Deloitte.

Deloitte AI Agent Security Framework

Version 1.0

Deloitte.

Framework

Document Title:			
Deloitte AI Agent Security Framework			
Market Offering / Submarket Offering:		Service Offering	
Cyber		**Cyber AI**	
Information Security Classification:	Document Reference #:	Review Cycle:	Revision No:
Proprietary and Confidential	**01**	**1 Year**	**0.1**

First Released Date:	Posted Date:	Last Reviewed Date:	Last Revised Date:
(YYYY/MM/DD)	(YYYY/MM/DD)	(YYYY/MM/DD)	(YYYY/MM/DD)
Approved by (name and position):		Approved on:	
TBD		**(YYYY/MM/DD)**	

History of Changes

Date	Revision #	Description of Changes
(2025/06/10)	0.1	First draft for review, update, and approval.

1 Introduction

Artificial Intelligence (AI) technologies, particularly AI agents, are transforming business and IT operations, enabling automation, efficiency, and innovation across virtually every organizational domain. AI agents autonomously or semi-autonomously perform tasks, make decisions, and execute processes traditionally handled by human operators. As reliance on these advanced technologies grows, so do the complexities and risks associated with their use, including cybersecurity threats, ethical dilemmas, compliance challenges, and potential societal impacts.

Securing AI agents and ensuring their responsible and compliant use has become critical. Recognizing these challenges, Deloitte has developed this comprehensive **AI Agent Security Standard Framework**, aligned explicitly with international standards including ISO/IEC 27001:2022 (Information Security Management), ISO/IEC 42001:2023 (AI Management Systems), and ISO/IEC 42005:2025 (AI Impact Management). This framework provides organizations with structured guidelines, robust controls, and clear implementation pathways to proactively secure AI agents, manage associated risks, ensure regulatory compliance, and foster stakeholder trust.

The adoption of this framework will empower organizations to confidently leverage AI technologies, securing their benefits while responsibly managing their risks.

1.1 Purpose

The purpose of this AI Agent Security Standard Framework is to establish a comprehensive and structured approach for securely managing AI agents across their entire lifecycle—from initial design and development, through deployment, operational usage, and ultimately retirement. Specifically, the framework aims to:

- Establish clear guidelines and controls for secure and ethically responsible AI usage.
- Proactively identify, assess, and manage unique security, ethical, compliance, operational, and societal risks inherent to AI technologies.
- Align organizational practices with internationally recognized cybersecurity and AI management standards, including ISO/IEC 27001:2022, ISO/IEC 42001:2023, and ISO/IEC 42005:2025.
- Provide comprehensive mechanisms for logging, monitoring, auditing, and explaining AI decisions to stakeholders and regulatory bodies.
- Institutionalize continuous improvement practices to adapt to emerging threats, technologies, and regulatory changes.

1.2 Scope

This framework applies universally to all AI agents and AI-driven processes developed, implemented, operated, or utilized within the organization. It includes internally developed AI agents as well as externally sourced AI components from third-party vendors, suppliers, cloud services, or open-source repositories. The framework covers AI-driven automated processes across diverse business and IT domains, such as cybersecurity, finance, customer services, operational processes, internal governance, and compliance management. All internal teams, personnel, and external entities involved in the AI lifecycle shall fully adhere to the requirements specified herein.

1.3 How to Read This Framework

This framework is structured to facilitate clarity, ease of reference, and practical applicability. Each major section of the document is organized as follows:

- ***Objective***: Clearly defines the purpose and goals of each specific section.
- ***Requirements***: Enumerates mandatory, clearly articulated controls and practices organizations must implement.
- ***Implementation Guidelines***: Provides practical guidance, recommendations, and best practices for effectively meeting specified requirements.
- ***ISO Control Mapping***: Explicitly references relevant clauses from ISO/IEC 27001:2022, ISO/IEC 42001:2023, and ISO/IEC 42005:2025 standards, facilitating traceability and compliance validation.

2 Governance and Oversight of AI Agents

Objective

Establish a robust governance framework providing clear strategic oversight, accountability, and control mechanisms for managing AI agents securely, ethically, and compliantly throughout their lifecycle.

Requirements

2.1 The organization shall implement and maintain an AI Agent Security Governance Program, integrated into cyber security program. The program shall oversee risk management, compliance, and lifecycle governance of AI agents across the organization.

2.2 Top management shall formally approve and communicate a comprehensive AI Security Policy, clearly articulating the objectives, principles, responsibilities, acceptable risk thresholds, and ethical standards guiding AI agent use.

2.3 Clearly defined roles and responsibilities shall be formally documented and assigned, specifically including: 1) AI System Owners accountable for agent operations and compliance; 2) AI Risk Managers responsible for conducting and overseeing risk and impact assessments; 3) AI Ethics Officer or Committee overseeing ethical alignment, fairness evaluations, and compliance with ethical standards; and 4) Technical Leads ensuring technical implementation adheres to security and ethical standards.

2.4 The organization shall explicitly enforce segregation of duties to prevent conflicts of interest and unauthorized activities, ensuring that: 1) development and deployment roles are separated from approval and monitoring roles; 2) no single individual or AI agent can independently develop, approve, deploy, and monitor AI agents.

2.5 A cross-functional AI Oversight Committee shall be established to provide strategic oversight, specifically responsible for 1) reviewing and approving high-risk or high-impact AI agent use cases; 2) ensuring comprehensive completion and review of risk and impact assessments; and 3) monitoring ongoing AI-related incidents and ensuring lessons learned inform governance improvements.

