📘 MCQs: Guardrails – Purpose, Timing & Tripwires in Fundamentals of Agentic AI

### 1. What is the ****main purpose of guardrails**** in Agentic AI?

A) To make responses faster  
B) To ensure safety, compliance, and control in agent behavior  
C) To reduce token usage  
D) To improve creativity

**Answer:** B  
**Explanation:** Guardrails are designed to **keep AI safe, reliable, and aligned** with rules.

### 2. Which of the following best describes ****guardrails****?

A) Static instructions only  
B) Runtime policies that enforce safety and compliance  
C) Token sampling parameters  
D) Error logs

**Answer:** B  
**Explanation:** Guardrails are **active policies** applied at runtime.

### 3. When are guardrails typically applied?

A) Before execution  
B) During execution  
C) After execution  
D) All of the above

**Answer:** D  
**Explanation:** Guardrails can apply **pre-execution (input filtering)**, **during execution (tripwires)**, and **post-execution (output filtering)**.

### 4. What is a ****tripwire**** in guardrails?

A) A token counter  
B) A mechanism that detects violations and stops unsafe behavior immediately  
C) A retry function  
D) A logging system

**Answer:** B  
**Explanation:** Tripwires **intercept unsafe content** before it continues.

### 5. Example of a ****tripwire****?

A) Blocking sensitive data exposure like credit card numbers  
B) Changing temperature  
C) Logging inputs  
D) Resetting context

**Answer:** A  
**Explanation:** Tripwires detect and **block unsafe or restricted patterns**.

### 6. Which type of guardrail ensures that ****inputs**** are safe before execution?

A) Output guardrails  
B) Pre-execution guardrails  
C) Context guards  
D) Retry handlers

**Answer:** B  
**Explanation:** Pre-execution guardrails validate **user input safety**.

### 7. Which type of guardrail ensures that ****outputs**** are compliant?

A) Post-execution guardrails  
B) Pre-execution guardrails  
C) Tripwires only  
D) Token guards

**Answer:** A  
**Explanation:** Post-execution guardrails filter **unsafe responses** before returning to the user.

### 8. Why are ****timing layers**** important in guardrails?

A) They decide when to apply rules for maximum effectiveness  
B) They replace system messages  
C) They reduce memory usage  
D) They block retries

**Answer:** A  
**Explanation:** Proper timing ensures **guardrails catch issues at the right stage**.

### 9. Which of these is an example of a ****purpose-driven guardrail****?

A) Ensuring model never outputs personally identifiable information (PII)  
B) Increasing randomness  
C) Reducing latency  
D) Disabling context

**Answer:** A  
**Explanation:** Purpose is about **safety & compliance goals**, like preventing PII leaks.

### 10. What happens if a ****tripwire is triggered****?

A) Execution continues normally  
B) The agent halts, retries, or routes to a fallback  
C) Logs are ignored  
D) Tokens are reduced

**Answer:** B  
**Explanation:** Tripwires **stop unsafe flow** and apply **fallbacks or re-routing**.

### 11. Which of the following is a ****limitation**** of guardrails?

A) They cannot validate structured data  
B) They may block safe inputs if rules are too strict  
C) They always slow down agents drastically  
D) They disable retries

**Answer:** B  
**Explanation:** Overly strict guardrails can **cause false positives**.

### 12. How do guardrails interact with ****dynamic instructions****?

A) They override them if unsafe  
B) They always ignore them  
C) They merge automatically  
D) They reduce creativity

**Answer:** A  
**Explanation:** Guardrails **override unsafe dynamic instructions** to protect the system.

### 13. Why are ****tripwires lightweight****?

A) To minimize inference delay while still detecting violations  
B) To reduce token counts  
C) To disable retries  
D) To log faster

**Answer:** A  
**Explanation:** Tripwires are **light checks** that run **quickly** to avoid slowing execution.

### 14. Which best describes a ****layered guardrail system****?

A) One guardrail for all cases  
B) Multiple checks at input, reasoning, and output stages  
C) Randomized blocking  
D) Context-only validation

**Answer:** B  
**Explanation:** A layered approach applies guardrails **across all stages**.

### 15. What is a ****practical use case**** of post-execution guardrails?

A) Ensuring AI responses never output offensive language  
B) Preventing API timeouts  
C) Reducing token usage  
D) Increasing temperature

**Answer:** A  
**Explanation:** Post-execution guardrails filter **unsafe final outputs**.

### 16. What is the difference between ****guardrails and error handling****?

A) Guardrails focus on safety/compliance, error handling manages failures  
B) Guardrails log errors only  
C) Guardrails retry automatically  
D) They are the same

**Answer:** A  
**Explanation:** Guardrails = **safety**, Error handling = **failure recovery**.

### 17. How do guardrails relate to ****trust**** in AI systems?

A) They reduce trust  
B) They build user trust by preventing unsafe or biased outputs  
C) They disable logging  
D) They slow execution

**Answer:** B  
**Explanation:** Guardrails ensure **trustworthiness** of agents.

### 18. Which type of guardrail protects ****system messages**** from injection attacks?

A) Input sanitization tripwires  
B) Post-execution validators  
C) Retry loops  
D) Temperature guards

**Answer:** A  
**Explanation:** Input sanitization stops **malicious prompt injection**.

### 19. What is the role of ****fallbacks**** when guardrails block execution?

A) To retry with adjusted safe responses  
B) To ignore the error  
C) To log only  
D) To reset the agent

**Answer:** A  
**Explanation:** Fallbacks ensure **user still gets a safe, meaningful output**.

### 20. Best practice for implementing guardrails in production?

A) Use a single universal rule  
B) Layer pre-, during-, and post-execution guardrails with clear tripwires  
C) Disable retries  
D) Keep them hidden from logs

**Answer:** B  
**Explanation:** Best practice = **layered guardrails** + **tripwires** for maximum safety.