image 4 of blog post titled 
 
 
  
 
 
 
    
  
   
    
      
       Responsible AI dashboard: A one-stop shop for operationalizing Responsible AI in practice
       
      
     
   
  
 
   
 
 
 
 
 


Figure 7. Cohorts of data with higher error rate in Error Analysis

* + Fairness Assessment powered by [Fairlearn](https://github.com/fairlearn/fairlearn), which identifies which groups of people may be disproportionately negatively impacted by an AI system and in what ways.
  + Model Interpretability powered by [InterpretML](https://github.com/interpretml/interpret-community), which incorporates state-of-the-art machine learning interpretability techniques under one roof, and explains blackbox models, helping users understand their model's global behavior, or the reasons behind individual predictions.
  + Counterfactual Analysis powered by [DiCE](https://github.com/interpretml/DiCE), which shows feature-perturbed versions of the same datapoint, which would have received a different prediction outcome.
  + Causal Analysis/Inference powered by [EconML](https://github.com/microsoft/EconML), which focuses on answering What If-style questions to apply data-driven decision-making.
  + Data Explorer powered by the [Responsible AI Dashboard](https://github.com/microsoft/responsible-ai-toolbox/blob/main/docs/databalance-README.md), which helps teams explore their dataset statistics and distributions, gain an overall understanding of their data, identify features receiving the positive outcome more than others, and visualize feature distributions. They can further compare the cohort data stats with other cohorts or the overall benchmark data.

The different components available in open source as well as through Azure Machine Learning (Azure ML) in the Azure AI platform, have been designed with each stage of the model improvement lifecycle in mind, informing targeted model improvement through error analysis, fairness assessment, data exploration, and interpretability.

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

Figure 8. ML model debugging via Responsible AI dashboard

For more information, see [Responsible AI: The research collaboration behind new open-source tools offered by Microsoft](https://www.microsoft.com/en-us/research/blog/responsible-ai-the-research-collaboration-behind-new-open-source-tools-offered-by-microsoft/?culture=en-us&country=us).

To this end, these components can easily communicate with each other. You can interactively pass cohorts of data and insights from one component to another for deep-dive investigations, without having to manually save and reload results in different dashboards. Tabular, image, and text data are supported. For more information, see [Responsible AI dashboard: A one-stop shop for operationalizing Responsible AI in practice](https://techcommunity.microsoft.com/t5/ai-machine-learning-blog/responsible-ai-dashboard-a-one-stop-shop-for-operationalizing/ba-p/3030944).

To learn how to use the Responsible AI Dashboard in Azure AI to assess AI systems and make data-driven decisions, see:

* [Assess AI systems and make data-driven decisions with Azure Machine Learning Responsible AI dashboard](https://docs.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai-dashboard).
* [How to use the Responsible AI dashboard in studio (preview)](https://docs.microsoft.com/en-us/azure/machine-learning/how-to-responsible-ai-dashboard) .

#### Tools to protect AI systems data

These tools are used to make AI systems more secure and privacy-preserving:

* [Presidio](https://microsoft.github.io/presidio/). A data protection and anonymization software development kit (SDK).
* [SmartNoise](https://smartnoise.org/). A differential privacy (DP) toolkit for analytics and machine learning helping to protect data used in AI systems by preventing re-identification.
* [Counterfit](https://github.com/Azure/counterfit/). A command-line tool and a generic automation layer for assessing the security of AI systems.

#### Tools to establish control, monitoring, validation, and governance

These are used to make AI systems more reliable and allows people who design and deploy AI systems to be held accountable for how their systems operate.

The [Datasheets for datasets](https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4t8QB), and the proposed revised [Aether Data Documentation Template](https://www.microsoft.com/en-us/research/uploads/prod/2022/07/aether-datadoc-082522.pdf) are yet another illustration of such a management tool. They are a tool for documenting datasets, helping creators and consumers of datasets think through underlying assumptions, potential risks, and implications of use.

Datasheets indeed contain questions about dataset motivation, composition, collection, pre-processing, labeling, intended uses, distribution, and maintenance. Crucially, and unlike other tools for meta-data extraction, datasheets are not automated, but are intended to capture information known only to the dataset creators and often lost or forgotten over time

Although not examined here any further for the sake of the length of this guide, you can learn more about governance and control tools by referring to the Policy section in [Microsoft Responsible AI center](https://aka.ms/RAI).

Regardless of the above three broad categories, this set of tools continues to evolve and to be enriched.