2.6 AI governance practices shall be integrated into broader enterprise risk management and compliance functions. An up-to-date, centrally managed AI Agent Inventory shall be maintained, documenting each agent's purpose, owner, risk level, and lifecycle status.

2.7 All documentation relating to AI governance, policies, risk assessments, impact assessments, approvals, and oversight activities must be managed centrally in a secure, accessible repository to support transparency, auditability, and compliance.

Implementation Guidelines

- Regularly review governance roles and responsibilities to maintain relevance to evolving AI use cases and regulatory requirements.
- Maintain clear escalation and approval pathways, particularly for high-risk or sensitive AI agent deployments.
- Provide structured training programs for governance roles, emphasizing security, ethical standards, and regulatory compliance.

ISO Control Mapping

Standard	Clause / Control	Description
ISO/IEC 27001:2022	• Clause 5.1, Annex A.5.1.1, A.5.2.1 • Annex A.5.3.1	• Leadership, security policy, roles, and responsibilities • Segregation of duties
ISO/IEC 42001:2023	• Clause 5.1, Annex A.2.2, A.3.1	• AI leadership, AI governance policy, defined AI roles and responsibilities
ISO/IEC 42005:2025	• Core Process 4.1, Annex A	• Integration of AI impact assessment into governance, clearly defined oversight roles

3 Ethical and Responsible AI Practices

Objective

Embed ethical principles and responsible practices foundationally within AI governance, ensuring AI agents are developed, deployed, and operated in alignment with fairness, transparency, accountability, human rights, regulatory compliance, and societal values.

Requirements

3.1 The organization shall formally document, approve, and communicate a clear set of ethical AI principles, including fairness, transparency, accountability, explainability, privacy, non-discrimination, and respect for human rights. This Ethical AI Policy must guide all AI agent-related activities.

3.2 Regular ethical and fairness assessments shall be conducted for all AI agents throughout their lifecycle—especially during initial design, before deployment, and periodically during operations—to identify and mitigate biases, discriminatory outcomes, or unintended ethical consequences.

3.3 AI agents used in critical or sensitive contexts shall provide transparent, understandable explanations of decisions and actions, allowing impacted stakeholders to clearly comprehend the rationale and processes behind significant outcomes.

3.4 AI agents must be designed, developed, and operated explicitly considering potential human rights impacts and harm prevention, consistent with international human rights standards, local laws, and recognized ethical guidelines (e.g., OECD AI Principles, IEEE standards).

3.5 Clearly defined mechanisms must be established for stakeholders impacted by AI decisions to provide feedback, request decision explanations, challenge outcomes, and escalate concerns. These processes must be accessible, documented, and regularly reviewed for effectiveness.

3.6 AI agents shall undergo periodic assessments to ensure compliance with relevant ethical standards, industry best practices, legal requirements (e.g., GDPR, EU AI Act, CCPA), and other regulatory obligations, with documented evidence of adherence.

3.7 An independent Ethics Oversight Committee or Ethics Officer shall provide oversight, guidance, and review of AI agent development, deployment, and operational activities, particularly for high-risk AI scenarios.

3.8 Personnel involved in the AI lifecycle (design, development, deployment, monitoring) must regularly participate in structured training programs covering ethical AI practices, bias detection and mitigation, human rights considerations, and relevant regulatory requirements.

3.9 Ethical assessments, fairness evaluations, stakeholder feedback, regulatory compliance audits, and corrective actions must be systematically documented and transparently reported to internal governance committees and, where applicable, external stakeholders.

Implementation Guidelines

- Incorporate internationally recognized ethical frameworks (e.g., OECD AI Principles, IEEE Ethically Aligned Design) into organizational policies.
- Utilize automated and manual tools for fairness evaluation (e.g., fairness-aware ML frameworks, AI Fairness 360 toolkit).
- Ensure explainability tools or model interpretability techniques are implemented
- Develop accessible stakeholder feedback portals and clearly documented escalation processes to ensure stakeholder trust and responsiveness.

ISO Control Mapping

Standard	Clause / Control	Description
ISO/IEC 42001:2023	• Clause 5.2, Annex A.2.1, A.2.2 • Annex A.3.1, Annex A.9.1	• Ethical AI governance, principles, and policies • Fairness, bias detection, and ethical validation
ISO/IEC 42005:2025	• Core Process (Sections 4.2, 4.3), Annex A • Transparency & Stakeholder Engagement	• Ethical and societal impact assessments • Mechanisms for stakeholder transparency and accountability

4 Risk Management and Impact Assessment

Objective

Establish systematic, comprehensive processes to proactively identify, evaluate, manage, and mitigate security, ethical, compliance, and operational risks, as well as societal impacts associated with AI agents throughout their lifecycle.

Requirements

4.1 A clearly documented AI-specific risk management process shall be established and integrated into the organization's broader enterprise risk management framework. This process must cover AI-related threats, vulnerabilities, ethical concerns, compliance obligations, and potential societal impacts.

4.2 AI Impact Assessments shall be systematically conducted for all AI agents, explicitly evaluating 1) impacts on individuals (e.g., privacy, fairness, discrimination); 2) organizational operations and resources; 3) broader societal implications and stakeholder interests. Assessments must occur during initial design, prior to deployment, upon significant changes, and periodically throughout the AI agent's operational lifecycle.

4.3 As part of the broader risk management process, ethical risk evaluations must be explicitly conducted to identify potential ethical issues, biases, or human rights concerns associated with AI agent use.

4.4 High-risk or high-impact AI agents, as determined through the AI-specific risk and impact assessments, shall require explicit, formal approval by the AI Oversight Committee or designated senior management before deployment.

