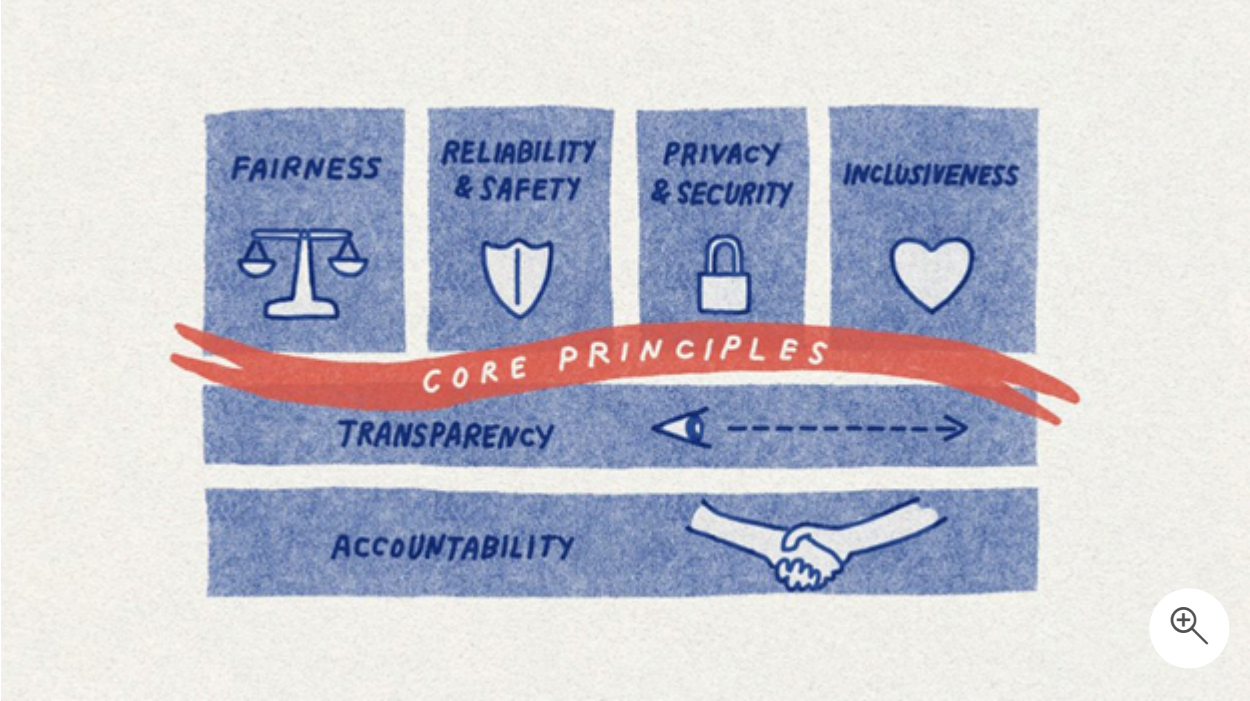
100 XP

Microsoft and GitHub's six principles of responsible AI

3 minutes



Microsoft is a global leader in Responsible AI, identifying six key principles that should guide AI development and usage. These principles are:

- **Fairness**: AI systems should treat all people fairly.
- **Reliability and safety**: AI systems should perform reliably and safely.
- **Privacy and security**: AI systems should be secure and respect privacy.
- **Inclusiveness**: AI systems should empower everyone and engage people.
- **Transparency**: AI systems should be understandable.
- **Accountability**: People should be accountable for AI systems.

Now, let's explore each of these principles in greater detail to understand how they guide responsible AI practices.

Fairness: AI systems should treat all people fairly

AI systems should treat everyone fairly, avoiding differential impacts on similarly situated groups. In contexts like medical treatment, loan applications, or employment, AI should

provide consistent recommendations to individuals with similar symptoms, financial situations, or qualifications.

Employs techniques to detect bias and mitigate unfair impacts such as:

- Reviewing training data.
- Testing models with balanced demographic samples.
- Using adversarial debiasing.
- Monitoring model performance across user segments.
- Implementing controls to override unfair model scores.

Training AI models on diverse and balanced data can help reduce biases, ultimately promoting fairness.

Reliability and safety: AI systems should perform reliably and safely

To build trust, AI systems must operate reliably, safely, and consistently.

These systems need to function as designed, respond safely to unexpected conditions, and resist harmful manipulation. Their behavior and the variety of conditions they handle reflect the foresight of developers during design and testing.