As an illustration, the above Responsible AI toolbox benefits from the following [newest additions](http://aka.ms/rai-mitigationstracker-blog):

* The [Responsible AI Mitigations Library](https://github.com/microsoft/responsible-ai-toolbox-mitigations), which enables you to explore different measurements and mitigation steps that may be most appropriate when the model underperforms for a given data cohort, and thus more easily experiment with different techniques for addressing failures.
* The [Responsible AI Tracker](https://github.com/microsoft/responsible-ai-toolbox-tracker), which uses visualizations to show the effectiveness of the different techniques for more informed decision-making within the same framework therefore enabling fast model iteration and evaluation processes.
* The [Responsible AI Gender Bias (GenBit)](https://github.com/microsoft/responsible-ai-toolbox-genbit), which helps you measure gender bias in Natural Language Processing (NLP) datasets. The main goal is to analyze your text corpora and compute metrics that give insights into the gender bias present in a corpus.

For more information, see [Share Responsible AI insights and make data-driven decisions with Azure Machine Learning](https://learn.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai-scorecard?view=azureml-api-2).

Likewise, the Responsible AI scorecard now allows to share your RAI key insights from the above Responsible AI Dashboard in Azure ML. As such, this is a PDF report you can extract and share with summaries of key data and model performance and fairness insights.

Une image contenant texte, capture d’écran, diagramme, nombre

Description générée automatiquement Une image contenant texte, capture d’écran, nombre, logiciel

Description générée automatiquement

Figure 8. Summaries from the Responsible AI scorecard

The first summary segment of the scorecard gives you an overview of the machine learning model and the key target values you have set to help all stakeholders determine if your model is ready to be deployed.

For more information, see [Responsible AI Dashboard and Scorecard in Azure Machine Learning](https://techcommunity.microsoft.com/t5/ai-machine-learning-blog/responsible-ai-dashboard-and-scorecard-in-azure-machine-learning/ba-p/3391068).

For more information, see [Building AI responsibly from research to practice](https://blogs.microsoft.com/ai-for-business/building-ai-responsibly-from-research-to-practice/), as well as the guide [Implementing a Responsible AI Lifecycle for MLOps processes](https://github.com/microsoft/responsible-ai-workshop/blob/main/nongen-ai-lifecycle-walkthrough/docs/implementing_responsible_ai_lifecycle.docx), also part of this Responsible AI Workshop, as an illustration for non-Generative AI.

# Module 3: Initiating/Improving your own Responsible AI journey

“I would argue that perhaps the most productive debate we can have isn’t one of good versus evil: The debate should be about the values instilled in the people and institutions creating this technology.”

*- Satya Nadella*

## Understanding Microsoft’s AI commitments

Our goal is not only to build (non-Generative vs. Generative) AI responsibly ourselves, but also to empower our customers to do the same

As part of the [Microsoft’s AI Customer Commitments](https://blogs.microsoft.com/blog/2023/06/08/announcing-microsofts-ai-customer-commitments/) announced last June 2023, Microsoft is helping organizations take a more holistic approach to responsible AI across their own development lifecycle by sharing what we are learning about developing and deploying AI responsibly, and thus also making available leading research, practical guidance and tooling innovation – early in development for others if relevant to build upon and broadly (through Azure AI) for organizations to use at scale.

As already touched, we’ve already made available our suite of 30 responsible AI tools that include more than 100 features, see section Releasing responsible AI tools and practices above, and our transparency documentation - which includes publishing 33 Transparency Notes since 2019.

While we recognize that every individual, organization, and region will have their own beliefs and standards that should be reflected in their own AI journey, we hope the previous modules provide you with a set of ideas to initiate your own journey and create a holistic approach to Responsible AI if you haven’t already done so.

## Defining your own AI principles

Defining your own AI principles constitutes your initial step as we introduced our six core principles back in 2018.

At the international level, the Organization for Economic Co-operation and Development (OECD) has also facilitated discussions on these topics. The resulting [OECD AI Principles](https://oecd.ai/en/ai-principles) were [initially adopted in May 2019](https://www.oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm) and [updated in May 2024](https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449).

These principles promote use of AI that is innovative and trustworthy and that respects human rights and democratic values, and set standards for AI that are practical and flexible enough to stand the test of time:

* [Inclusive growth, sustainable development and well-being](https://oecd.ai/en/dashboards/ai-principles/P5).
* Human rights and democratic values, including fairness [and privacy](https://oecd.ai/en/dashboards/ai-principles/P6).
* [Transparency and explainability](https://oecd.ai/en/dashboards/ai-principles/P7).
* [Robustness, security and safety](https://oecd.ai/en/dashboards/ai-principles/P8).
* [Accountability](https://oecd.ai/en/dashboards/ai-principles/P9).