4.5 Identified AI risks shall be explicitly addressed through documented risk treatment and mitigation plans, including preventive, detective, corrective, and contingency measures, clearly specifying responsibilities and timelines.

4.6 AI agent risks and impacts shall be periodically reviewed and updated at defined intervals and after significant incidents, operational changes, regulatory updates, or environmental shifts.

4.7 All risk assessments, ethical evaluations, impact assessments, mitigation plans, approval decisions, and periodic reviews shall be systematically documented and managed in a centralized repository to ensure auditability, transparency, and compliance oversight.

Implementation Guidelines

- Apply standardized risk management methodologies such as ISO 31000, supplemented by AI-specific threat modeling (e.g., STRIDE).
- Clearly define risk thresholds (low, moderate, high) for categorizing AI agents, aligning appropriate oversight, approvals, and mitigation measures accordingly.
- Engage multidisciplinary expertise (security, ethics, compliance, legal, technical) in comprehensive risk and impact assessment processes.
- Maintain clear documentation templates and checklists to standardize risk and impact assessments organization-wide.

ISO Control Mapping

Standard	Clause / Control	Description
ISO/IEC 27001:2022	• Clause 6.1.2, Annex A.5.4.1, A.5.4.2	• Information security risk assessments and treatments
ISO/IEC 42001:2023	• Clause 6.1, Annex A.5.1, A.5.2	• AI-specific risk identification, assessment, and mitigation
ISO/IEC 42005:2025	• Core Process Sections 4.1, 4.2, Annex A • Documentation & Traceability	• Structured AI Impact Assessments throughout the AI lifecycle • Documenting AI risk and impact decisions and mitigation plans

5 Secure Design and Development

Objective

Integrate security, privacy, ethical considerations, and regulatory compliance systematically into the AI agent design and development phases, ensuring secure and resilient outcomes from the earliest lifecycle stages.

Requirements

5.1 Security and resilience shall be explicitly incorporated into AI agent designs from initial stages. Development practices must comply with recognized secure software engineering standards, including secure coding practices, input validation, robust error handling, and defense against adversarial scenarios.

5.2 A structured, documented threat modeling exercise (e.g., STRIDE methodology) shall be conducted during the design phase to proactively identify, prioritize, and mitigate AI-specific vulnerabilities, risks, and threats.

5.3 AI agents must incorporate privacy-enhancing technologies and methodologies (e.g., data minimization, anonymization, pseudonymization, differential privacy) from initial design, ensuring adherence to relevant privacy regulations (e.g., GDPR, CCPA).

5.4 Design processes must include proactive ethical assessments and fairness evaluations, identifying potential biases or unintended discriminatory outcomes. Mitigation measures shall be explicitly documented and integrated into AI agent development.

5.5 AI models, code artifacts, datasets, and configurations shall be managed under rigorous version control. Clear documentation must be maintained, detailing model lineage, version histories, change logs, and approval records.

5.6 Datasets for AI testing, validation, and fairness evaluations must be representative, protected, regularly updated, and reviewed for biases. Synthetic or anonymized datasets shall be preferred for testing to ensure data privacy.

5.7 AI agents shall undergo documented independent validation and peer review processes prior to deployment, ensuring adherence to security, privacy, ethical requirements, and compliance standards.

5.8 Significant design decisions, assumptions, limitations, ethical considerations, and security controls must be systematically documented, centrally stored, and transparently available for audits and reviews.

Implementation Guidelines

- Utilize secure coding guidelines (e.g., OWASP Top Ten, OWASP AI Security Guidelines) and industry best practices for secure development.
- Engage security and privacy specialists early in the design process to provide targeted threat modeling and privacy input.
- Integrate automated security tools (static/dynamic analysis, dependency scanning, adversarial testing) into development workflows to detect vulnerabilities proactively.
- Clearly define ethical and fairness testing frameworks and integrate fairness metrics into standard testing routines

ISO Control Mapping

Standard	Clause / Control	Description
ISO/IEC 27001:2022	• Annex A.5.8.1, A.8.28.1, A.8.29.1	• Security integration in projects, secure coding, security testing
ISO/IEC 42001:2023	• Annex A.6.2, A.6.2.4, A.8.3.1	• Secure design, secure AI model lifecycle management, AI model robustness
ISO/IEC 42005:2025	• Design-phase Impact Assessments • Documentation & Traceability	• Conducting early ethical and societal impact assessments • Comprehensive documentation of design rationale and ethical impacts

6 Identity, Access, and Privilege Management

Objective

Ensure that AI agents have unique digital identities, appropriately controlled access privileges, and secure authentication mechanisms, operating strictly within the minimum access required to perform their designated tasks securely and effectively.

Requirements

6.1 Each AI agent shall be assigned a unique, traceable digital identity or service account clearly associated with a responsible owner or accountable group, documented within a centralized identity repository.

6.2 AI agent access must strictly adhere to the least privilege principle, granting only the minimal rights and permissions necessary to fulfill the defined operational tasks.

6.3 Access management for AI agents shall employ defined RBAC or ABAC models, clearly documenting assigned roles, attributes, permissions, and the rationale for access provisioning.

6.4 AI agents shall use robust authentication mechanisms (e.g., cryptographic keys, certificates, secure tokens). Authentication credentials must be securely managed, regularly rotated, protected, and promptly revoked upon compromise or decommissioning.

6.5 Periodic, documented reviews and audits of AI agent access privileges shall be conducted to confirm continued appropriateness, compliance with the least privilege principle, and timely revocation of unnecessary privileges.

