INTERNATIONAL STANDARD

OpenGuardrails AI Runtime Security Management System

(AI-RSMS)

OG-AI-RSMS-001

Version 1.0

Draft

Prepared by the OpenGuardrails Community

2025-01

---

1 OPENGUARDRAILS AI RUNTIME SECURITY MANAGEMENT SYSTEM (AI-RSMS)

International Standard

1.1 Foreword

This document was prepared by the OpenGuardrails community.

AI systems and large language models introduce new classes of runtime risks that are not fully addressed by traditional information security management systems. This International Standard defines a runtime-centric management system to govern, enforce, and audit AI behavior in enterprise environments.

1.2 Introduction

This International Standard specifies requirements for establishing, implementing, maintaining, and continually improving an **AI Runtime Security Management System (AI-RSMS)**.

The AI-RSMS enables organizations to operationalize AI governance, policy enforcement, and regulatory compliance directly within the runtime operation of AI systems, including prompts, models, agents, tools, and outputs.

This document is intended to be compatible with ISO/IEC 27001, ISO/IEC 27701, the EU Artificial Intelligence Act, and related frameworks.

2 1 SCOPE

This International Standard applies to organizations that design, develop, deploy, operate, or manage:

- AI applications (e.g. chatbots, copilots, RAG systems);
- AI agents with autonomous or semi-autonomous behavior;
- AI model services provided internally or by third parties.

The scope includes AI behavior during runtime, including prompts, retrieved context, inference, tool invocation, and outputs.

3 2 NORMATIVE REFERENCES

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document applies.

- ISO/IEC 27000, *Information security management systems — Overview and vocabulary*

- ISO/IEC 27001, *Information security management systems — Requirements*
- ISO/IEC 27701, *Privacy information management*
- ISO/IEC 23894, *Artificial intelligence — Risk management*
- Regulation (EU) 2024/… *Artificial Intelligence Act*

4 3 TERMS AND DEFINITIONS

For the purposes of this document, the following terms and definitions apply.

4.1 3.1 AI runtime

phase during which an AI system processes input, performs inference, invokes tools or functions, and produces output

4.2 3.2 AI control

measure implemented to modify or mitigate risks arising from AI runtime behavior

4.3 3.3 policy

formalized rule defining organizational, regulatory, or ethical constraints applicable to AI behavior

4.4 3.4 enforcement action

action taken when a control condition is met, including blocking, masking, rewriting, escalation, or model switching

5 4 CONTEXT OF THE ORGANIZATION

5.1 4.1 Understanding the organization and its context

The organization shall determine internal and external issues relevant to AI runtime security.

5.2 4.2 Understanding the needs and expectations of interested parties

The organization shall identify interested parties and their requirements related to AI behavior, safety, and compliance.

5.3 4.3 Determining the scope of the AI-RSMS

The organization shall define and document the scope of the AI-RSMS.

6 5 LEADERSHIP

6.1 5.1 Leadership and commitment

Top management shall demonstrate leadership and commitment to AI runtime security.

6.2 5.2 AI runtime security policy

The organization shall establish, maintain, and communicate an AI runtime security policy.

6.3 5.3 Organizational roles, responsibilities and authorities

Roles and responsibilities for AI governance, policy approval, and runtime enforcement shall be assigned and communicated.

7 6 PLANNING

7.1 6.1 Actions to address risks and opportunities

The organization shall identify and assess AI runtime risks considering:

- data sensitivity;
- model autonomy and agency;
- potential impact on individuals, organizations, or systems.

7.2 6.2 AI runtime risk treatment

The organization shall determine and implement risk treatment options using AI runtime controls.

7.3 6.3 AI runtime security objectives and planning

Measurable AI runtime security objectives shall be established and monitored.

8 7 SUPPORT

8.1 7.1 Resources

The organization shall provide adequate resources to implement and operate AI runtime controls.

8.2 7.2 Competence

Personnel involved in AI governance and operations shall be competent.

8.3 7.3 Awareness

Relevant personnel shall be aware of AI runtime security policies and controls.

8.4 7.4 Documented information

Documented information required by the AI-RSMS shall be controlled and maintained.

9 8 OPERATION

9.1 8.1 Operational planning and control

The organization shall implement AI runtime controls as defined in Annex A.

9.2 8.2 Policy-driven runtime enforcement

AI behavior shall be monitored and enforced in real time according to defined policies.

