ISO/IEC 42001:2023

AI Security & Pentesting Requirements

A practical guide for CTOs, security leaders, and AI teams navigating ISO 42001 certification, with AI security testing guidance, Annex A control mapping, and practical pentest scoping considerations.

Contents

01 Quick Summary

WHAT CTOS AND SECURITY LEADERS NEED TO KNOW

ISO/IEC 42001:2023 is the first international management-system standard specifically designed for artificial intelligence. It establishes requirements for creating and continually improving an Artificial Intelligence Management System (AIMS) that governs how an organization develops, provides, or uses AI systems.

ISO 42001 does not explicitly require penetration testing. Instead, it requires organizations to identify AI risks, establish controls, verify and validate AI systems, monitor their operation, manage incidents, and evaluate whether the AIMS is effective. Technical security testing can provide valuable evidence that some of those controls work in practice rather than only on paper.

HOW TO READ THIS GUIDE

GOVERNANCE REQUIREMENT Documentation, policy, or process the AIMS itself must have in place.

TECHNICAL EVIDENCE Where a penetration test or other technical validation can directly demonstrate a control works.

AT A GLANCE

- ISO 42001 applies to organizations that develop, provide, or use AI systems.
- Certification focuses on the organization's AI Management System, not simply one AI application.
- AI risk assessment and risk treatment are central to the standard.
- Annex A contains 38 reference controls across 9 control areas (A.2-A.10), covering AI lifecycle management, data, system monitoring, transparency, responsible use, and suppliers.
- Security testing can support evidence for controls related to verification and validation, operation and monitoring, event logging, data protection, intended use, and third-party AI systems.
- Traditional application-security testing remains important because most AI functionality is delivered through web applications, APIs, cloud services, retrieval systems, and identity layers.
- AI-specific testing may additionally examine prompt injection, excessive agency, tool abuse, retrieval authorization failures, sensitive-data disclosure, and AI trust boundaries.
- Vulnerability scanning alone cannot evaluate many AI-specific attack paths or business-logic failures.
- AI security testing should be risk-driven, based on the AIMS scope, AI architecture, intended use, and applicable Annex A controls.

ISO specifically identifies risk management, transparency, traceability, and accountability among the intended benefits of the standard.

02 What Is ISO/IEC 42001:2023?

BACKGROUND AND CONTEXT

ISO/IEC 42001:2023 specifies requirements for establishing, implementing, maintaining, and continually improving an Artificial Intelligence Management System (AIMS).

An AIMS is the governance system an organization uses to manage its responsibilities around AI. It connects policies, roles, risk management, AI lifecycle processes, data governance, monitoring, transparency, and continual improvement.

ISO 42001 is designed for organizations of any size and sector involved in developing, providing, or using AI-based products or services.

Who Does ISO 42001 Apply To?

Organization Type	Examples
AI product developers	SaaS companies adding copilots, agents, recommendation engines, or generative AI features
AI platform providers	Companies providing models, APIs, orchestration platforms, or AI infrastructure
Organizations integrating AI	Software companies embedding third-party models into customer-facing products
Organizations using AI internally	Companies using AI for customer support, HR, analytics, security, development, or decision support
Regulated organizations	Financial services, healthcare, government, education, and other organizations requiring formal AI governance
Enterprise suppliers	Vendors facing customer requirements for documented AI governance and responsible-use controls

Key Terms

Term	Plain-English Meaning
AIMS	Artificial Intelligence Management System - the policies, processes, responsibilities, and controls used to govern AI
AI system	A system using AI techniques to generate outputs such as predictions, content, recommendations, or decisions
AI risk	Potential harm, failure, misuse, security issue, operational impact, or other uncertainty associated with AI
AI system lifecycle	The stages from requirements and development through validation, deployment, operation, monitoring, and retirement
Impact assessment	Evaluation of how an AI system may affect individuals, groups, organizations, or society
Interested party	A stakeholder affected by or interested in the organization's use or provision of AI

Term	Plain-English Meaning
Statement of Applicability	Documentation showing which Annex A controls are applicable to the AIMS and why
AI supplier	A third party supplying models, datasets, services, platforms, tooling, or other components used in an AI system

03 How the AIMS Is Structured

THE MANAGEMENT SYSTEM + CONTROL MODEL

ISO 42001 follows the same general Harmonized Structure used by other ISO management-system standards such as ISO 27001. It uses the same management-system concept - mandatory governance clauses supported by risk-selected reference controls - but applies it specifically to AI governance.

