Microsoft

Responsible AI Guiding principles

Abstract

In the last unit, we discussed the societal implications of AI and the responsibility of businesses, governments, NGOs, and academic researchers to anticipate and mitigate unintended consequences of AI technology. In light of this responsibility, organizations are finding the need to create internal policies and practices to guide their AI efforts, whether they are deploying third-party AI solutions or developing their own.

At Microsoft, we've recognized six principles that we believe should guide AI development and use: fairness, reliability and safety, privacy and security, inclusiveness, transparency, and accountability. For us, these principles are the cornerstone of a responsible and trustworthy approach to AI, especially as intelligent technology becomes more prevalent in the products and services we use every day.

Microsoft's six guiding principles

Fairness

AI systems should treat everyone fairly and avoid affecting similarly situated groups of people in different ways. 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.

We believe that mitigating bias starts with people understanding the implications and limitations of AI predictions and recommendations. Ultimately, people should supplement AI decisions with sound human judgment and be held accountable for consequential decisions that affect others.

When designing and building AI systems, developers should understand how bias can be introduced and how it can affect AI-based recommendations. To help mitigate bias, they should use training datasets that reflect the diversity of society. They should also design AI models in ways that allow them to learn and adapt over time without developing biases. To help them develop AI systems that treat everyone fairly, developers can leverage tools, methodologies, techniques, and other resources that help detect and mitigate biases.

Reliability and safety

To build trust, it's critical that AI systems operate reliably, safely, and consistently under normal circumstances and in unexpected conditions. These systems should be able to operate as they were originally designed, respond safely to unanticipated conditions, and resist harmful manipulation. It's also important to be able to verify that these systems are behaving as intended under actual operating 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 anticipate during design and testing.

We believe that rigorous testing is essential during system development and deployment to ensure AI systems can respond safely in unanticipated situations and edge cases, don't have unexpected performance failures, and don't evolve in ways that are inconsistent with original expectations. After testing and deployment, it's equally important that organizations properly operate, maintain, and protect their AI systems over the lifespan of their use. If not maintained properly, AI systems can become unreliable or inaccurate, so it's crucial to account for long-term operations and monitoring in every AI implementation. Ultimately, because AI should augment and amplify human capabilities, people need to play a critical role in making decisions about how and when an AI system is deployed, and whether it's appropriate to continue to use it over time. Human judgment will be key to identifying potential blind spots and biases in AI systems.

Privacy and security

As AI becomes more prevalent, protecting privacy and securing important personal and business information is becoming more critical and complex. With AI, privacy and data security issues require especially close attention because access to data is essential for AI systems to make accurate and informed predictions and decisions about people. AI systems must comply with privacy laws that require transparency about the collection, use, and storage of data and mandate that consumers have appropriate controls to choose how their data is used. At Microsoft, we are continuing to research privacy and security breakthroughs (see next unit) and invest in robust compliance processes to ensure that data collected and used by our AI systems is handled responsibly.

Inclusiveness

At Microsoft, we firmly believe everyone should benefit from intelligent technology, meaning it must incorporate and address a broad range of human needs and experiences. 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. Intelligent solutions such as real-time speech-to-text transcription, visual recognition services, and predictive text functionality are already empowering those with hearing, visual, and other impairments.

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.

Transparency

Underlying the preceding values are two foundational principles that are essential for ensuring the effectiveness of the rest: transparency and accountability. When AI systems are used to help inform decisions that have tremendous impacts on people's lives, it is critical that people understand how those decisions were made. For example, a bank might use an AI system to decide whether a person is creditworthy, or a company might use an AI system to determine the most qualified candidates to hire.

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.

 Accountability

The people who design and deploy AI systems must be accountable for how their systems operate. Organizations should draw upon industry standards to develop accountability norms. These norms can ensure that AI systems are not the final authority on any decision that impacts people's lives and that humans maintain meaningful control over otherwise highly autonomous AI systems.

Organizations should also consider establishing a dedicated internal review body. This body can provide oversight and guidance to the highest levels of the company on which practices should be adopted to help address the concerns discussed above and on particularly important questions regarding the development and deployment of AI systems. They can also help with tasks like defining best practices for documenting and testing AI systems during development or providing guidance when an AI system will be used in sensitive cases (like those that may deny people consequential services like healthcare or employment, create risk of physical or emotional harm, or infringe on human rights).

We recognize that every individual, company, and region will have their own beliefs and standards that should be reflected in their AI journey. We share our perspective with you as you consider developing your own guiding principles.

In the next unit you'll hear from Matt Fowler, VP and Head of Machine Learning, Enterprise Data and Analytics at TD Bank Group, as he describes how his business is approaching responsible AI.