9.3 8.3 AI security incident handling

Procedures shall be established for identifying, responding to, and learning from AI runtime security incidents.

10 9 PERFORMANCE EVALUATION

10.1 9.1 Monitoring, measurement, analysis and evaluation

The organization shall monitor and measure the effectiveness of AI runtime controls.

10.2 9.2 Internal audit

The organization shall conduct internal audits of the AI-RSMS at planned intervals.

10.3 9.3 Management review

Top management shall review the AI-RSMS to ensure its continuing suitability, adequacy and effectiveness.

11 10 IMPROVEMENT

11.1 10.1 Nonconformity and corrective action

The organization shall address nonconformities and take corrective actions.

11.2 10.2 Continual improvement

The organization shall continually improve the AI-RSMS.

12 ANNEX A (NORMATIVE)

12.1 OpenGuardrails AI Runtime Control Library

This annex defines a normative set of AI runtime controls. Organizations claiming conformance to this standard SHALL implement all applicable controls defined in this annex and document their applicability in the AI Statement of Applicability (AI-SoA).

12.2 A.1 Governance & Scope Controls

12.2.1 OG-A1.1 AI Application Definition

The organization SHALL define and document each AI application or AI agent within the scope of the AI-RSMS, including:

- intended purpose and use cases;
- deployment context and user groups;
- boundaries of acceptable behavior.

This information SHALL be reviewed and kept up to date.

12.2.2 OG-A1.2 AI Risk Profiling

The organization SHALL assign a risk profile to each AI application based on:

- data sensitivity;
- level of autonomy or agency;
- potential impact on individuals, organizations, or systems.

Risk profiles SHALL be reviewed when significant changes occur.

12.2.3 OG-A1.3 Tenant and Boundary Isolation

The organization SHALL ensure logical or physical isolation between tenants, applications, or business units to prevent unauthorized access, data leakage, or policy crossover during AI runtime.

12.2.4 OG-A1.4 Policy Ownership and Approval

The organization SHALL assign ownership for AI runtime security policies. Policies SHALL be approved by authorized personnel and reviewed at planned intervals.

12.3 A.2 Policy & Rule Controls

12.3.1 OG-A2.1 Enterprise Policy Enforcement

The organization SHALL define enterprise-specific AI policies and ensure that AI behavior during runtime complies with these policies.

Policy violations SHALL be detected and handled according to defined enforcement rules.

12.3.2 OG-A2.2 Regulatory Policy Enforcement

The organization SHALL identify applicable legal and regulatory requirements related to AI behavior and SHALL enforce corresponding policies during AI runtime.

12.3.3 OG-A2.3 Scope and Off-Topic Control

The organization SHALL ensure that AI outputs remain within the defined scope of the AI application. Off-topic or out-of-scope interactions SHALL be detected and handled according to policy.

12.3.4 OG-A2.4 Restricted Topic Control

The organization SHALL define restricted topics, entities, or domains. AI systems SHALL be prevented from generating content related to restricted topics during runtime.

12.3.5 OG-A2.5 Policy Versioning and Change Control

The organization SHALL maintain version control for AI runtime policies. Policy changes SHALL be documented, traceable, and auditable.

12.4 A.3 Runtime Interaction Controls

12.4.1 OG-A3.1 Prompt Pre-Execution Control

The organization SHALL inspect prompts before AI model execution to detect policy violations, security threats, or misuse.

Non-compliant prompts SHALL be blocked or modified according to policy.

12.4.2 OG-A3.2 Context and Memory Control

The organization SHALL inspect retrieved context, memory, or external knowledge sources used during inference to prevent policy violations or data leakage.

12.4.3 OG-A3.3 Tool and Function Call Control

The organization SHALL restrict and validate AI-initiated tool or function calls. Unauthorized or unsafe tool usage SHALL be prevented during runtime.

12.4.4 OG-A3.4 Output Pre-Delivery Control

The organization SHALL inspect AI-generated outputs before delivery to users. Outputs violating policy or posing unacceptable risk SHALL be blocked, modified, or replaced.

12.4.5 OG-A3.5 Streaming Interruption Capability

Where streaming outputs are used, the organization SHALL implement the capability to interrupt or terminate responses when violations are detected mid-stream.

12.5 A.4 Detection Controls

12.5.1 OG-A4.1 Prompt Injection Detection

The organization SHALL detect attempts to manipulate or override system instructions, policies, or safeguards through prompt injection techniques.