Mandatory Management-System Clauses

GOVERNANCE REQUIREMENT

Clause	Focus
Clause 4	Context of the organization and AIMS scope
Clause 5	Leadership, policy, roles, and responsibilities
Clause 6	Planning, AI risk assessment, risk treatment, objectives
Clause 7	Resources, competence, awareness, communication, documentation
Clause 8	Operational planning and control
Clause 9	Monitoring, measurement, internal audit, and management review
Clause 10	Nonconformity, corrective action, and continual improvement

ISO describes ISO 42001 as a management-system standard using a Plan-Do-Check-Act approach, rather than a standard that evaluates individual AI applications in isolation.

Annex A Reference Control Areas

Annex A groups 38 AI-specific reference controls across nine control areas. It is a catalogue organizations select from, not a checklist implemented top to bottom - selections are driven by the AIMS scope and the AI risk and impact assessments, and documented in the Statement of Applicability.

GOVERNANCE REQUIREMENT

Area	Focus
A.2	Policies related to AI
A.3	Internal organization
A.4	Resources for AI systems

Area	Focus
A.5	Assessing impacts of AI systems
A.6	AI system lifecycle
A.7	Data for AI systems
A.8	Information for interested parties
A.9	Use of AI systems
A.10	Third-party and customer relationships

04 ISO 42001 Controls and Security Testing Mapping

WHERE TECHNICAL TESTING PROVIDES USEFUL EVIDENCE

Not every ISO 42001 control should be tested through a penetration test. Many controls concern policy, accountability, impact assessment, organizational responsibilities, documentation, or stakeholder communication - those require governance evidence.

However, where an organization claims that technical safeguards prevent misuse, unauthorized access, unintended AI behavior, data exposure, or insecure system operation, security testing can verify whether those safeguards actually work.

A.6 - AI SYSTEM LIFECYCLE

SECURITY TESTING RELEVANCE

A.6 contains some of the strongest connections to technical testing.

Control Area	Security Testing Relevance
AI system requirements	Security requirements should include relevant abuse cases, authorization boundaries, data restrictions, and AI-specific threat scenarios
Design and development	Architecture testing can validate trust boundaries between the application, model, retrieval systems, APIs, tools, and external services
Verification and validation (A.6.2.4)	AI security testing can form part of the evidence that security assumptions and safeguards were actually validated
Deployment	Testing can identify insecure configuration, exposed interfaces, weak credentials, overly permissive APIs, or unsafe deployment defaults
Operation and monitoring (A.6.2.6)	Adversarial testing can determine whether malicious prompts, unauthorized tool calls, data-access attempts, or other attacks are detected
Technical documentation (A.6.2.7)	Pentest results can identify undocumented attack surfaces, integrations, or system behaviors
Event logging (A.6.2.8)	Testing can verify whether AI activity, tool calls, authorization failures, and security-relevant events are captured and attributable

A.6.2.4 - AI SYSTEM VERIFICATION AND VALIDATION

This is one of the strongest technical-security hooks in ISO 42001. Verification and validation should demonstrate that an AI system meets its defined requirements. Where those requirements include security, privacy, access control, intended use, or resistance to misuse, technical security testing can provide direct evidence - for example: can users bypass authorization boundaries through the AI interface? Can one tenant retrieve another tenant's data? Can untrusted content influence privileged AI actions? Can external integrations be abused to perform actions outside the user's permissions?

