

Responsible AI applies to the entire system

Responsible AI guides how organizations build and operate AI—not only how they choose a model.

It guides how organizations

- Design AI solutions
- Select and prepare data
- Evaluate system behavior
- Deploy and monitor applications
- Respond when harm occurs

Key message

A capable model can still produce an irresponsible system.

The six principles

Principle	Central question
Fairness	Does the system treat comparable people consistently?
Reliability and safety	Does it operate safely under expected and unexpected conditions?
Privacy and security	Does it protect people, data, models, and credentials?
Inclusiveness	Can people with different needs and abilities use it?
Transparency	Do people understand the system’s purpose and limitations?
Accountability	Who remains responsible for its decisions and effects?

Exam strategy

Several principles can apply. Select the principle that best matches the question’s primary risk.

Fairness: comparable cases deserve comparable treatment

Ask whether differences have a valid, appropriate basis—not whether every outcome is identical.

Scenario • “A résumé-screening system advances equally qualified applicants at different rates based on gender.”

AI systems should

- Avoid unjustified differences between groups
- Measure performance across relevant populations
- Use sufficiently representative data
- Investigate unequal error rates or outcomes

Primary principle: Fairness

Exam cue

Disparate treatment, demographic bias, or unequal outcomes usually signal fairness.

Reliability and safety: predictable behavior within defined limits

Reliability is whether the system performs as designed. Safety is the harm when it fails.

Scenario · “A medical assistant gives confident treatment advice when the available evidence is incomplete.”

AI systems should

- Perform consistently in intended conditions
- Handle unusual inputs safely
- Fail without causing unacceptable harm
- Resist manipulation and misuse
- Support testing, monitoring, and recovery

Primary principle: Reliability and safety

Exam cue

Unsafe failures, inconsistent performance, adversarial inputs, or inadequate testing point here.

Privacy and security: protect data and the AI system

Privacy governs appropriate data use. Security protects data and systems from unauthorized access.

Scenario · “A support assistant includes another customer’s account details in its response.”

Protect

- Personal and confidential information
- Stored data and transmitted data
- Model endpoints, credentials, and connected tools
- User consent and control over data
- Systems against unauthorized access or data leakage

Primary principle: Privacy and security

Exam cue

Personal data, permissions, leakage, credentials, or unauthorized access usually indicate privacy and security.

Inclusiveness: design for the full range of users

Inclusiveness asks who the system unintentionally excludes.

Scenario · “A customer assistant supports typed input but offers no practical option for users who cannot use a keyboard.”

Inclusive AI should

- Consider people with different abilities and backgrounds
- Support alternative interaction methods
- Avoid unnecessary barriers
- Include affected communities in design and testing

Primary principle: Inclusiveness

Exam cue

Accessibility, accommodation, or excluded users point to inclusiveness.

Transparency: help people understand the system

Users need disclosure, limitations, and a way to question or appeal an outcome.

Scenario · “An assistant presents an AI-generated eligibility decision as a final human judgment.”

Users should know

- That they are interacting with AI
- What the system can and cannot do
- What information influenced an output
- When uncertainty or limitations matter
- How to question or appeal an outcome

Primary principle: Transparency

Exam cue

Explanations, disclosure, traceability, or communicated limitations usually indicate transparency.

Accountability: humans and organizations retain responsibility

Responsibility cannot be delegated to the model.

Scenario · “An autonomous agent issues customer refunds, but no team owns its policies or reviews questionable actions.”

Accountable organizations

- Assign clear ownership
- Define approval and escalation paths
- Review high-impact decisions
- Audit system behavior
- Correct failures and provide recourse

Primary principle: Accountability

Exam cue

Ownership, oversight, governance, auditing, or responsibility point to accountability.

One failure can involve several principles

“A voice assistant performs poorly for some accents and gives no warning about its limitations.”

Problem	Principle
Unequal recognition performance	Fairness
Some users cannot complete the task	Inclusiveness
Limitations remain undisclosed	Transparency
No process exists for reviewing failures	Accountability

Exam strategy

Identify the specific defect in the question, then choose the principle that matches that primary risk.

Match the scenario to the principle

Scenario	Principle
A model reveals confidential records.	Privacy and security
A hiring model disadvantages one demographic group.	Fairness
A chatbot conceals that its answers come from AI.	Transparency
A vision application cannot be used by people with impaired vision.	Inclusiveness
An automated decision has no assigned human owner.	Accountability
A system behaves dangerously when given an unexpected input.	Reliability and safety

Memory hooks

1. Fairness → Comparable treatment
2. Reliability and safety → Dependable, safe behavior
3. Privacy and security → Protected data and systems
4. Inclusiveness → Access and participation
5. Transparency → Understanding and disclosure
6. Accountability → Human ownership

Ask first

Have learners answer before revealing the mappings. Choose the principle that matches the primary risk.