12.5.2 OG-A4.2 Jailbreak Detection

The organization SHALL detect attempts to bypass alignment, safety, or policy constraints using jailbreak or evasion techniques.

12.5.3 OG-A4.3 Sensitive Data Detection

The organization SHALL detect personal, sensitive, or regulated data in AI inputs and outputs during runtime.

12.5.4 OG-A4.4 Enterprise Confidential Data Detection

The organization SHALL detect leakage or misuse of enterprise confidential or proprietary information during AI runtime.

12.5.5 OG-A4.5 Off-Topic and Abuse Detection

The organization SHALL detect irrelevant, abusive, or manipulative interactions that fall outside acceptable AI usage.

12.5.6 OG-A4.6 Agent Abuse Detection

The organization SHALL detect malicious or unintended behaviors by AI agents, including code execution abuse, command injection, or unsafe automation.

12.6 A.5 Enforcement and Response Controls

12.6.1 OG-A5.1 Content Blocking

The organization SHALL block AI inputs or outputs that violate defined policies or exceed acceptable risk thresholds.

12.6.2 OG-A5.2 Content Masking and Redaction

Where appropriate, the organization SHALL mask or redact sensitive information instead of fully blocking content.

12.6.3 OG-A5.3 Response Rewriting

The organization MAY rewrite AI responses to ensure compliance with policies while preserving usability.

12.6.4 OG-A5.4 Human-in-the-Loop Escalation

The organization SHALL provide mechanisms for escalating high-risk cases to human reviewers.

12.6.5 OG-A5.5 Safe Model Switching

The organization SHALL support routing requests to alternative models when higher safety or compliance guarantees are required.

12.6.6 OG-A5.6 Risk-Based Enforcement

The organization SHALL apply enforcement actions based on defined risk levels, policy thresholds, and contextual factors.

12.7 A.6 Evidence, Logging, and Audit Controls

12.7.1 OG-A6.1 Decision Traceability

The organization SHALL ensure that AI runtime decisions are traceable to policies, detections, and enforcement actions.

12.7.2 OG-A6.2 Detection Evidence Logging

The organization SHALL log detection results with sufficient detail to support investigation and audit.

12.7.3 OG-A6.3 Enforcement Evidence Logging

The organization SHALL log enforcement actions taken during AI runtime.

12.7.4 OG-A6.4 Evidence Retention

The organization SHALL retain AI runtime security evidence in accordance with legal, regulatory, and organizational requirements.

12.7.5 OG-A6.5 Compliance and Audit Export

The organization SHALL provide mechanisms to export evidence and control status for audits, assessments, or regulatory review.

13 ANNEX B (INFORMATIVE)

13.1 Mapping to ISO/IEC 27001 and EU Artificial Intelligence Act

This annex provides an informative mapping between the AI-RSMS controls defined in Annex A and selected controls from ISO/IEC 27001 Annex A and requirements of the EU Artificial Intelligence Act. This mapping is provided for guidance only and does not replace formal compliance obligations.

13.2 B.1 Mapping to ISO/IEC 27001 Annex A

Table B.1 — Mapping of AI-RSMS controls to ISO/IEC 27001 Annex A

AI-RSMS Control	Description	ISO/IEC 27001 reference
OG-A1.1	AI application definition	A.5.9 Information security inventory
OG-A1.2	AI risk profiling	A.5.8 Information security risk management
OG-A1.3	Tenant and boundary isolation	A.8.1 User access control
OG-A1.4	Policy ownership	A.5.1 Information security policies
OG-A2.1	Enterprise policy enforcement	A.5.1, A.5.36 Compliance
OG-A2.2	Regulatory policy enforcement	A.18.1 Compliance with legal requirements
OG-A2.3	Scope and off-topic control	A.12.1 Operational procedures
OG-A2.4	Restricted topic control	A.8.3 Information access restriction
OG-A2.5	Policy versioning	A.7.5 Documented information

AI-RSMS Control	Description	ISO/IEC 27001 reference
OG-A3.1	Prompt pre-execution control	A.14.2 Secure development
OG-A3.2	Context and memory control	A.8.2 Privileged access rights
OG-A3.3	Tool and function call control	A.8.2 Privileged access rights
OG-A3.4	Output pre-delivery control	A.12.2 Change management
OG-A3.5	Streaming interruption	A.12.6 Technical vulnerability management
OG-A4.x	Detection controls	A.12.4 Logging and monitoring
OG-A5.x	Enforcement controls	A.16.1 Incident management
OG-A6.x	Evidence and audit controls	A.12.4 Logging, A.18.2 Audits