A.7 - DATA FOR AI SYSTEMS

SECURITY TESTING RELEVANCE

AI applications often introduce data flows that traditional application-security assessments miss. ISO 42001's data controls address management of AI data, data acquisition, quality, provenance, and preparation. Security testing does not replace data-governance evidence, but it can validate whether technical controls actually protect AI data.

Testing Area	Relevant Security Question
Training or reference datasets	Can unauthorized users access, modify, or replace datasets?
RAG knowledge bases	Are retrieval permissions enforced per user, role, organization, or tenant?
Vector databases	Can embeddings or metadata expose information users should not access?
Data ingestion	Can malicious documents or content influence AI behavior?
Data provenance	Could attackers introduce or substitute untrusted data without detection?
Data preparation pipelines	Are secrets, personal information, or sensitive customer data exposed during processing?

A.8 - INFORMATION AND INCIDENT COMMUNICATION

A.8 addresses system documentation, reporting, communication of incidents, and information supplied to interested parties. Consider an AI agent that is exploited to access another customer's information:

GOVERNANCE REQUIREMENT

Does the organization have a documented incident-communication process?

TECHNICAL EVIDENCE

Would the organization actually know the unauthorized access occurred?

The first is answered by policy review. The second is answered by testing whether the AI system's logging and monitoring would actually surface the event.

A.9 - USE OF AI SYSTEMS

SECURITY TESTING RELEVANCE

A.9 addresses responsible-use processes, objectives for responsible use, and ensuring systems are used according to their intended use. Security testing can challenge safeguards designed to enforce those boundaries. Statements such as “the assistant cannot access another customer's records” or “the agent may only execute actions available to the authenticated user” are testable security claims.

05 Highlighted AI Security Control Areas

WHERE PENTEST VULNERABILITIES MAP TO ANNEX A

AI APPLICATION ACCESS AND AUTHORIZATION

SECURITY TESTING RELEVANCE

Relevant areas: A.6, A.9

AI features do not replace traditional authorization controls - they create additional ways to reach them. A conventional application may expose data through a predictable API call. An AI assistant may choose which API to call, construct the parameters, and return the result to the user. If authorization is enforced only through the AI's instructions rather than at the underlying data or tool layer, the control can fail.

Testing Area	Example
User-to-user authorization	Can User A ask the AI to retrieve User B's records?
Tenant isolation	Can one customer retrieve information belonging to another organization?
Role enforcement	Can a standard user convince the AI to perform an administrator-only operation?
Tool authorization	Does the backend independently verify permission before an AI tool executes?
Object-level authorization	Are IDs, filenames, document references, or resource identifiers revalidated server-side?
AI/API trust boundaries	Does the application assume an AI-generated API request is automatically trusted?

THE KEY PRINCIPLE

The model should not be the security boundary. Instructions such as “never expose another user's information” are useful behavioral controls, but the application should still enforce authorization at the data, API, and tool layers. This is why AI security testing should include conventional application and API authorization testing rather than only adversarial prompting.

PROMPT INJECTION AND INSTRUCTION MANIPULATION

SECURITY TESTING RELEVANCE

Relevant areas: A.6.2.4, A.6.2.6, A.9

Prompt injection occurs when untrusted instructions influence how an AI system behaves. The attack does not necessarily have to come directly from a user - it can also arrive indirectly through content the AI retrieves or processes, including uploaded files, webpages, email, support tickets, CRM records, knowledge-base articles, third-party API content, or database records. The security impact depends on what the AI can access or do: an assistant that only generates text creates a different risk than an agent that can send emails, query customer systems, issue refunds, modify accounts, deploy code, or access internal documents.

Scenario	Security Question
Direct prompt injection	Can user instructions override security-sensitive system behavior?
Indirect prompt injection	Can malicious retrieved content control the AI?
System prompt disclosure	Can hidden instructions expose secrets or sensitive architecture?
Tool manipulation	Can an attacker influence which tool executes or its parameters?
Instruction hierarchy	Are high-privilege system rules reliably enforced?
Unsafe output handling	Is model output trusted by downstream systems without validation?

