AI Safety and Risk Management Framework

Global Standards-Aligned Version

Comprehensive Guidelines for the Identification, Assessment, Treatment and Continuous Governance of Risks in Artificial Intelligence Systems

Issued by
International Federation for AI Standards (IFAIS)

IFAIS-GSF-001

Version 1.0 — 12 February 2025

This framework is fully harmonised with ISO/IEC 42001:2023, Regulation (EU) 2024/1689 (AI Act), NIST AI Risk Management Framework 1.0, OECD AI Principles (2019, revised 2024), and UNESCO Recommendation on the Ethics of AI (2021).

Contents

Foreword **2**

1 Scope **2**

Scope **2**

2 Normative References **2**

3 Terms and Definitions **2**

4 Global Principles of Trustworthy AI **3**
4.1 Lawfulness, Compliance and Accountability 3
4.2 Human-Centric Design and Effective Oversight 3
4.3 Transparency and Explainability 3
4.4 Fairness and Non-Discrimination 3
4.5 Technical Mitigations 3
4.6 Privacy and Data Governance 3
4.7 Safety, Security and Robustness 4
4.8 Transparency of Governance 4

5 Risk Classification and Categorisation **4**

6 AI Risk Management Lifecycle **4**
6.1 Context Establishment and System Mapping (MAP) 4
6.2 Risk Assessment (MEASURE) 4
6.3 Risk Treatment (MANAGE) 5
6.4 Monitoring, Measurement and Continuous Improvement (GOVERN) 5

7 Organisational Roles and Responsibilities **5**

8 AI Incident Response and Reporting **5**

9 Documentation and Record-Keeping Requirements **6**

Foreword

The rapid integration of artificial intelligence into critical societal functions has created an urgent need for robust, internationally harmonised governance mechanisms. This AI Safety and Risk Management Framework has been developed to provide organisations with a practical, compliance-ready system that satisfies the most stringent global requirements while remaining flexible enough for organisations of any size or sector.

This document is intended to serve as both a normative framework (containing "shall" statements that establish requirements) and an informative guide (providing explanations, examples and implementation advice).

The International Federation for AI Standards (IFAIS) issues this framework as a public good and encourages its adoption, adaptation and extension by any entity deploying or developing AI systems.

Scope

This framework applies to all organisations — public or private — that design, develop, deploy, operate, procure or audit AI systems. It covers the entire AI lifecycle, from conception and data acquisition through training, testing, deployment, monitoring and decommissioning.

It is applicable to all AI technologies, including machine learning, deep learning, rule-based systems, generative AI, autonomous agents and embedded AI.

Excluded from scope are purely research prototypes that are never placed on the market or put into service.

Normative References

The following documents are indispensable for the application of this framework:

- ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system
- Regulation (EU) 2024/1689 (EU AI Act) and its amendments
- NIST AI Risk Management Framework (AI RMF 1.0), January 2023
- OECD Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449, revised 2024)
- UNESCO Recommendation on the Ethics of Artificial Intelligence (2021)
- ISO/IEC 23894:2023 — Artificial intelligence — Guidance on risk management
- ISO/IEC 27001:2022 — Information security management systems

Terms and Definitions

For the purposes of this document, the following terms apply:

AI system Software using one or more AI techniques (EU AI Act Art. 3(1))

High-risk AI system AI system listed in Annex III of the EU AI Act or meeting the criteria therein

Risk Combination of the probability of occurrence of harm and the severity of that harm

Residual risk Risk remaining after risk treatment

Human oversight Human supervision with the ability to intervene, override or halt the system

Global Principles of Trustworthy AI

Organisations shall implement AI systems in accordance with the following principles. These principles are derived directly from the normative references and are mandatory for compliance.

Lawfulness, Compliance and Accountability

AI systems shall comply with all applicable laws and regulations. Organisations shall appoint accountable persons and maintain complete records sufficient to demonstrate conformity during regulatory inspections or audits.

Accountability shall be allocated at board level and cascaded through the organisation via clear roles (e.g., Chief AI Safety Officer).

Human-Centric Design and Effective Oversight

AI systems shall be designed to augment, not replace, human judgment in high-stakes decisions. Appropriate forms of human oversight (human-in-the-loop, human-on-the-loop or human-in-command) shall be implemented according to risk level.