They focus on how governments and other actors can shape a human-centered approach to safe, secure, and trustworthy AI. As an OECD legal instrument, they represent a common aspiration for the countries that adhere[[1]](#footnote-2) to it.

* In June 2020, the Group of the Twenty (G20), i.e., the main forum for international economic cooperation, adopted AI principles centered on humans, which are based on the OCDE’s AI principles. They call for users and developers of AI to be fair and accountable, with transparent decision-making processes and to respect the rule of law and values including privacy, equality, diversity and internationally recognized labor rights.
* In October 2023, the [Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI System](https://microsofteur-my.sharepoint.com/personal/philber_microsoft_com/Documents/Drafts/Responsible%20AI%20Workshop%202024/Hiroshima%20Process%20International%20Guiding%20Principles%20for%20Organizations%20Developing%20Advanced%20AI%20System) from G20 discusses and elaborates a non-exhaustive list of guiding principles as a living document to build on the existing OECD AI Principles in response to recent developments in advanced AI systems, i.e., Generative AI systems, and are meant to help seize the benefits and address the risks and challenges brought by these technologies.

To this extend, one should note that over the past few years, principles around developing AI responsibly have proliferated and, for the most part, there is overwhelming agreement on the need to prioritize issues like transparency, fairness, accountability, privacy, safety and security, see [Principled Artificial Intelligence - Mapping consensus in ethical and rights-based approaches to principles for AI](https://cyber.harvard.edu/publication/2020/principled-ai).