PROMPT INJECTION IS NOT ALWAYS A VULNERABILITY

Being able to make a model ignore a benign formatting instruction is different from being able to access customer records or perform an unauthorized action. A useful AI pentest should focus on security impact, not simply whether model behavior can be manipulated.

RETRIEVAL-AUGMENTED GENERATION AND DATA SEGREGATION

SECURITY TESTING RELEVANCE

Relevant areas: A.6, A.7, A.9

Retrieval-Augmented Generation (RAG) allows AI systems to retrieve information from external data sources before generating a response. This creates an important authorization question: does the retrieval layer enforce the same permissions as the source system?

Testing Area	What to Validate
Tenant isolation	Users cannot retrieve another tenant's documents
User-level permissions	Source document access restrictions survive ingestion into the AI system
Deleted / revoked content	Removed access is reflected in the retrieval system
Search filters	Metadata filters cannot be bypassed or manipulated
Vector database access	Direct access to embeddings or metadata is properly restricted
Source citations	References do not reveal hidden filenames, URLs, IDs, or restricted content
Cached responses	Sensitive retrieved information is not exposed to subsequent users

COMMON DESIGN FAILURE

A document platform may correctly restrict access to a file. But once the file is copied into a shared vector database, the original authorization model can disappear - the AI may then retrieve content that the underlying application would never have shown the user. This is not fundamentally a “model problem.” It is an authorization failure in the retrieval architecture.

AI testing should also determine whether sensitive information can leak through prompts and responses, model context, retrieval results, error messages, logs and traces, tool responses, cached sessions, system prompts, training or fine-tuning data, or third-party model APIs. Controls should be evaluated at every stage where sensitive information enters, passes through, or exits the AI system.

AI AGENTS, TOOL USE AND EXCESSIVE AGENCY

SECURITY TESTING RELEVANCE

Relevant areas: A.6, A.9, A.10

AI agents create a substantially different attack surface because the system can take actions rather than only generate responses. An agent may be able to access internal records, update CRM data, issue refunds, send emails, create tickets, run queries, modify files, execute code, call internal or third-party APIs, or trigger business workflows.

Security Control	Test
User authorization	Can the agent perform an action the authenticated user cannot?
Tool permissions	Does each tool receive the minimum required privilege?
Parameter validation	Can the model provide unsafe or manipulated arguments?
Human approval	Are high-impact actions actually blocked until approval occurs?
Transaction boundaries	Can several individually permitted actions combine into an unauthorized outcome?
Tool chaining	Can attackers create dangerous sequences across multiple tools?
External content	Can retrieved content trigger unauthorized actions?
Auditability	Can investigators determine who caused an action, what the AI decided, and which tool executed it?

HIGHER CAPABILITY = HIGHER SECURITY REQUIREMENT

The more authority an AI system receives, the less appropriate it becomes to rely on prompt-level restrictions alone. Sensitive actions should be protected using conventional controls such as server-side authorization, scoped service credentials, deterministic validation, rate limits, transaction limits, approval workflows, sandboxing, logging, and monitoring.

THIRD-PARTY AI AND MODEL SUPPLY CHAIN

GOVERNANCE REQUIREMENT TECHNICAL EVIDENCE

Relevant area: A.10

Most AI applications depend heavily on third parties: foundation-model providers, hosted inference APIs, embedding providers, vector databases, AI gateways, orchestration frameworks, data suppliers, evaluation platforms, and plugins or connectors. A.10 specifically addresses allocation of responsibilities, suppliers, and customer relationships across the AI lifecycle.

