Day 12

Generative AI Security

ISO 42001 — AI Management System (AIMS):

ISO 42001 is the world's first standard focused exclusively on AI governance, risk management, and operational controls. It helps organizations ensure their AI systems are safe, secure, responsible, and compliant.

Understanding the Scope of ISO 42001:

What ISO 42001 Covers?

ISO 42001 defines:

- how to manage AI systems throughout their lifecycle
- how to control AI risks
- how to ensure responsible and ethical AI use
- governance rules, documentation, audits, and monitoring
- requirements for security, privacy, fairness, transparency

In short: it creates a management system for AI similar to ISO 27001 for security.

Which Organizations Need It?

ISO 42001 applies to:

- companies building AI models
- companies deploying AI in products
- enterprises using third-party AI systems
- government, healthcare, finance, manufacturing—any high-risk AI users

Any organization that wants strong control over AI risks or needs regulatory trust should use ISO 42001.

Scope Definition for AI Systems:

Before implementing ISO 42001, the org must define:

- which AI systems, models, and datasets fall under the scope
- which teams/departments use AI
- which processes are affected
- what exclusions exist (and why)

Clear scope ≡ clear accountability.

Leadership & Organizational Responsibilities:

Leadership Commitments:

Top leaders must:

- approve AI policies
- assign budgets
- define strategic goals for safe AI
- support the AI governance program
- ensure compliance with laws (EU AI Act, data protection, sector rules)

Management commitment is mandatory.

Roles & Accountability Framework:

ISO 42001 requires defining:

- Responsible — operators, ML engineers
- Accountable — AI product owners, governance heads
- Consulted — legal, HR, ethics, cybersecurity
- Informed — business stakeholders

This aligns closely with a RACI model.

Policy Approvals & Governance Structure:

Organizations must have:

- AI governance board
- AI risk review committee
- Policy review structure
- Approval workflow for AI development and deployment
- Formal escalation routes

Governance must be documented.

AI Lifecycle-based Risk Management:

ISO 42001 requires risk management built into every stage of the AI lifecycle.

Risk Assessment Frameworks Under ISO 42001

Organizations must:

- identify AI risks
- assess likelihood × impact
- classify critical AI systems
- define controls for security, privacy, fairness, ethics
- document risk decisions

This connects to NIST AI RMF and ISO 27005-style risk management.

Mapping AI Risks to Controls:

Examples:

- Prompt injection → input validation + output filtering
- Model drift → continuous monitoring + periodic evaluation
- Bias → fairness metrics + dataset review
- Data poisoning → dataset verification + anomaly checks
- Model theft → access controls + rate limiting

Secure Handling at Each AI Lifecycle Stage.

Documentation Requirements (Cross-Lifecycle)

At each stage you must maintain:

- risk register
- model cards
- data lineage
- change logs
- validation & test reports
- deployment approvals
- monitoring records

ISO 42001 places heavy emphasis on documentation.

Data Governance Requirements:

Dataset Sourcing Requirements:

Organizations must:

- verify where data came from
- confirm legal rights to use it
- check for copyright or licensing issues

Data Quality Assurance:

Practices include:

- removing corrupted or low-quality data
- checking representativeness
- preventing duplicated or fabricated data
- ensuring balanced datasets for fairness

Data Lineage Documentation:

Must capture:

- data origin
- processing steps
- transformations
- storage locations
- who modified it and when

This helps during audits and investigations.

PII Handling & Privacy Mandates:

ISO 42001 requires:

- GDPR-style privacy controls
- anonymization/pseudonymization
- restricting sensitive data in training
- purpose limitation
- consent verification (if applicable)

Restrictions Around Copyrighted or Synthetic Data:

- copyrighted data cannot be used without authorization
- synthetic data should be validated for quality and bias
- synthetic data cannot violate privacy or re-identify individuals

Documentation & Audit Requirements

ISO 42001 has a very strong documentation and audit focus, similar to ISO 27001.

Mandatory Documentation

Includes:

- AI governance policy
- AI risk management policy
- data governance policy
- model lifecycle documentation
- technical controls
- training records
- monitoring records
- incident management documentation

Audit Preparation

Auditors check:

- if controls exist
- if they are implemented
- if they are effective
- if monitoring is continuous
- if evidence is available

Internal audits must be done before certification audits.

Control Evidence Collection

Evidence examples:

- logs
- screenshots

- test results
- approvals
- risk assessments
- model cards
- data lineage records
- training pipeline documentation

Reporting Structures

- governance board reports
- metrics reports (drift, incidents, safety)
- compliance reports
- escalation reports for high-risk incidents

Continuous Improvement Loops

ISO 42001 follows the PDCA cycle:

- Plan → policies, risk framework
- Do → build and deploy AI safely
- Check → audits, monitoring
- Act → update controls, fix gaps

This ensures AI governance evolves over time.

Basically,

- ISO 42001 is a global standard for safe and responsible AI management.
- It requires strong leadership, policies, and governance structures.
- AI risks must be managed across every lifecycle stage.
- Data governance must cover sourcing, quality, lineage, privacy, and copyright.
- Documentation and audits are critical for certification.
- The framework uses continuous improvement to keep AI systems safe and compliant.

Interoperability: Combining NIST CSF, NIST AI RMF, ISO 42001:

Basically,

- NIST CSF handles cybersecurity
- NIST AI RMF handles AI-specific risks and trustworthiness
- ISO 42001 manages governance, documentation, and audits Together they give complete AI security + governance + risk control.
- Organizations combine them into one unified system using shared documents, shared dashboards, common controls, and a single governance board.
- This reduces duplication, improves compliance, and strengthens AI safety.

--The End--