6.6 AI agents performing sensitive or high-risk functions must implement just-in-time privileged access management, requiring explicit human approval, secondary verification, or multi-factor authorization for sensitive actions.

6.7 Detailed logging of all authentications, access requests, granted privileges, and actions performed by AI agents shall be maintained. While general logging is managed under Section 12 (Logging and Monitoring), this section emphasizes logging specifically related to identity, access, and privilege activities.

6.8 Documentation related to AI agent identities, access rights, privilege approvals, and review activities must be centrally managed, ensuring transparency, auditability, and compliance oversight.

Implementation Guidelines

- Integrate AI agent identity management into centralized Identity and Access Management (IAM) systems to maintain oversight and control.
- Utilize Privileged Access Management (PAM) solutions and secrets-management platforms (e.g., HashiCorp Vault, Azure Key Vault, AWS Secrets Manager) for securely managing AI agent credentials and privileged actions.
- Automate periodic access reviews and credential rotation processes to reduce human error and streamline privilege management.
- Ensure clearly defined escalation procedures for unusual privilege or access events detected through logging

ISO Control Mapping

Standard	Clause / Control	Description
ISO/IEC 27001:2022	• Annex A.5.15.1, A.5.16.1, A.5.17.1, A.5.18.1, A.8.2.1	• Access control policies, identity lifecycle, authentication controls, access reviews, privileged access
ISO/IEC 42001:2023	• Annex A.6.2.4, Annex A.8.3.2	• Secure AI agent identity management, technical security controls
ISO/IEC 42005:2025	• Accountability and Traceability	• Documented access control, traceability, and compliance of AI decisions

7 Cryptographic Controls for AI Models and Components

Objective

Ensure the confidentiality, authenticity, integrity, and resilience of AI models, configurations, data, and associated components by applying robust, standardized, and secure cryptographic controls throughout their lifecycle.

Requirements

7.1 AI models, associated code, and configuration artifacts shall be protected using digital signatures and integrity checks (e.g., cryptographic hashes), ensuring detection of unauthorized alterations or tampering.

7.2 AI models and sensitive AI-related artifacts stored within the organization shall be encrypted at rest, employing robust encryption algorithms compliant with industry standards and regulatory requirements.

7.3 Transmission of AI models, data, and related artifacts across networks must employ secure communication protocols (e.g., TLS) with strong encryption algorithms and secure key exchanges.

7.4 Cryptographic keys used for securing AI models and components must be securely generated, managed, stored, rotated regularly, and revoked when compromised or obsolete, following documented key management procedures.

7.5 AI agents shall implement cryptographic validation checks at initialization, load, or execution time, ensuring only authorized and verified models and components are utilized.

7.6 Sensitive or proprietary AI models must employ cryptographic watermarking or similar techniques, embedding traceable and verifiable identifiers to detect unauthorized use, copies, or distribution.

7.7 Logs must be maintained for critical cryptographic activities (e.g., key usage, key rotation, digital signing events), centrally managed and protected from unauthorized access, ensuring transparency and accountability.

7.8 The organization shall proactively evaluate and integrate cryptographic agility and quantum-resistant algorithms to ensure the long-term resilience and security of AI models and associated cryptographic mechanisms.

Implementation Guidelines

- Use standardized, industry-accepted cryptographic algorithms (e.g., AES-256, RSA-2048 or higher, ECC, SHA-256 or higher).
- Integrate cryptographic operations with centralized, secure Key Management Systems (KMS) or Hardware Security Modules (HSMs).
- Regularly assess cryptographic algorithms and configurations against current threats and emerging technologies (e.g., quantum computing threats).
- Document and centralize cryptographic policy, key custodianship, rotation schedules, and incident handling procedures

ISO Control Mapping

Standard	Clause / Control	Description
ISO/IEC 27001:2022	• Annex A.8.24.1, A.8.25.1	• Cryptographic protection, key management controls
ISO/IEC 42001:2023	• Annex A.8.3.2, Annex A.8.4.2	• Secure AI model cryptographic protection and component integrity
ISO/IEC 42005:2025	• Security of AI Configurations	• Cryptographic and integrity protections throughout AI lifecycle

8 Segregation of Duties and Human Oversight

Objective

Establish clear segregation of duties among AI agents and human personnel, accompanied by explicit human oversight mechanisms, ensuring accountability, transparency, and controlled execution of high-risk or critical AI-driven tasks.

Requirements

8.1 Critical functions related to AI agent lifecycle—such as design, development, deployment, monitoring, and approval—shall have clear separation among roles and responsibilities, preventing any single individual or AI agent from independently controlling the entire process.

8.2 AI agents performing critical or sensitive tasks shall be explicitly segregated, preventing agents that develop or configure processes from autonomously approving or deploying them, thereby avoiding potential conflicts or misuse.

8.3 Tasks identified as high-risk, sensitive, or having significant ethical, operational, or compliance impacts must include clearly documented human oversight mechanisms, including 1) human-in-the-loop: mandatory human approval required before executing critical decisions; and 2) human-on-the-loop: continuous or periodic human monitoring with predefined intervention triggers and procedures.

8.4 Clearly documented approval workflows shall be implemented for high-risk AI-driven actions, requiring explicit human approval or secondary verification by authorized personnel prior to execution.

8.5 Automated alerts and escalation mechanisms shall promptly notify human overseers if AI agents deviate from predefined operational, ethical, or performance thresholds, triggering timely human review and corrective action.

8.6 All human interventions, approvals, oversight activities, and escalation events shall be systematically documented, including timestamp, decision rationale, involved personnel, and outcomes, ensuring full auditability.