Area	What to Validate
API credentials	Keys are not exposed in applications, prompts, logs, or client-side code
Model-provider permissions	Service credentials cannot access unnecessary models, data, or functions
Data transmission	Sensitive information sent to providers follows intended restrictions
Webhooks / callbacks	Third-party integration endpoints authenticate requests correctly
Provider failure modes	Errors do not expose internal data or cause unsafe fallback behavior
Model changes	Application assumptions remain valid when model versions change
External tools	Plugins / connectors cannot expand the agent's privilege unexpectedly

THIRD-PARTY GOVERNANCE + TECHNICAL VALIDATION

Supplier questionnaires and contracts answer what the supplier is supposed to do. Technical testing answers what happens when the integration is attacked. Organizations often need both.

LOGGING AND MONITORING

SECURITY TESTING RELEVANCE

Relevant areas: A.6.2.6, A.6.2.8

Testing should determine whether security-relevant AI activity is observable - for example, repeated prompt-injection attempts, unauthorized retrieval attempts, unexpected tool invocation, blocked admin actions, high-risk model outputs, changes to AI configuration, unusual data access, privilege escalation attempts, or sensitive information returned in responses. A control that blocks an attack but leaves no useful audit trail may still create an investigation and accountability gap.

06 AI Risk Assessment vs. Security Testing

THEY ANSWER DIFFERENT QUESTIONS

ISO 42001 places significant emphasis on identifying and managing AI risks and impacts. An AI risk or impact assessment asks: what could go wrong, who could be affected, and what controls should exist? Security testing asks: can an attacker actually make one of those scenarios happen?

RISK & IMPACT ASSESSMENT | SECURITY TESTING

Example

Step	Detail
Identified risk (Governance)	An AI assistant can access customer account information.
Planned control (Governance)	The assistant may only access records belonging to the authenticated user's organization.
Security test (Technical)	Attempt to manipulate the assistant, retrieval layer, API requests, resource IDs, or tool calls to access another organization's information.
Possible result (Technical)	The model refuses the request, but the underlying retrieval API accepts a manipulated tenant ID. The policy and model guardrail existed - the technical control did not.

RISK ASSESSMENT SHOULD DRIVE THE TEST

A good ISO 42001-aligned security test should start with the organization's AI inventory, intended uses, identified risks, architecture, applicable Annex A controls, and existing safeguards. Testing should then challenge the safeguards protecting the highest-risk scenarios.

ISO itself positions risk assessment and treatment as central to the management-system approach, while ISO/IEC 42005:2025 - published in 2025 - provides a separate, complementary standard specifically for AI system impact assessment.

07

Pentest Scope for ISO 42001

WHAT SHOULD BE IN SCOPE?

Unlike ISO 27001, ISO 42001 does not define pentest scope through an information-security boundary alone. Testing should follow the AI system architecture and AIMS scope.

The first question should be: which AI systems, data flows, integrations, users, and business processes are included in the AIMS? Then determine which components enforce security-relevant controls.

SECURITY TESTING RELEVANCE

Component	Typical Testing Focus
Web application	Authentication, authorization, session management, business logic, conventional web vulnerabilities
Public / private APIs	Object-level authorization, privilege enforcement, input validation, rate limits
LLM interface	Prompt injection, system instruction exposure, sensitive output
RAG / retrieval layer	Tenant isolation, document authorization, metadata leakage
AI agents	Tool abuse, excessive agency, unauthorized actions, approval bypass
Tools and integrations	API permissions, trust boundaries, parameter validation
Model provider	Credentials, data exposure, API configuration, provider trust assumptions
Vector database	Access control, tenant separation, metadata and document leakage
AI data pipeline	Unauthorized modification, sensitive-data handling, ingestion trust
Cloud infrastructure	IAM, secrets, storage, service configuration
Identity systems	SSO, MFA, roles, service identities, delegated authorization
Logging / monitoring	Detection coverage, forensic visibility, event integrity
Admin interfaces	Configuration changes, model settings, privileged functions
CI/CD and deployment	Secrets, model / config changes, deployment authorization

What Should Not Automatically Be Included

ISO 42001 certification does not mean every AI-related component must undergo every possible type of offensive testing. Scope should be based on the AIMS boundary, system architecture, identified risk, control applicability, user privileges, data sensitivity, AI capabilities, exposure to untrusted input, and potential impact of misuse.

