Microsoft

# Implications of responsible AI Practical guide

## Abstract

AI has transformed from vision to reality, creating tangible benefits for people and enterprises around the world. But like any technology, it poses the risk of negative consequences when used improperly or irresponsibly. We want to share what we're learning in our own journey in hopes that it can provide a useful perspective for other organizations navigating similar challenges.

Learn about Microsoft's perspective on the importance of engaging with AI in a responsible manner.

## The transformative potential of AI

AI is the defining technology of our time. It is already enabling faster and more profound progress in nearly every field of human endeavor and helping to address some of society's most daunting challenges—like providing remote students with access to education and helping farmers produce enough food for our growing global population.

At Microsoft, we believe that the computational intelligence of AI should be used to amplify the innate creativity and ingenuity of humans. Our vision for AI is to empower every developer to innovate, empower organizations to transform industries, and empower people to transform society.

## Societal implications of AI

As with all great technological innovations in the past, the use of AI technology will have broad impacts on society, raising complex and challenging questions about the future we want to see. AI will have implications on decision-making across industries, data security and privacy, and the skills people need to succeed in the workplace. As we look to this future, we must ask ourselves: How do we design, build, and use AI systems that create a positive impact on individuals and society? How can we best prepare workers for the impact of AI? How can we attain the benefits of AI while respecting privacy?

## The importance of a responsible approach to AI

It's important to recognize that as new intelligent technology emerges and proliferates throughout society, with its benefits will come unintended and unforeseen consequences, some with significant ethical ramifications and the potential to cause serious harm. While organizations can't predict the future just yet, it's our responsibility to make a concerted effort to anticipate and mitigate the unintended consequences of the technology we release into the world through deliberate planning and continual oversight.

### Novel threats

We were reminded of this responsibility in 2016 when we released a chatbot on Twitter called Tay. We taught Tay to learn unsupervised from interactions with Twitter users, so she could better replicate human communication and personality traits. However, within 24 hours users realized that she could learn and began to feed her bigoted rhetoric, turning her from a polite bot into a vehicle for hate speech. This experience taught us that while technology may not be unethical on its own, people do not always have good intentions and we must consider the human element when designing AI systems. We learned to prepare for new types of attacks that influence learning datasets, especially for AI systems that have automatic learning capabilities. To help ensure a similar experience does not happen again, we developed technology such as advanced content filters and introduced supervisors for AI systems with automatic learning capabilities.

## Biased outcomes

Another unintended consequence that organizations should keep in mind is that AI may reinforce societal or other biases without deliberate planning and design. For example, Microsoft partnered with a large financial lending institution to develop a risk scoring system for loan approvals. We trained an existing industry model using the customer's data. When we conducted an audit of the system, we discovered that while it only approved low-risk loans, all approved loans were for male borrowers. The training data reflected the fact that loan officers historically favor male borrowers—and inspecting the system allowed us to identify and address that bias before the system was deployed. It's important for developers to understand how bias can be introduced into either training data or machine learning models. At Microsoft, our researchers are exploring tools and techniques for detecting and reducing bias within AI systems. Explore more about this and more in the summary and resources unit of this module.

## Sensitive use cases

Another illustration of our responsibility to mitigate unintended consequences is with sensitive technologies like facial recognition. Recently, there has been a growing demand for facial recognition technology, especially from law enforcement organizations that see the potential of the technology for use cases like finding missing children. However, we recognize that these technologies could potentially be used by a government to put fundamental freedoms at risk by, for example, enabling continuous surveillance of specific individuals. We believe society has a responsibility to set appropriate boundaries for the use of these technologies, which includes ensuring governmental use of facial recognition technology remains subject to the rule of law.

While we believe that new laws and regulations are indispensable, we also recognize that they are not a substitute for the responsibility that needs to be exercised by businesses, governments, NGOs, and academic researchers engaging with AI. This is why, in July 2018, we announced that we would assess and develop principles to govern our work with facial recognition technologies. We anticipate these principles will evolve over time as we continue to learn and partner with customers, other tech companies, academics, civil society, and others on this issue. Review them in the summary and resources unit of this module.

Facial recognition technology highlights the importance of preparing for and remaining vigilant of shortcomings and unintended consequences with all emerging AI. We consider it a shared responsibility across the public and private sector to engage with AI responsibly. It's essential that we continue to foster open dialogue among businesses, governments, NGOs, academic researchers, and all other interested individuals and organizations.

## Applying these ideas in your organization

The following three questions can help you start to consider the ways your organization can develop and deploy AI in a responsible manner.

1. How can you use a human-led approach to drive value for your business?
2. How will your organization’s foundational values affect your approach to AI?
3. How will you monitor AI systems to ensure they are evolving responsibly?

## Coming up

In the next two units, we have outlined some of the steps Microsoft and TD Bank are taking to prioritize responsible AI in hopes that our experience can help as you consider these questions for your own organization. The next unit will outline six guiding principles we developed to guide our development and use of AI.