8.7 Personnel responsible for AI oversight and intervention shall regularly receive targeted training on identifying anomalous behaviors, executing timely interventions, and effectively managing escalation processes.

8.8 Human oversight records, approvals, and escalation documentation must be centrally managed, facilitating transparency, accountability, and compliance assurance.

Implementation Guidelines

- Clearly define roles and responsibilities in organizational charts, job descriptions, and procedural documentation to ensure effective segregation.
- Utilize automated workflow and approval management systems (e.g., ServiceNow, Jira, approval frameworks) for streamlined, auditable human oversight processes.
- Regularly review escalation and intervention criteria, adjusting thresholds based on operational insights, ethical considerations, and stakeholder feedback.
- Conduct periodic oversight drills or tabletop exercises to verify human overseers' readiness and procedural effectiveness

ISO Control Mapping

Standard	Clause / Control	Description
ISO/IEC 27001:2022	• Annex A.5.3.1, A.5.11.1	• Segregation of duties, accountability, and monitoring
ISO/IEC 42001:2023	• Annex A.3.2.2, A.6.3.1	• Explicit human oversight responsibilities, approval mechanisms for AI
ISO/IEC 42005:2025	• Mitigation Planning and Control Design • Accountability & Control Traceability	• Documentation of human interventions and oversight

9 Pre-Deployment Testing and Approval

Objective

Ensure rigorous testing, validation, and formal approval processes are systematically conducted prior to AI agent deployment, verifying security, ethical compliance, resilience, accuracy, and operational readiness, thus minimizing risks associated with AI operational environments.

Requirements

9.1 Prior to deployment, AI agents must undergo documented security testing, including 1) vulnerability assessments; 2) penetration and adversarial input testing; 3) robustness evaluation against known and potential threats.

9.2 AI models shall be validated using representative and high-quality test datasets, clearly defined acceptance criteria, and recognized statistical performance benchmarks (e.g., accuracy, precision, recall).

9.3 AI agents shall undergo adversarial scenario testing, stress testing, and resilience assessments to evaluate and ensure robustness against adversarial attacks, manipulation, unexpected inputs, and operational disruptions.

9.4 Prior to deployment, AI agents must complete documented ethical validations and impact assessments, verifying alignment with the organization's ethical standards, fairness principles, regulatory requirements, and stakeholder expectations.

9.5 High-risk or high-impact AI agents shall require explicit, formal approval by the AI Oversight Committee or senior management, with documented risk acceptance prior to deployment into operational environments.

9.6 Comprehensive documentation—including security tests, ethical evaluations, impact assessments, validation outcomes, and formal approvals—must be systematically maintained in a centralized repository to support auditability, transparency, and traceability.

9.7 Clear rollback procedures and contingency plans shall be documented and tested prior to deployment, ensuring the ability to swiftly revert to a stable state if significant security, ethical, or operational issues are identified post-deployment.

Implementation Guidelines

- Employ standardized testing frameworks and tools (e.g., automated vulnerability scanners, penetration testing, AI model robustness frameworks).
- Integrate AI-specific resilience and adversarial testing into organizational testing methodologies.
- Establish clear risk acceptance thresholds and approval criteria, documented and communicated to relevant governance committees and technical teams.
- Regularly conduct rollback drills to ensure contingency plans are operationally effective.

ISO Control Mapping

Standard	Clause / Control	Description
ISO/IEC 27001:2022	• Annex A.8.28.1, A.8.29.1, A.8.32.1	• Secure coding and testing, operational system testing, separation of environments
ISO/IEC 42001:2023	• Clause 8.3.1, Annex A.9.2.1	• AI system validation and verification, explainability, bias, and ethical testing
ISO/IEC 42005:2025	• Mitigation Planning and Control Design • Accountability & Control Traceability	• Documented impact assessments and ethical validation prior to operational deployment

10 Deployment and Configuration Hardening

Objective

Ensure secure, consistent, and resilient deployment and configuration of AI agents, minimizing security vulnerabilities and operational risks through rigorous adherence to defined hardening baselines and secure deployment practices.

Requirements

10.1 AI agents shall be deployed in secure, isolated, and controlled environments explicitly separated from development, testing, and other operational systems, reducing risk of unauthorized interactions and access.

10.2 AI agent configurations—including underlying infrastructure, operating systems, containers, virtual environments—must adhere to documented, standardized hardening baselines. These shall include 1) disabling unnecessary services, ports, and functionalities; 2) enforcing secure defaults; 3) implementing security patches and updates promptly.

10.3 Deployments shall utilize Infrastructure-as-Code (IaC) practices, ensuring consistency, repeatability, traceability, and secure configuration management. IaC scripts and deployment artifacts must be securely managed and version-controlled.

10.4 AI agents must be deployed strictly with essential functionalities required to fulfill their defined tasks. Any additional functionality or services shall require explicit justification and formal approval.

10.5 Secure provisioning processes shall be implemented, including 1) secure boot mechanisms; 2) cryptographic validation of deployment artifacts; 3) secure injection of configuration parameters and credentials.

10.6 All deployment and configuration changes must follow formal, documented change management processes, including impact analysis, security review, documented approvals, deployment validation, and rollback readiness.

10.7 Continuous monitoring shall be rigorously applied immediately following AI agent deployment (initial stabilization period), rapidly detecting configuration errors, operational anomalies, and security issues. This differs from ongoing operational drift monitoring (Section 13).

10.8 Deployment and configuration activities, decisions, validations, and approvals must be clearly documented and managed centrally, ensuring auditability, compliance, and traceability.