08 ISO 42001 and ISO 27001

WHERE AI GOVERNANCE AND INFORMATION SECURITY OVERLAP

ISO 42001 and ISO 27001 solve different problems. ISO offers the two standards together because one addresses AI management while the other addresses information security management.

GOVERNANCE REQUIREMENT

ISO 42001	ISO 27001
Governs responsible development, provision, and use of AI	Governs information-security risks
AI risk and impact assessment	Information-security risk assessment
AI lifecycle governance	Secure development and operations
AI-specific data considerations	Confidentiality, integrity, availability
AI transparency and interested parties	Information-security communication
Responsible use	Access control and acceptable use
AI suppliers	ICT supply-chain security
AI operation and monitoring	Logging and security monitoring

Where the Pentest Fits

TECHNICAL EVIDENCE

An AI pentest may produce evidence relevant to both standards. For example, a cross-tenant RAG exposure vulnerability is relevant to ISO 42001 AI verification and validation, responsible use, AI system operation, and AI data governance - and equally relevant to ISO 27001 access restriction, security testing, technical vulnerability management, and application security.

PRACTICAL TAKEAWAY

Organizations pursuing both certifications should avoid running two disconnected security programs. The same technical evidence can often support information-security assurance under ISO 27001 and AI-control assurance under ISO 42001.

09 What Certification Auditors Expect

EVIDENCE OF CONTROL OPERATION, NOT A MANDATORY "AI PENTEST" CHECKBOX

Certification auditors are assessing whether the AIMS conforms to ISO 42001 and whether applicable controls are implemented and operating effectively. The standard itself does not establish a mandatory annual AI penetration test. Where security-related controls are applicable, however, organizations should be prepared to show credible evidence that those controls have been validated.

Typically Relevant Evidence

GOVERNANCE REQUIREMENT	TECHNICAL EVIDENCE WHERE APPLICABLE
Defined AIMS scope; AI inventory; AI policies	AI verification and validation results
Roles and responsibilities	Application / API pentest reports
AI risk assessments; risk-treatment documentation	AI-specific security-test results
Impact assessments; Statement of Applicability	Threat modeling; authorization testing
Supplier assessments; AI lifecycle documentation	Data-isolation testing; agent / tool permission testing
Incident-management records; internal audit; management review	Vulnerability-management evidence; monitoring and logging validation; remediation and retest evidence

What an ISO 42001-Aligned Security Test Report Should Include

- AIMS / AI-system scope tested
- Architecture and trust boundaries
- Systems, models, APIs, retrieval sources, and tools included
- User roles tested
- Testing methodology and AI-specific attack scenarios
- Conventional web / API testing coverage
- Clear reproduction evidence
- Business and security impact, affected technical controls
- Relationship to relevant AI risks; severity and prioritization
- Remediation guidance and retest evidence
- Limitations and components not tested

IMPORTANT

The report should not claim that a pentest "makes you ISO 42001 compliant." Certification applies to the organization's AIMS as a whole. A better statement is: "This assessment provides technical evidence supporting the organization's evaluation of security-related risks and controls within its AI Management System."

10 When Should AI Security Testing Happen?

TEST WHEN THE RISK CHANGES

ISO 42001 is built around continual management and improvement of AI systems. AI security testing should therefore be triggered by meaningful changes rather than treated exclusively as an annual compliance event.

SECURITY TESTING RELEVANCE

Trigger	Why Retest
Before initial certification	Establish technical evidence for applicable security controls
New customer-facing AI feature	New input, data, authorization, and abuse paths
New RAG implementation	Retrieval introduces new data-access boundaries
AI agent gains tools	System can now take actions rather than only generate text
New model / provider	Model behavior and trust assumptions may change
New data source	Changes retrieval, privacy, provenance, and poisoning risk
New tenant model	Data-segregation boundaries change
Major architecture change	Trust boundaries or authorization may move
Significant AI incident	Verify corrective actions and identify related attack paths
Major privilege expansion	AI receives greater access to internal systems
Before surveillance / recertification	Refresh evidence for controls and confirm remediation

ANNUAL TESTING MAY STILL MAKE SENSE

Many organizations already perform annual penetration testing for ISO 27001, SOC 2, customer assurance, or internal security programs. For those organizations, the practical approach may be an annual application / API pentest plus AI-specific scope whenever material AI functionality exists - avoiding a completely separate AI security program.