Safety in AI refers to minimizing unintended harm, including physical, emotional, and financial harm to individuals and societies. Reliability means that AI systems perform consistently as intended without unwanted variability or errors. Safe and reliable systems are robust, accurate, and behave predictably under normal conditions.

Privacy and security: AI systems should be secure and respect privacy

As artificial intelligence (AI) becomes more common, it's important to protect user privacy and data security. Microsoft and GitHub are aware of this tenet and both companies include privacy and security as key parts of their Responsible AI plan. This plan focuses on using principles to guide data practices.

Microsoft and GitHub’s approach to Responsible AI aims to stop abuse and keep user trust. Key points include:

- Getting users' permission before collecting their data. Clearly explain how the AI uses their data and get their consent. Don't collect data secretly. Let users choose if they want to share personal data and inform them through clear prompts and policies.

- Collecting only the data needed for the AI to work. Avoid gathering extra information and remove sensitive data once the AI is in use. Regularly check data inputs to ensure only essential data is collected.
- Anonymizing personal data. Use methods like pseudonymization and aggregation to protect identities. Pseudonymization replaces personal details with random identifiers, while aggregation groups data into summaries, removing specific individual details.

Encrypt sensitive data both during transfer and when stored. Use strong encryption methods and secure keys through:

- Hardware Security Modules (HSMs) which store keys in a tamper-proof environment.
- Secure vaults like Microsoft Azure for key storage with controlled access.
- Envelope encryption, which uses two keys for added security.
- Organizations should control who can access keys and models, rotate keys regularly, and securely back up keys. They should also limit employee access to sensitive models and data, classify them based on sensitivity, and conduct regular security audits to prevent unauthorized access.

Inclusiveness: AI systems should empower everyone and engage people

Inclusiveness means ensuring that AI systems are fair, accessible, and empower everyone. Microsoft's Responsible AI standard recognizes that AI creators (including GitHub) must proactively design AI to include all people, communities, and geographies - especially those areas of society historically underrepresented.

Microsoft's Responsible AI standard for inclusiveness means:

- AI systems work well for diverse users and groups. They don't disadvantage some people.
- AI systems are accessible. Anyone can use AI systems easily, regardless of physical or mental abilities.
- AI systems are available worldwide, even in developing countries/regions. AI systems can't exclude certain geographies.
- People from different backgrounds and communities provide input into the development of AI systems.
- AI systems allow all users to benefit equally from their capabilities. They must empower everyone.

Examples of inclusive AI include:

- Facial recognition that works across skin tones, ages, and genders.
- Interfaces that support screen readers for the visually impaired.

- Language translation that supports small regional dialects.
- Teams that seek diverse perspectives when designing systems.

Microsoft's Responsible AI standard requires that everyone can access AI systems, regardless of their disability, language, or infrastructure barriers. Responsible AI solutions must enable full global inclusion by:

- Offering alternative modes of interaction such as voice control, captions, and screen readers.
- Supporting adaptation into different languages and local cultural contexts.
- Working offline and with limited connectivity and computing resources.

Transparency: AI systems should be understandable

Microsoft's Responsible AI principle of Transparency emphasizes that AI systems must be understandable and interpretable. AI creators should:

- Explain how their systems operate clearly through a clear validation framework.
- Justify the design choices behind AI systems.
- Be honest about the capabilities and limitations of AI systems.
- Enable auditability with logging, reporting, and auditing capabilities.

Transparency is essential to build trust, ensure accountability, promote fairness, enhance safety, and support inclusiveness. Implementing transparency involves documenting data and models, creating explanatory interfaces, using AI debugging tools, constructing testing dashboards, and enabling logging and auditing. By being transparent, AI creators can foster trust and responsible AI use.

Accountability: People should be accountable for AI systems

The Accountability principle states that AI creators should be responsible for how their systems operate. They need to continuously monitor system performance and mitigate risks. Accountability in the AI industry is becoming a pressing issue as high-profile cases of algorithmic harm, bias, and abuse come to light. Critics increasingly argue that without accountability, AI creators hold too much power over opaque systems impacting lives.

Microsoft emphasizes accountability in AI development and deployment through its Responsible AI Standard, which considers accountability a foundational principle. According to

Microsoft, AI systems must be accountable to people, and companies deploying AI systems must take responsibility for their operation.

Now let's do a knowledge check to review what we just learned.

Next unit: Module assessment

Previous Next