# Principle 3: Transparency and explainability

Let's see how this principle guides responsible AI practice.  We'll look at a key risk this principle is designed to address and explore a brief practical example of how this principle relates to an AI use case.

**Definition:** “The Public Service needs to commit to transparency in its use of AI. People interacting with government AI systems or receiving AI-assisted services should be aware of and understand how AI is being used.

To support this, agencies should publicly disclose:

* when AI systems are used
* how they were developed
* how they affect outcomes — as relevant and appropriate according to the given use case.

Agencies should also enable people affected by the outcome of an AI system to understand how the outcome was determined.”

**Let’s focus on: Interpretability:** AI systems can feel like a veritable “black box” where the decision-making process is not easily interpretable by a layperson. The lack of transparency can make it challenging to audit AI systems for fairness, accountability and errors.

* To be transparent means to make the inner workings and decision-making processes of AI systems (and their data) understandable and accessible to relevant stakeholders.
* To be explainable means the models should provide clear, interpretable explanations for the decisions where possible, aligning with precepts of open data.

**Practical example:** A public university is using AI to determine student admissions. If faculty members see AI as a “black box”, they may raise questions over fairness and potential students may question accountability. The AI decision-making process should be clear to all parties involved.

To help ensure explainability, design choices can favour a deterministic system over a non-deterministic system:

* Deterministic systems: Always produce the same output for the same input.

Example: Student eligibility checker leveraging pre-determined rules.

* Non-deterministic (generative) systems: Can produce different outputs, even with the same input.

Example: A summary of application documents that may produce a different output based on a slight variance in the prompt.

When it comes to explainability, the AI system should provide clear reasons for each admissions decision it has commented on, so that educators can trust the process and ensure it aligns with equity and other standards. The selection of a deterministic approach helps enable this – while non-deterministic (generative) systems are not inherently unexplainable, interpreting their decisions requires a higher level of AI Governance maturity.

So what: Key takeaways for evaluating use cases.

**Does the use case... ·**

* Commit to transparency by publicly disclosing AI usage and development processes?
* Ensure stakeholders understand how AI affects outcomes and decisions?
* Make AI systems interpretable and explainable for fairness and accountability?
* Provide clear explanations for AI-driven decisions to maintain trust and equity standards?