Implementation Guidelines

- Leverage deployment automation tools (e.g., Kubernetes, Terraform, Ansible) to enforce standardized, repeatable, and secure deployments.
- Regularly audit and update hardened configuration baselines to address emerging threats and vulnerabilities.
- Integrate Continuous Integration/Continuous Deployment (CI/CD) pipelines with built-in security scanning, automated testing, and rollback capabilities.
- Conduct periodic configuration audits and assessments against documented hardening standards to maintain compliance and identify deviations.

ISO Control Mapping

Standard	Clause / Control	Description
ISO/IEC 27001:2022	• Annex A.8.9.1, A.8.10.1, A.8.32.1	• Configuration management, secure disposal of deployment artifacts, separation of operational environments
ISO/IEC 42001:2023	• Annex A.6.2.4, Annex A.8.3.2	• Secure AI model deployment, technical and operational security measures
ISO/IEC 42005:2025	• Integration with Controls	• Secure integration and deployment controls based on documented impact assessments

11 Third-Party and Supply Chain Controls for AI Components

Objective

Establish robust governance, security, and compliance measures over third-party AI components, suppliers, and related supply chains, effectively managing risks throughout procurement, integration, operation, and ongoing lifecycle management.

Requirements

11.1 Prior to procurement or integration, AI components sourced from third parties must undergo formal, documented risk assessments, evaluating supplier reliability, security posture, ethical alignment, regulatory compliance, and operational resilience.

11.2 Supplier contracts must explicitly document security, ethical, data privacy, compliance obligations, data ownership rights, incident reporting requirements, audit rights, and clearly defined service-level agreements (SLAs).

11.3 Third-party AI components shall be securely integrated into organizational processes following 1) security testing and vulnerability scanning; 2) cryptographic integrity verification; 3) compliance checks with organizational standards.

11.4 Clear traceability records must be maintained, documenting sources, responsible suppliers, origins of AI components, and history of changes or updates, facilitating transparency and auditability.

11.5 Third-party AI components and suppliers shall undergo ongoing security, compliance, and performance monitoring, including regular reassessments, vulnerability tracking, and risk evaluations to detect emerging threats promptly.

11.6 Clear incident response processes and communication channels shall be established with third-party suppliers, ensuring coordinated, timely management and reporting of incidents involving externally sourced AI components.

11.7 Formal exit strategies and transition plans must be developed for critical third-party AI suppliers or components, detailing secure transition, data deletion procedures, continuity measures, and risk mitigation in case of contract termination or supplier replacement.

11.8 All documentation related to supplier assessments, contracts, integration validation, monitoring activities, and incident coordination shall be managed in a centralized repository, ensuring traceability, transparency, and compliance.

Implementation Guidelines

- Utilize standardized third-party risk management frameworks (e.g., ISO/IEC 27036, SOC 2 audits) for evaluating supplier maturity and control effectiveness.
- Regularly conduct supplier audits and compliance assurance exercises to maintain oversight.
- Integrate supplier management and monitoring into enterprise risk management and procurement processes, supported by automation tools and threat intelligence feeds.
- Develop clearly documented and regularly reviewed incident response procedures with third-party suppliers

ISO Control Mapping

Standard	Clause / Control	Description
ISO/IEC 27001:2022	• Annex A.5.19.1, A.5.20.1, A.5.21.1, A.5.23.1	• Supplier relationships, contracts, monitoring, cloud and third-party component security
ISO/IEC 42001:2023	• Annex A.8.4.1, A.8.4.2	• Supplier management, third-party AI component controls and security
ISO/IEC 42005:2025	• Model Source Review	• Third-party AI sourcing impact assessments and continuous evaluations

12 AI Data Protection, Classification, and Privacy

Objective

Establish robust measures for protecting, classifying, and managing data utilized or generated by AI agents, ensuring compliance with privacy regulations, data security standards, and ethical principles throughout the entire AI lifecycle.

Requirements

12.1 Data utilized by AI agents shall be classified according to organizational standards, with clearly documented handling guidelines per classification, specifying required protective measures and restrictions.

12.2 AI agents shall collect, process, retain, and access only the minimum data necessary for their explicitly defined operational purposes. Data not essential must be promptly anonymized, aggregated, or securely deleted.

12.3 Privacy-enhancing measures (e.g., anonymization, pseudonymization, differential privacy, secure multi-party computation) shall be systematically embedded into AI agent designs and operational processes, ensuring adherence to applicable privacy regulations (GDPR, CCPA, PIPEDA).

12.4 Where applicable, explicit and documented consent processes (or alternative lawful bases) must be established and transparently communicated for AI-driven personal data processing, aligning with relevant privacy regulations.

12.5 Data handled by AI agents must be encrypted at rest and during transit, employing secure encryption standards compliant with organizational cryptographic policies and regulatory requirements.

12.6 Data processed or generated by AI agents must follow defined retention schedules compliant with regulatory requirements and organizational policies. Upon expiry, data shall be securely disposed of using approved methods (cryptographic erasure, overwriting, physical destruction).

12.7 Regular PIAs must be conducted throughout the AI lifecycle, explicitly evaluating and documenting potential privacy risks, impacts, mitigation measures, and compliance actions.

12.8 Clear processes for promptly detecting, containing, managing, and reporting AI-related data breaches or privacy incidents shall be established, complying with regulatory requirements and internal incident response frameworks.

12.9 Documentation related to data classification, privacy assessments, retention schedules, breach responses, and privacy compliance activities shall be centrally managed, ensuring auditability and transparency.

