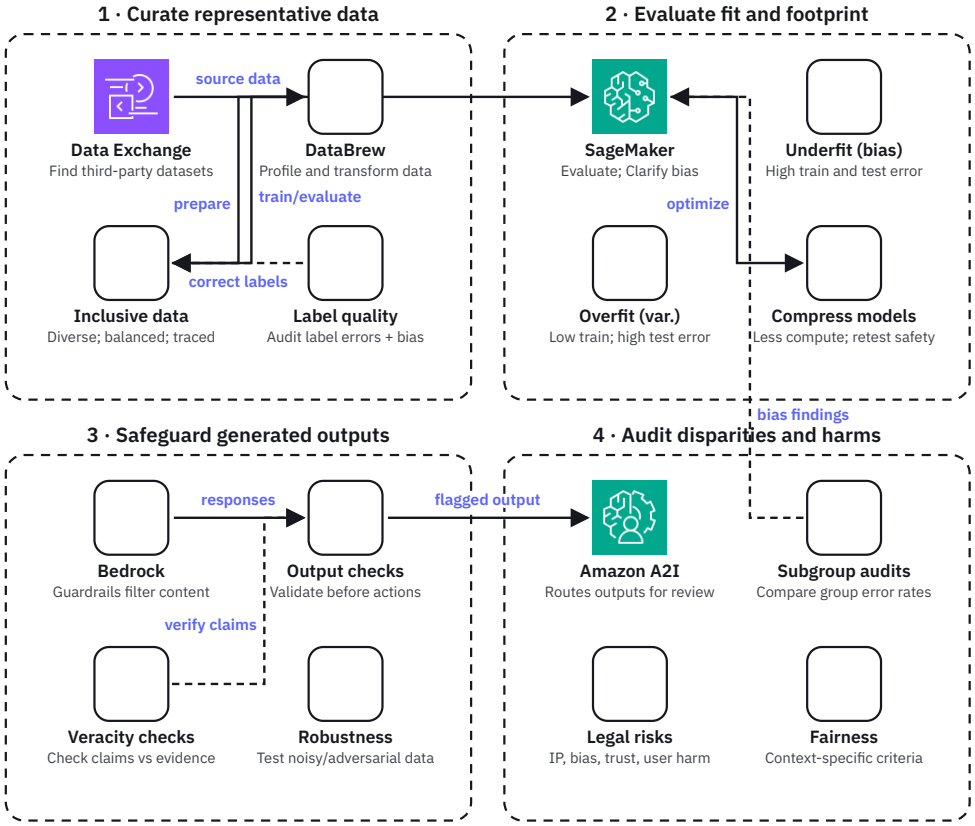
DOMAIN 4 · GUIDELINES FOR RESPONSIBLE AI · 14% OF THE EXAM · TASK STATEMENT 4.1

Explain the development of AI systems that are responsible

Curate, evaluate, safeguard, audit



1 Curate representative data

Choose **representative data** covering minority groups and rare cases, with curated sources, provenance, and permissions. Check class imbalance; DataBrew preparation and Data Exchange access do not guarantee fairness.

2 Evaluate fit and footprint

Select through **model evaluation**: compare quality, fairness, safety, latency, and environmental impact. Quantization, pruning, and distillation reduce resources; recheck accuracy, safety, and energy/water impacts.

3 Safeguard generated outputs

Apply **output validation** at application boundaries: check schema, permissions, sensitive data, and unsafe content. Configure Guardrails for filtering and grounding checks; safeguards do not establish factual truth.

4 Audit disparities and harms

Use **human review** for ambiguous or high-risk outputs; Amazon A2I requires configured review workflows. Audit labels and subgroups over time. Review licenses and outputs to reduce infringement risk and customer harm.