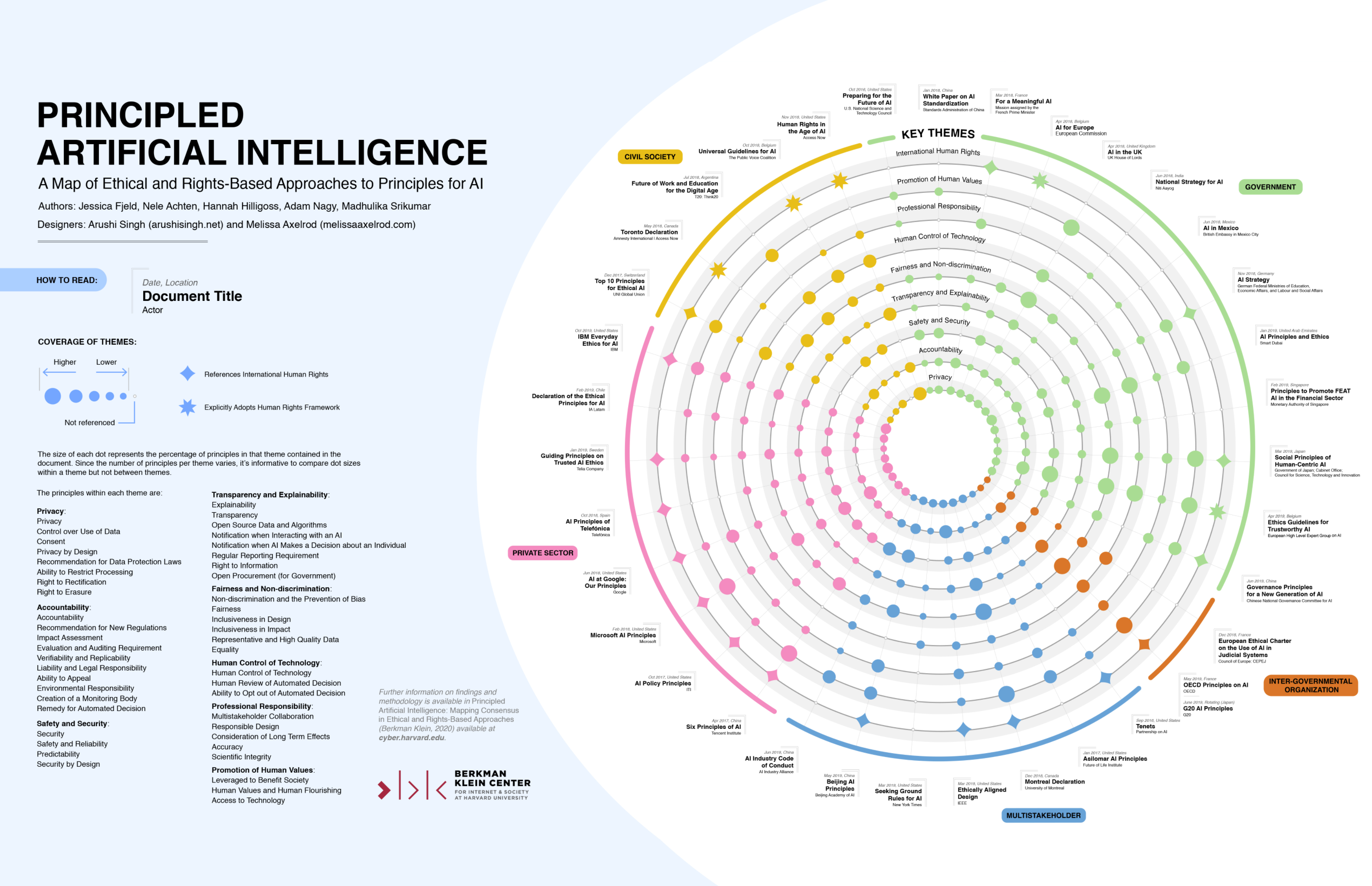


Figure 9. Principled Artificial Intelligence, source: <https://wilkins.law.harvard.edu/misc/PrincipledAI_FinalGraphic.jpg>

Once you have your Responsible AI principles – to take (yet) another example, Novartis, a leading, focused medicines company, has [eight principles for ethical use of AI](https://www.novartis.com/about/strategy/data-and-digital/artificial-intelligence/our-commitment-ethical-and-responsible-use-ai) in place -, next comes how to put these principles to work, i.e., into practice as previously discussed.

## Assessing your Responsible AI maturity

Although the vast majority of people believe in the importance of RAI, many organizations aren’t sure how to cross what is commonly referred to as the “Responsible AI Gap” between such above principles and tangible actions in their day-to-day development of AI systems.

Many organizations could even actually overestimate their responsible AI maturity, partially because they lack clarity on how to make their principles operational, see [Are You Overestimating Your Responsible AI Maturity?](https://www.bcg.com/publications/2021/the-four-stages-of-responsible-ai-maturity).

Additionally, while we do recognize that turning such principles into practice must consider engineering realities, we also collectively need new kinds of engineering tools aimed at helping to better understand and refine AI technologies.

The [Responsible AI Maturity Model (RAI MM)](https://www.microsoft.com/en-us/research/publication/responsible-ai-maturity-model/) (RAI MM) is a framework design to help organizations identify their current and desired levels of RAI maturity.

To this extend, RAI MM provides a map to the complex responsible AI territory, so our customers/teams can identify where they are and where they could go next on their own responsible AI journey.

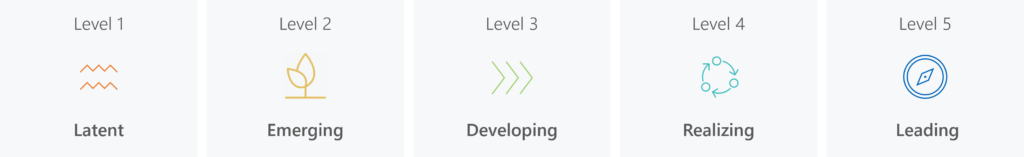


Figure 9. RAI MM levels

Whilst the previously discussed RAIS in Module 1: Building Responsible AI solutions above sets goals each AI systems should achieve at least from Microsoft’s perspective, the RAI MM is a tool for organizations to decide what investments might be needed to support RAI work and how people and teams can best work together to create RAI.

As such, the RAI MM contains twenty-four empirically derived dimensions that are key to an organization’s RAI maturity. Each dimension has five levels, going from low (Level 1: Latent) to high (Level 5: Leading) maturity. The dimensions are organized into three main categories:

1. Organizational Foundations.
2. Team Approach.
3. RAI Practice.

We recommend thinking of the RAI MM as a high-level map of the complex and evolving territory of RAI. Its use helps navigate what it means to be a mature RAI organization:

* Orientation: Read the dimension to understand what it takes to grow in RAI maturity.
* Goal-setting: Pick dimensions that matter to your organization or team; Discuss what levels you want to aim for on those dimensions; Plan for how to get there.

## Paving the road of your own journey

To help you proactively establish your principles, and in turn operationalize them for your own AI systems, and thus both anticipate and mitigate their AI risks, and maximize their benefits, you can:

Visit our [AI Business School page](https://www.microsoft.com/en-us/ai/ai-business-school?rtc=1) to help you establish and/or adapt a strategy which is suitable for you.

