**Responsible AI Principles**

Before beginning the technical steps, it's essential to understand Microsoft's Responsible AI Principles, an ethical framework designed to guide the responsible development, deployment, and operation of AI systems. These principles guide the responsible design, development, and deployment of AI systems, ensuring that AI technologies are built in a way that is fair, transparent, and inclusive. These principles are the foundation for evaluating the safety of AI models.

Microsoft's Responsible AI Principles include:

* **Fairness and Inclusiveness**: 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 who has similar symptoms, financial circumstances, or professional qualifications.
* **Reliability and Safety**: To build trust, it's critical that AI systems operate reliably, safely, and consistently. These systems should be able to operate as they were originally designed, respond safely to unanticipated conditions, and resist harmful manipulation. How they behave and the variety of conditions they can handle reflect the range of situations and circumstances that developers anticipated during design and testing.
* **Transparency**: When AI systems help inform decisions that have tremendous impacts on people's lives, it's 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. A company might use an AI system to determine the most qualified candidates to hire.
* **Privacy and Security**: As AI becomes more prevalent, protecting privacy and securing personal and business information are becoming more important and complex. With AI, privacy and data security require close attention because access to data is essential for AI systems to make accurate and informed predictions and decisions about people.
* **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 aren't the final authority on any decision that affects people's lives. They can also ensure that humans maintain meaningful control over otherwise highly autonomous AI systems.