13.3 B.2 Mapping to EU Artificial Intelligence Act

Table B.2 — Mapping of AI-RSMS controls to EU Artificial Intelligence Act

AI-RSMS Control	EU AI Act article	Alignment description
OG-A1.1	Article 9	Risk management system
OG-A1.2	Article 9	Risk identification and mitigation
OG-A2.2	Article 10	Data governance and misuse prevention
OG-A3.x	Article 15	Accuracy, robustness and cybersecurity
OG-A4.x	Article 15	Detection of anomalous behaviour
OG-A5.4	Article 14	Human oversight
OG-A6.x	Articles 12, 17	Record-keeping and technical documentation

13.4 B.3 Usage guidance

Organizations may use this mapping to: - support internal compliance analysis; - prepare regulatory documentation; - explain alignment to auditors or regulators.

This annex is informative and non-normative.

14 ANNEX C (INFORMATIVE)

14.1 AI Statement of Applicability (AI-SoA)

This annex provides a template for documenting the applicability and implementation status of AI runtime controls defined in Annex A.

14.2 C.1 General information

Organization name: _____

AI application / system: _____

AI-RSMS version: _____

Assessment date: _____

Responsible owner: _____

Table C.1 — Applicability and implementation status

Control ID	Control name	App.	Impl.
OG-A1.1	AI Application Definition	Y	Y
OG-A1.2	AI Risk Profiling	Y	Y
OG-A2.3	Scope Control	Y	Y
OG-A3.3	Tool Call Control	N	N
OG-A5.4	Human-in-the-loop	Y	N
OG-A6.5	Audit Export	Y	Y

Table C.2 — Justification and evidence

Control ID	Justification E	vidence
OG-A3.3	No tools used	N/A
OG-A5.4	Planned in phase 2	Roadmap

14.3 C.3 Summary and approval

Any applicable control that is not implemented shall have documented justification and an associated remediation plan.

Approved by:

Name / role: _____

Signature: _____

Date: _____

15 ANNEX D (INFORMATIVE)

15.1 AI-RSMS Auditor Checklist

15.2 Purpose

This checklist supports audits and assessments of conformity with the **OpenGuardrails AI Runtime Security Management System (AI-RSMS)**.

This annex is informative. Normative requirements are defined in: - Clauses 5–11 of OG-AI-RSMS-001 - Annex A (Normative): AI Runtime Control Library

15.3 D.1 Audit Information

- **Organization**: _____
- **AI System / Application**: _____
- **Audit Type**: □ Internal □ External □ Readiness
- **Audit Date(s)**: _____
- **Auditor(s)**: _____
- **AI-RSMS Version**: _____

15.4 D.2 Conformance Scale

Auditors SHOULD record one of the following for each item:

- **C** – Conformant
- **PC** – Partially Conformant
- **NC** – Nonconformant
- **NA** – Not Applicable

15.5 D.3 Context & Governance Controls

(Clause 5, Annex A.1)

15.5.1 OG-A1.1 AI Application Definition

☐ All AI applications and agents in scope are identified

☐ Purpose and intended use are documented

☐ Acceptable and unacceptable behaviors are defined

☐ Documentation is current and reviewed

Result: □ C □ PC □ NC □ NA

Evidence: _____

15.5.2 OG-A1.2 AI Risk Profiling

☐ Each AI application has an assigned risk profile

☐ Risk profile considers data sensitivity and autonomy

☐ Risk profile is reviewed after significant changes

Result: □ C □ PC □ NC □ NA
Evidence: _____

15.5.3 OG-A1.3 Tenant and Boundary Isolation

☐ Logical or physical isolation is implemented between tenants/applications

☐ Controls prevent cross-tenant data or policy leakage

Result: □ C □ PC □ NC □ NA
Evidence: _____

15.5.4 OG-A1.4 Policy Ownership and Approval

☐ AI runtime security policies are formally approved

☐ Policy ownership is clearly assigned

☐ Policies are reviewed at planned intervals

Result: □ C □ PC □ NC □ NA
Evidence: _____

15.6 D.4 Policy & Rule Controls

(Annex A.2)