Exam traps

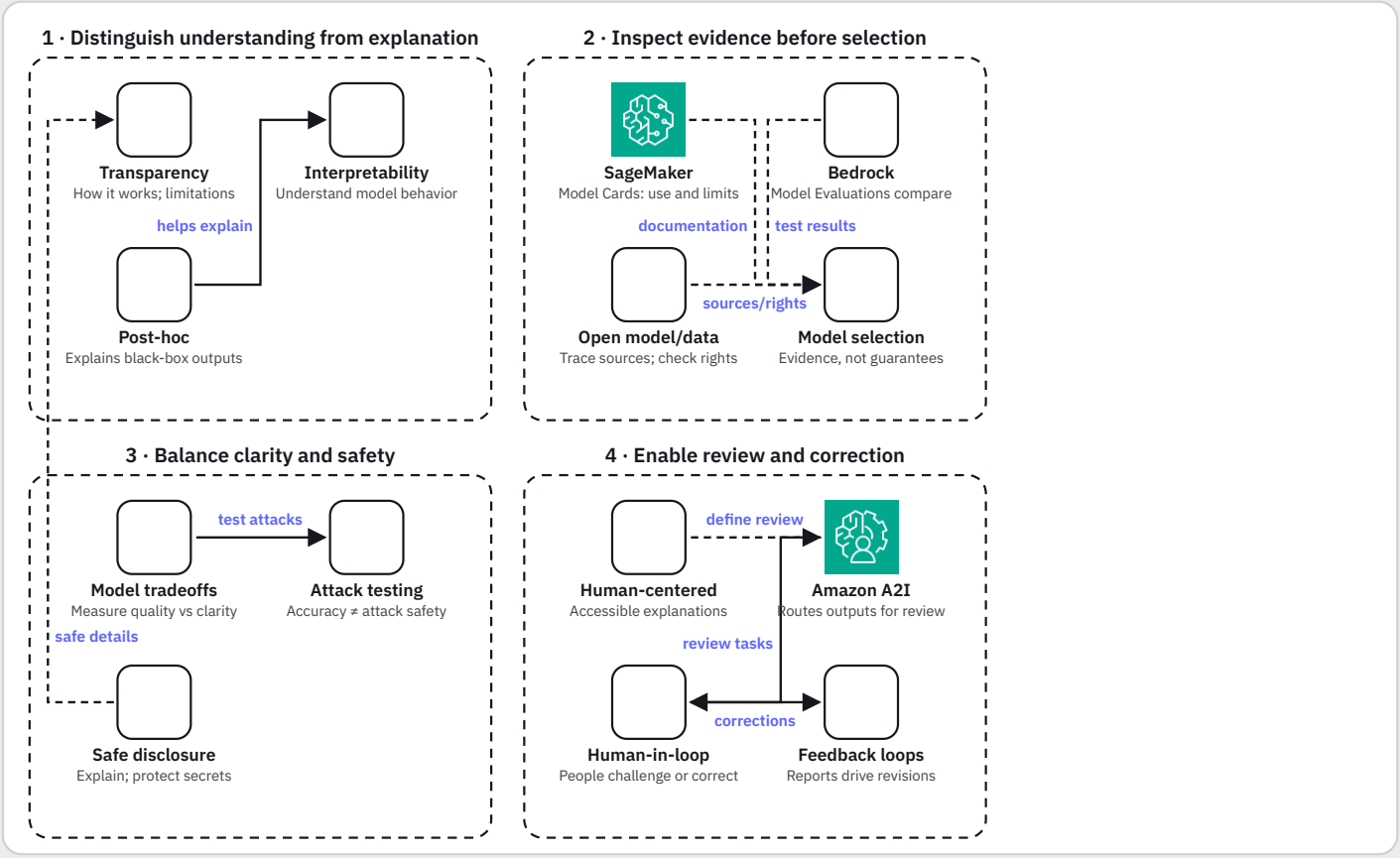
- **High overall accuracy** does not prove fairness: minority-group errors and rare-class failures can remain hidden. Statistical bias in underfitting is distinct from discriminatory algorithmic bias.
- **Guardrails do not eliminate hallucinations**. Plausible answers and citations can be fabricated; grounding checks, evidence verification, and human review reduce rather than eliminate risk.
- **Managed model access is not IP clearance**. Training data, prompts, retrieved documents, and generated outputs can carry rights restrictions; provider access does not remove infringement risk.

Get practice exams and lessons for AWS Certified AI Practitioner

DOMAIN 4 · GUIDELINES FOR RESPONSIBLE AI · TASK STATEMENT 4.2

Recognize the importance of transparent and explainable models

Distinguish, inspect, balance, involve



1 Distinguish understanding from explanation

Choose **post-hoc explainability** to explain an opaque model after training. Transparent models expose how they work; post-hoc explanations can approximate behavior without making internals transparent.

2 Inspect evidence before selection

Choose **documented evidence**: inspect available code/weights, data provenance, and license entitlements. Model Cards document intended use and limits; Bedrock evaluations compare models, not prove transparency.

3 Balance clarity and safety

Use **selective disclosure** when details expose sensitive data, intellectual property, or attack paths. Measure interpretability, performance, and adversarial robustness; greater complexity is not automatically better.

4 Enable review and correction

Choose **human review** for uncertain or consequential predictions. A2I supports configured review workflows; people supply judgment. Feedback improves models or explanations only when evaluated and acted on.

Exam traps

- **Explainable does not mean correct or fair.** A plausible post-hoc explanation is not proof of accuracy, fairness, or causality; explanations themselves require evaluation.
- **Open-source does not mean unrestricted or fully transparent.** Available code or weights do not guarantee disclosed training data, clear provenance, or permission for every use.
- **Human review is not automatic retraining.** A2I routes configured review tasks; corrections require evaluation and an explicit update process before model behavior changes.

Get practice exams and lessons for AWS Certified AI Practitioner

DOMAIN 4 · GUIDELINES FOR RESPONSIBLE AI · WHAT TO REMEMBER

Concepts and services of Guidelines for Responsible AI

Concepts to know

Algorithmic Fairness

Fairness addresses harmful disparities across populations. **Compare subgroup outcomes**, not just overall accuracy; choose representative data, appropriate fairness metrics, and bias checks throughout the AI lifecycle.

Model Transparency

Transparency provides visibility into model design, data, and limitations; explainability addresses particular outputs. **Match disclosure to stakeholder needs** without exposing sensitive information or enabling misuse.

Human-in-the-Loop

Human review adds contextual judgment to AI decisions. **Route high-risk or uncertain outputs for review** and provide meaningful opportunities to correct or challenge decisions; automated safeguards do not replace accountability.

Core services

Amazon SageMaker

Choose **Clarify for bias analysis and prediction explanations**; use Model Cards to document intended use, evaluation results, and limitations.

Amazon Bedrock

Choose **Guardrails for configurable safety controls** and model evaluations to compare foundation models. Guardrails filter configured risks but do not guarantee truthful or harmless outputs.

Amazon Augmented AI (Amazon A2I)

Choose **human review workflows** when predictions need human judgment, such as uncertain or consequential results; reviewer decisions support correction and oversight.

Get practice exams and lessons for AWS Certified AI Practitioner