

Scenario drill: principle and mitigation

For each scenario, identify the primary principle and the strongest mitigation.

Available principles

- Fairness
- Reliability and safety
- Privacy and security
- Inclusiveness
- Transparency
- Accountability

Key message

Choose the primary principle, even when several apply. Focus on the mitigation that most directly addresses the stated failure.

Scenarios 1 and 2

1. Biased loan recommendations

A loan recommendation system approves qualified applicants from one demographic group less often than similarly qualified applicants from another group.

Ask

- Which principle applies most directly?
- What should the organization do first?

Reveal

Fairness

Evaluate approval and error rates across relevant demographic groups, then address unjustified performance differences.

2. Unexplained medical output

A medical assistant recommends urgent treatment but provides no supporting evidence, confidence information, or explanation.

Ask

- Which principle applies most directly?
- What mitigation would help the clinician evaluate the recommendation?

Reveal

Transparency

Provide supporting evidence, communicate uncertainty, and explain the system's limitations.

Scenarios 3 and 4

3. Exposure of personal information

A support assistant includes information from another customer's account in its response.

Ask

- Which principle applies most directly?
- What control should prevent this?

Reveal

Privacy and security

Enforce user-level authorization and isolate retrieved customer data between sessions.

4. Inaccessible speech-only interface

A public-service assistant accepts and returns only spoken language. Users who cannot hear or speak have no alternative.

Ask

- Which principle applies most directly?
- What should the application provide?

Reveal

Inclusiveness

Add accessible text input and output, captions, and assistive-technology support.

Scenarios 5 and 6

5. Unsafe autonomous action

An AI agent interprets “reduce cloud costs” as permission to delete production resources without approval.

Ask

- Which principle applies most directly?
- What control should constrain the agent?

Reveal

Reliability and safety

Limit the agent’s permissions and require human approval before destructive or high-impact actions.

6. No human escalation path

A customer disputes an AI-generated refund denial, but the application offers no review process and no accountable decision owner.

Ask

- Which principle applies most directly?
- What should the organization establish?

Reveal

Accountability

Assign a human owner and provide a review, appeal, and escalation process.

Scenarios 7 and 8

7. Hidden AI involvement

Customers believe they are chatting with a human support representative because the application never discloses that it generates responses with AI.

Ask

- Which principle applies most directly?
- What should change in the user experience?

Reveal

Transparency

Clearly disclose AI involvement before users rely on the assistant's responses.

8. Unequal speech recognition

A voice assistant performs well for most users but frequently misinterprets customers with particular regional accents.

Ask

- Which principle applies most directly?
- What should the team do?

Reveal

Fairness

Test recognition performance by accent, improve representative data, and provide an alternative input method.

Scenarios 9 and 10

9. Confident response during an outage

The assistant cannot reach the order database but tells the customer that an order has shipped.

Ask

- Which principle applies most directly?
- What should happen when the data source is unavailable?

Reveal

Reliability and safety

Detect the dependency failure, avoid unsupported claims, and degrade to a clear status message or human escalation.

10. Sensitive data retained without need

The assistant permanently stores complete customer conversations, including payment and health information, even though the service no longer needs them.

Ask

- Which principle applies most directly?
- What data practice should change?

Reveal

Privacy and security

Minimize collected data and enforce purpose-based retention and deletion policies.

Answer key

Scenario	Primary principle	Best mitigation
1. Biased loans	Fairness	Test and correct group-level disparities
2. Unexplained medical output	Transparency	Explain evidence, uncertainty, and limits
3. Customer-data exposure	Privacy and security	Enforce authorization and data isolation
4. Speech-only interface	Inclusiveness	Provide accessible alternatives
5. Unsafe autonomous action	Reliability and safety	Constrain permissions and require approval
6. No escalation path	Accountability	Assign ownership and provide review
7. Hidden AI involvement	Transparency	Disclose AI use
8. Unequal speech recognition	Fairness	Evaluate cohorts and improve representation
9. Database outage	Reliability and safety	Fail safely and degrade predictably
10. Unnecessary retention	Privacy and security	Minimize and delete data

Use after the drill

Confirm each pair, then keep the key visible while you teach the decision method.

How to solve responsible-AI questions

Question focus	Primary principle
Unequal treatment or performance	Fairness
Unsafe or unreliable behavior	Reliability and safety
Data exposure or unauthorized access	Privacy and security
Barriers to participation	Inclusiveness
Missing disclosure or explanation	Transparency
Missing ownership, review, or recourse	Accountability

Then mitigate

Choose the mitigation that directly removes or controls that failure.