15.6.1 OG-A2.1 Enterprise Policy Enforcement

☐ Enterprise-specific AI policies are defined

☐ Policies are enforced during AI runtime

☐ Policy violations are detected and handled

Result: □ C □ PC □ NC □ NA
Evidence: _____

15.6.2 OG-A2.2 Regulatory Policy Enforcement

☐ Applicable laws and regulations are identified

☐ Regulatory requirements are translated into runtime policies

☐ Enforcement is demonstrated during operation

Result: □ C □ PC □ NC □ NA
Evidence: _____

15.6.3 OG-A2.3 Scope and Off-Topic Control

☐ AI application scope is clearly defined

☐ Off-topic or out-of-scope interactions are detected

☐ Appropriate enforcement actions are applied

Result: □ C □ PC □ NC □ NA
Evidence: _____

15.6.4 OG-A2.4 Restricted Topic Control

☐ Restricted topics/entities/domains are defined

☐ Restricted content generation is prevented at runtime

Result: □ C □ PC □ NC □ NA
Evidence: _____

15.6.5 OG-A2.5 Policy Versioning and Change Control

☐ Policies are version-controlled

☐ Changes are documented and traceable

☐ Historical versions can be reviewed

Result: □ C □ PC □ NC □ NA
Evidence: _____

15.7 D.5 Runtime Interaction Controls

(Annex A.3)

15.7.1 OG-A3.1 Prompt Pre-Execution Control

☐ Prompts are inspected before model execution

☐ Non-compliant prompts are blocked or modified

Result: □ C □ PC □ NC □ NA
Evidence: _____

15.7.2 OG-A3.2 Context and Memory Control

☐ Retrieved context or memory is inspected

☐ Controls prevent leakage via context

Result: □ C □ PC □ NC □ NA
Evidence: _____

15.7.3 OG-A3.3 Tool and Function Call Control

☐ Tool/function calls initiated by AI are restricted

☐ Unauthorized tool usage is prevented

Result: □ C □ PC □ NC □ NA
Evidence: _____

15.7.4 OG-A3.4 Output Pre-Delivery Control

☐ AI outputs are inspected before delivery

☐ Violating outputs are blocked, modified, or replaced

Result: □ C □ PC □ NC □ NA
Evidence: _____

15.7.5 OG-A3.5 Streaming Interruption Capability

☐ Streaming responses can be interrupted

☐ Interruption is triggered by detected violations

Result: □ C □ PC □ NC □ NA
Evidence: _____

15.8 D.6 Detection Controls

(Annex A.4)

Verify existence, configuration, and effectiveness of each detection capability:

☐ OG-A4.1 Prompt Injection Detection

☐ OG-A4.2 Jailbreak Detection

☐ OG-A4.3 Sensitive Data Detection

☐ OG-A4.4 Enterprise Confidential Data Detection

☐ OG-A4.5 Off-Topic and Abuse Detection

☐ OG-A4.6 Agent Abuse Detection

Result: □ C □ PC □ NC □ NA
Evidence: _____

15.9 D.7 Enforcement & Response Controls

(Annex A.5)

☐ Enforcement actions are defined and documented

☐ Enforcement is risk-based and configurable

☐ Human-in-the-loop escalation is supported

☐ Safe model switching is supported where required

Result: □ C □ PC □ NC □ NA
Evidence: _____

15.10 D.8 Evidence, Logging, and Auditability

(Annex A.6)

☐ Detection decisions are logged

☐ Enforcement actions are logged

☐ Decision traceability is ensured

☐ Evidence is retained per policy

☐ Evidence can be exported for audit

Result: □ C □ PC □ NC □ NA
Evidence: _____

15.11 D.9 AI Statement of Applicability (AI-SoA)

☐ An AI-SoA exists

☐ All Annex A controls are listed

☐ Non-applicable controls are justified

☐ AI-SoA is approved and current

Result: □ C □ PC □ NC □ NA
Evidence: _____

15.12 D.10 Audit Summary

- **Total Controls Reviewed**: _____
- **Conformant**: _____
- **Partially Conformant**: _____
- **Nonconformant**: _____

15.12.1 Key Findings

_____

15.12.2 Required Corrective Actions

_____

15.13 D.11 Auditor Conclusion

☐ The AI-RSMS **conforms** to OG-AI-RSMS-001

☐ The AI-RSMS **does not conform** and corrective actions are required

Auditor Name / Signature: _____

Date: _____