Transparency and Explainability

Organisations shall provide clear, appropriate and timely information to affected persons about the operation, logic, capabilities and limitations of AI systems. For high-risk systems, technical documentation shall be maintained in accordance with Annex IV of the EU AI Act.

Explainability techniques (model-based or post-hoc) shall be applied where the absence of explainability would create material risk.

Fairness and Non-Discrimination

Organisations shall systematically assess and mitigate bias throughout the AI lifecycle. Demographic parity testing, equalised odds analysis and counterfactual fairness testing shall be conducted on all systems affecting natural persons.

Technical Mitigations

include diverse training data, fairness-aware algorithms, re-weighting and post-processing techniques.

Privacy and Data Governance

Personal data processing shall comply with GDPR, Sri Lanka PDPA or equivalent jurisdictional requirements. Privacy-by-design and privacy-by-default shall be applied. Techniques such as differential privacy, federated learning and synthetic data generation shall be used where appropriate.

Safety, Security and Robustness

AI systems shall be secure by design and resilient to adversarial attacks, data poisoning, prompt injection, model theft and supply-chain compromise. Robustness testing (including red-teaming) shall be performed before deployment and after any material update.

Transparency of Governance

Organisations shall establish and document clear governance structures, policies and procedures for AI risk management.

Risk Classification and Categorisation

Organisations shall classify every AI system according to the following categories:

Risk Level	Criteria (examples)	Required Controls
Prohibited	Deepfakes for blackmail, social scoring, predictive policing using protected characteristics	Prohibited — shall not be developed o
High-risk	EU AI Act Annex III systems, credit scoring, medical diagnosis, autonomous vehicles	Full conformity assessment, third-part
Limited risk	Chatbots, emotion recognition, generative AI	Transparency obligations, user notifica
Low/Minimal risk	Recommendation systems, spam filters	Basic governance and documentation

Table 1: AI System Risk Classification

AI Risk Management Lifecycle

The organisation shall implement a systematic risk management process consisting of the following phases:

Context Establishment and System Mapping (MAP)

The organisation shall:

- Define the AI system's intended purpose, scope, boundaries and context of use
- Identify all internal and external stakeholders
- Conduct an AI Impact Assessment (AIIA) and, where personal data are processed, a Data Protection Impact Assessment (DPIA)
- Classify the system according to Section 6

Risk Assessment (MEASURE)

Risks shall be evaluated using at minimum a 5×5 matrix considering:

- Severity of harm (negligible to catastrophic)
- Likelihood (rare to almost certain)
- Detectability
- Affected population size

Residual risk shall be documented in a Risk Register together with justification for acceptance.

Risk Treatment (MANAGE)

For each identified risk exceeding the organisation's risk appetite, appropriate treatment options shall be applied:

1. Avoid — discontinue the practice
2. Mitigate — implement technical/organisational measures
3. Transfer — insurance or contractual indemnities
4. Accept — only with documented senior management approval

Specific technical controls are detailed in Annex A.

Monitoring, Measurement and Continuous Improvement (GOVERN)

The organisation shall:

- Establish continuous monitoring of model performance, drift and bias
- Conduct quarterly reviews for high-risk systems
- Perform annual management review of the AI management system
- Maintain a change log and perform impact assessment before any material change

Organisational Roles and Responsibilities

- Board of Directors — ultimate responsibility for AI risk appetite and oversight
- AI Governance Committee — meets at least quarterly, chaired by the Chief AI Safety & Compliance Officer (CAISO)
- Chief AI Safety & Compliance Officer (CAISO) — single point of accountability for the AI management system
- Data Protection Officer (DPO) — where required by law
- Model Risk Management Team — technical validation, red-teaming, drift monitoring

AI Incident Response and Reporting

Organisations shall establish and maintain an AI Incident Response Plan that includes:

1. Detection and classification of incidents (including bias incidents, security breaches, safety failures)
2. Containment procedures
3. Root-cause analysis using structured methodologies
4. Recovery and remediation
5. Mandatory reporting: – EU AI Act serious incidents: within 72 hours to the national authority – PDPA/GDPR data breaches: within 72 hours where required – Internal escalation to board level for material incidents

Documentation and Record-Keeping Requirements

Organisations shall maintain, for at least 10 years (or longer if required by law):

- Technical Documentation (EU AI Act Annex IV template)
- Risk Register and treatment plans
- Training records and impact assessments
- Audit logs and monitoring reports
- Conformity assessment records for high-risk systems