Implementation Guidelines

- Align AI data protection practices explicitly with relevant global privacy standards and regulations (e.g., GDPR, CCPA, PIPEDA, HIPAA).
- Employ Privacy-Enhancing Technologies (PETs) systematically throughout AI processes, particularly during design and data-processing phases.
- Regularly train development, operational, and compliance teams on evolving privacy requirements, secure data handling, and ethical implications of data usage.
- Integrate privacy compliance monitoring and automated data management solutions to ensure real-time adherence to defined privacy practices and policies.

ISO Control Mapping

Standard	Clause / Control	Description
ISO/IEC 27001:2022	• Annex A.5.10.1, A.8.3.1, A.8.7.1	• Information classification, handling, data encryption in transit and at rest
ISO/IEC 42001:2023	• Annex A.7.3.2, Annex A.8.1.1	• Data protection and privacy controls for AI datasets, privacy by design
ISO/IEC 42005:2025	• Transparency & Risk to Individuals	• Structured Privacy Impact Assessments and compliance

13 AI Logging, Monitoring, Transparency, and Auditability

Objective

Establish comprehensive logging, continuous monitoring, transparency, and auditability mechanisms for AI agents, ensuring traceable, explainable, and accountable operation to facilitate rapid detection, incident response, stakeholder trust, and regulatory compliance.

Requirements

13.1 AI agents shall maintain detailed, secure logs recording: 1) inputs received, outputs generated, decisions made, and associated context; 2) interactions with users, systems, and processes, including identities involved; 3) date/time stamps and model versions used.

13.2 AI agent operations shall be continuously monitored with automated tools designed specifically to detect anomalous behaviors, deviations from performance norms, unexpected outcomes, and security-related events. *(Differentiated clearly from operational drift monitoring in Section 13.)*

13.3 AI agents performing critical or sensitive tasks shall provide transparent explanations of their decision-making logic, enabling stakeholders and auditors to understand decision rationale and process.

13.4 Logs must be centrally aggregated, protected against unauthorized access, tampering, and deletion, and securely retained per defined retention policies, ensuring ongoing integrity, confidentiality, and availability.

13.5 The organization shall maintain comprehensive records linking AI decisions to specific models, configurations, and data states, ensuring clear traceability for investigations, audits, and regulatory compliance.

13.6 Logs generated by AI agents shall be regularly reviewed and analyzed to identify potential security incidents, compliance violations, or operational anomalies. Key insights shall be reported promptly to governance and oversight committees.

13.7 The organization shall regularly produce transparency reports detailing AI agent performance metrics, accuracy levels, identified biases, incidents, ethical assessments, and compliance statuses, enhancing stakeholder trust and accountability.

13.8 All logging standards, monitoring practices, explainability mechanisms, and auditability documentation shall be systematically centralized to facilitate transparency, compliance, and effective governance oversight.

Implementation Guidelines

- Utilize robust, secure logging and monitoring platforms (SIEM, log analytics tools, observability platforms) tailored to AI-specific event detection and anomaly identification.
- Define clear anomaly detection thresholds, escalation protocols, and response procedures specific to AI-driven systems.
- Employ explainability and interpretability frameworks for AI models (e.g., SHAP, LIME, integrated explainability methods) to facilitate transparent decision documentation.
- Ensure periodic independent audits of logging, monitoring, and transparency practices to validate operational effectiveness and regulatory compliance.

ISO Control Mapping

Standard	Clause / Control	Description
ISO/IEC 27001:2022	• Annex A.8.16.1, A.8.30.1, A.5.22.1	• System activity logging, protection of log information, accountability and traceability
ISO/IEC 42001:2023	• Clause 9.1, Annex A.9.2.1, Annex A.10.2.1	• AI monitoring, explainability and transparency, continuous improvement
ISO/IEC 42005:2025	• Monitoring & Review	• Continuous monitoring, explainability, transparency in AI decisions

14 Operational Security and Change Control for AI Agents

Objective

Maintain secure, stable, and resilient operational environments for AI agents by rigorously managing operational security risks, configuration changes, and ongoing operational drift, ensuring continuous performance alignment, resilience, and security.

Requirements

14.1 Implement continuous monitoring solutions specifically designed to detect operational drift, such as performance degradation, model accuracy reduction, unintended bias shifts, or gradual changes in operational behaviors of AI agents.

14.2 All operational changes, updates, model patches, or configuration adjustments for AI agents must follow documented, formal change management procedures, including: 1) impact assessments; 2) security and ethical evaluations; 3) testing and validation prior to deployment; 4) formal documented approvals.

14.3 AI agents and their operational environments must undergo regular vulnerability scanning, assessment, and prompt patching or updating of vulnerabilities based on a risk-prioritized schedule.

14.4 Clearly documented rollback procedures and contingency plans must be regularly tested and maintained, enabling rapid recovery to a stable operational state in case of operational anomalies or failed changes.

14.5 Automated checks shall regularly validate AI agent configuration integrity, detecting unauthorized or unintended modifications promptly and triggering alerts and corrective actions.

14.6 Operational anomalies, significant configuration changes, or identified operational drift events must be promptly communicated to relevant stakeholders, documenting root cause analyses, impact assessments, corrective actions, and preventive measures.

14.7 Insights gained from operational security incidents, drift detections, configuration issues, and audits shall be systematically integrated into continuous improvement processes, enhancing operational resilience and security posture.

14.8 Documentation related to operational security activities, drift monitoring results, change management records, vulnerability assessments, and patch management actions must be maintained in a centralized, accessible repository for auditability and governance oversight.