Do AI Pentesters Need to Be Independent?

GOVERNANCE REQUIREMENT

ISO 42001 does not establish a universal requirement that AI security testing be conducted by a specific type of accredited penetration-testing provider. However, independent testing can strengthen the credibility of technical evidence, particularly when it supports certification evidence, customers request third-party assurance, internal developers built the controls being tested, the AI functionality processes sensitive information, the system can perform high-impact actions, or testing requires specialized AI security expertise.

Testing Approach	Value
Development-team testing	Essential during development, but not independent
Internal security team	Strong ongoing assurance where sufficient expertise exists
Automated AI scanners	Useful for breadth and regression testing, but limited for authorization and business logic
Bug bounty	Useful supplementary evidence, but uncontrolled scope
Independent AI pentest	Strong evidence for complex AI attack paths and control validation
Traditional pentest with no AI methodology	May miss RAG, agent, model, and prompt-driven attack paths

THE PENTESTER SHOULD UNDERSTAND BOTH LAYERS

AI testing should not be performed as a prompt-only exercise. The pentester should be capable of evaluating AI behavior, application security, APIs, authorization, cloud / infrastructure, and business logic together - many of the highest-impact AI vulnerabilities occur where those layers intersect.

Does Your Pentest Cover Your AI Attack Surface?

Adding AI to an application can introduce new trust boundaries, data flows, tools, authorization paths, and attack scenarios that were not present during your last penetration test. Software Secured's AI security testing combines traditional application and API penetration testing with adversarial testing of the AI layer.

Assess your AI attack surface before certification, an enterprise security review, or your next major release.

Book an AI pentest scoping conversation ->

11 Resources

ISO 42001 AND AI SECURITY REFERENCES

ISO/IEC 42001:2023 - Artificial Intelligence Management System
iso.org/standard/42001
Primary standard for establishing, implementing, maintaining, and continually improving an AIMS. ISO describes the standard as applicable to organizations developing, providing, or using AI systems.

ISO/IEC 23894:2023 - Guidance on Risk Management for AI
iso.org/standard/77304
Guidance for managing AI-related risk, built on ISO 31000. ISO identifies it as a complementary AI risk-management standard.

ISO/IEC 42005:2025 - AI System Impact Assessment
iso.org/standard/42005
Published in 2025. Provides a structured approach to identifying and documenting intended and unintended impacts of AI systems throughout the lifecycle.

ISO/IEC 27001:2022 - Information Security Management Systems
iso.org/standard/27001
Relevant where AI-system security depends on conventional information-security controls including access management, vulnerability management, secure development, cloud security, and monitoring.

NIST AI Risk Management Framework
nist.gov/itl/ai-risk-management-framework
Useful complementary framework for identifying, measuring, managing, and governing AI risk.

OWASP Top 10 for LLM Applications
owasp.org/www-project-top-10-for-large-language-model-applications
Useful reference for security risks affecting generative-AI applications, including prompt injection, sensitive-information disclosure, excessive agency, and insecure output handling.

OWASP API Security Top 10
owasp.org/API-Security
Relevant because AI systems frequently expose or consume APIs and tools where authorization and object-level access controls must be independently enforced.

MITRE ATLAS
atlas.mitre.org
Threat knowledge base covering adversarial tactics and techniques affecting AI-enabled systems.

Software Secured: AI Penetration Testing
softwaresecured.com/service/ai-pentesting
Practical guidance on testing AI-enabled SaaS applications beyond automated scanning, including application, API, retrieval, authorization, and AI-agent attack surfaces.