Take the [Microsoft Learn](https://docs.microsoft.com/en-us/learn/) training freely available online:

* [Discover ways to foster an AI-ready culture in your business](https://docs.microsoft.com/en-us/learn/paths/foster-ai-ready-culture/).
* [Identify principles and practices for responsible AI](https://docs.microsoft.com/en-us/learn/paths/responsible-ai-business-principles/) (for private organizations).

-or-

* [Identify guiding principles for responsible AI in government](https://docs.microsoft.com/en-us/learn/paths/responsible-ai-government-principles/) (for public ones).

Read [The Ten Guidelines for Product Leaders to Implement AI Responsibly](https://aka.ms/RAITenGuidelines) whitepaper designed in collaboration with Boston Consulting Group (BCG). This whitepaper aims at helping you prompt important conversations to have about how to translate your responsible AI principles to work, with relevant standards, practices, guidelines and guardrails.

One should note that this last guidance is distinct from Microsoft’s internal processes (shortly) introduced earlier but reflects perspectives from both organizations.This said, while we are all still navigating our own journey as you now understand, we hope that these resources will help you take a thoughtful step toward implementing your own responsible AI journey.

This concludes this introduction of Responsible AI, the Microsoft’s Responsible AI principles and how they translate to work, into practices company-wide through a related governance framework and a forthcoming standard, and how you can build your own journey.

Ultimately, we believe that every organization that creates or uses advanced (non-Generative vs. Generative) AI systems will need to develop and implement its own AI governance systems. To be effective, these systems must go beyond standards and principles and invest in real-world tools and practices to support teams throughout the AI development lifecycle.

# As a conclusion

This concludes this starter guide, part of the Responsible AI Workshop. We hope you have enjoyed this guided tour on what a responsible AI journey is through the lens of Microsoft own investments in this space, and on (some) of the RAI tooling available to help you put responsible AI to work.

AI innovation is occurring at a rapid pace today and our customers are accelerating adoption of AI. From holistically transforming industries to addressing critical issues facing humanity, AI is already solving some of our most complex challenges and redefining how humans and technology interact. Societal and environmental expectations are evolving, and governments are regulating AI in response.

Organizations should therefore consider adopting a responsible AI approach to harness the power of (non-Generative vs. Generative) AI in a way which mitigates risks and is compliant with the policies, laws, and regulations that apply.

Responsible AI is both is a culture and a practice to make sure that AI systems are safe, secure, trustworthy and uphold an organization’s AI principles, which are the starting point of the journey.

Eight years ago, Microsoft embarked on this journey, but we’ve only scratched the surface. Even if we moved from initial ideas and research to a point where we are now operationalizing responsible AI at scale, we are indeed only at the beginning of this journey towards putting responsible AI into practice, continuing to both invest and expand our safe and responsible AI practices, working alongside other industry leaders, advocating for global governance, etc.

In line with our [Customers AI Commitments](https://blogs.microsoft.com/blog/2023/06/08/announcing-microsofts-ai-customer-commitments/), Microsoft is helping our customers put responsible AI into practice by building custom mitigations into our AI products, sharing our learnings and perspective, and providing purpose-built tooling to support customers that want to build their own AI (non-Generative vs. Generative) AI systems responsibly.

Now, it is high time to get our hands a little bit dirty with Responsible AI practices and tools. Together with Microsoft, let’s put Responsible AI into practice!

# To go beyond

To continue learning about the passionate subject of Responsible AI, you can follow the other tutorials and walkthroughs available in this workshop.

Une image contenant texte, motif, point

Description générée automatiquementYou can also scan this code or visit <https://aka.ms/RAIresources> where you can access the entirety of already available tools, guidelines, and other additional resources that will help you create your next AI solution in a (more) responsible manner.

Une image contenant texte, capture d’écran, Site web, Page web

Description générée automatiquement

Une image contenant bleu, brouillard, capture d’écran, bleu vert

Description générée automatiquement

1. The report [The state of implementation of the OECD AI Principles](https://www.oecd.org/publications/the-state-of-implementation-of-the-oecd-ai-principles-four-years-on-835641c9-en.htm) four years on takes stock of initiatives launched by countries worldwide to implement the OECD AI Principles which were reported to the [OECD.AI Policy Observatory](https://oecd.ai/en/) as of May 2023. [↑](#footnote-ref-2)