100 XP

Understand considerations for responsible AI

3 minutes

The previous unit introduced the need for considerations for responsible and ethical development of AI-enabled software. In this unit, we'll discuss some core principles for responsible AI that have been adopted at Microsoft.

Fairness

AI systems should treat all people fairly. For example, suppose you create a machine learning model to support a loan approval application for a bank. The model should make predictions of whether or not the loan should be approved without incorporating any bias based on gender, ethnicity, or other factors that might result in an unfair advantage or disadvantage to specific groups of applicants.

Fairness of machine learned systems is a highly active area of ongoing research, and some software solutions exist for evaluating, quantifying, and mitigating unfairness in machine learned models. However, tooling alone isn't sufficient to ensure fairness. Consider fairness from the beginning of the application development process; carefully reviewing training data to ensure it's representative of all potentially affected subjects, and evaluating predictive performance for subsections of your user population throughout the development lifecycle.

Reliability and safety

AI systems should perform reliably and safely. For example, consider an AI-based software system for an autonomous vehicle; or a machine learning model that diagnoses patient symptoms and recommends prescriptions. Unreliability in these kinds of system can result in substantial risk to human life.

As with any software, AI-based software application development must be subjected to rigorous testing and deployment management processes to ensure that they work as expected before release. Additionally, software engineers need to take into account the probabilistic nature of machine learning models, and apply appropriate thresholds when evaluating confidence scores for predictions.

Privacy and security

AI systems should be secure and respect privacy. The machine learning models on which AI systems are based rely on large volumes of data, which may contain personal details that must be kept private. Even after models are trained and the system is in production, they use new data to make predictions or take action that may be subject to privacy or security concerns; so appropriate safeguards to protect data and customer content must be implemented.

Inclusiveness

AI systems should empower everyone and engage people. AI should bring benefits to all parts of society, regardless of physical ability, gender, sexual orientation, ethnicity, or other factors.

One way to optimize for inclusiveness is to ensure that the design, development, and testing of your application includes input from as diverse a group of people as possible.

Transparency

AI systems should be understandable. Users should be made fully aware of the purpose of the system, how it works, and what limitations may be expected.

For example, when an AI system is based on a machine learning model, you should generally make users aware of factors that may affect the accuracy of its predictions, such as the number

of cases used to train the model, or the specific features that have the most influence over its predictions. You should also share information about the confidence score for predictions.

When an AI application relies on personal data, such as a facial recognition system that takes images of people to recognize them; you should make it clear to the user how their data is used and retained, and who has access to it.

Accountability

People should be accountable for AI systems. Although many AI systems seem to operate autonomously, ultimately it's the responsibility of the developers who trained and validated the models they use, and defined the logic that bases decisions on model predictions to ensure that the overall system meets responsibility requirements. To help meet this goal, designers and developers of AI-based solution should work within a framework of governance and organizational principles that ensure the solution meets ethical and legal standards that are clearly defined.

> **Note**
>
> Microsoft has released meaningful updates to the Responsible AI Standard in June 2022. As part of that, we've updated the approach to facial recognition including a new Limited Access policy for certain features as a safeguard for responsible use. You can apply for that limited access to enable those features for your application.

For more information about Microsoft's principles for responsible AI, visit the Microsoft responsible AI site .

Next unit: Understand capabilities of Azure Machine Learning

Continue >