Implementation Guidelines

- Use specialized tools (e.g., drift-detection frameworks, ML observability tools) to actively monitor AI agent performance and operational health.
- Establish well-defined, automated, and audited CI/CD pipelines that incorporate security scans, testing, validation, and rollback capabilities to manage secure operational changes.
- Regularly update vulnerability management practices in alignment with emerging threats and standards.
- Conduct regular drills of rollback and contingency procedures to validate their operational effectiveness.

ISO Control Mapping

Standard	Clause / Control	Description
ISO/IEC 27001:2022	• Annex A.8.23.1, A.8.29.1, A.8.31.1	• Technical vulnerability management, security testing, change management
ISO/IEC 42001:2023	• Annex A.8.4.2, Annex A.10.2.1	• Operational security measures for AI, continuous improvement through operational insights
ISO/IEC 42005:2025	• Change-Triggered Reassessments	• Operational change and impact reassessments

15 AI Incident Management and Response

Objective

Establish structured, robust incident management and response capabilities explicitly tailored for AI-related incidents, enabling rapid identification, containment, mitigation, recovery, communication, and ongoing improvements following AI-specific security, ethical, privacy, or compliance incidents.

Requirements

15.1 Proactive measures must be implemented for early detection of AI-specific incidents, including model compromise, adversarial attacks, significant ethical breaches, unintended biases, privacy violations, compliance failures, or unauthorized model manipulation.

15.2 AI incidents shall be formally classified and prioritized based on clearly defined severity and impact criteria, including operational, regulatory, ethical, financial, reputational, and stakeholder impacts.

15.3 Documented procedures explicitly designed for AI incident response shall cover: 1) immediate containment actions; 2) incident impact assessment and root-cause analysis; 3) corrective measures, model rollback, and operational recovery; 4) regulatory notifications and stakeholder communications.

15.4 Clear roles and responsibilities shall be assigned for AI incident response, including incident coordinators, technical specialists, compliance and ethical advisors, forensic analysts, and communications leads.

15.5 AI incident response shall include clearly documented escalation processes and communication protocols for rapid notification of stakeholders, regulators, affected parties, and internal governance bodies.

15.6 Detailed logging and audit trails of AI incidents shall be securely maintained, facilitating forensic analysis, supporting regulatory audits, and enabling accurate post-incident reviews.

15.7 Systematic post-incident reviews shall be conducted for AI incidents, clearly documenting findings, corrective actions, lessons learned, and recommendations for improvements, integrated into ongoing AI governance and risk management processes.

15.8 Regular AI-specific incident response simulations, drills, or tabletop exercises shall be conducted to ensure incident preparedness, team effectiveness, and procedural efficiency in real-world scenarios.

15.9 All documentation related to AI incidents—including incident logs, response actions, notifications, forensic analyses, and improvement actions—shall be centrally managed, transparent, and readily accessible for compliance and audit purposes.

Implementation Guidelines

- Integrate AI-specific incident response procedures within existing organizational incident management and cybersecurity frameworks, customizing for unique AI contexts.
- Leverage automation tools (e.g., SOAR platforms) and incident management systems to streamline detection, escalation, containment, and notification processes.
- Conduct regular specialized AI incident training for response teams to ensure preparedness and clarity of roles during incidents.
- Ensure clearly defined communication templates and escalation pathways to enable rapid stakeholder and regulatory communication during incidents.

ISO Control Mapping

Standard	Clause / Control	Description
ISO/IEC 27001:2022	• Annex A.5.25.1, A.5.26.1, A.8.27.1	• Incident management, response planning, forensic logging
ISO/IEC 42001:2023	• Clause 6.3.1, Annex A.9.3.1	• AI-specific incident management and response
ISO/IEC 42005:2025	• Response to Negative Impacts	• Incident management integrated with AI impact monitoring

16 Decommissioning, End-of-Life, and Knowledge Retention

Objective

Ensure secure, structured, and compliant processes for the retirement and decommissioning of AI agents, safeguarding data, revoking accesses, maintaining compliance, and systematically preserving knowledge and lessons learned for continuous improvement.

Requirements

16.1 The organization shall establish clear, documented procedures for secure AI agent decommissioning, including safe shutdown, data disposal, access revocation, component sanitization, and resource recovery.

16.2 AI-generated or processed data shall follow explicit retention schedules compliant with regulatory obligations. Data no longer required must be securely deleted through approved methods (e.g., cryptographic erasure, overwriting, physical destruction).

16.3 All digital identities, privileges, and accesses associated with decommissioned AI agents must be immediately revoked and securely deleted to prevent unauthorized residual access.

16.4 Hardware, software, cloud resources, and integrations associated with retired AI agents shall be securely sanitized, repurposed, or securely disposed of, ensuring no residual data or artifacts remain.

16.5 Key operational insights, performance metrics, incident histories, risk evaluations, compliance records, and ethical assessments related to AI agents shall be systematically documented, retained, and shared internally to inform future AI initiatives and improvements.

16.6 Clearly defined processes shall guide communication and notification of stakeholders, impacted parties, and regulators during AI agent retirement, detailing reasons, impacts, and transitional measures being taken.

16.7 A comprehensive final impact and risk assessment must be conducted at the retirement phase, confirming that no unresolved risks, compliance issues, or adverse impacts remain post-decommissioning.

16.8 All records relating to AI agent retirement—including decommissioning procedures, data disposal certificates, access revocation records, component sanitization documentation, and knowledge retention artifacts—must be maintained centrally, ensuring auditability and governance oversight.

Implementation Guidelines

- Develop standardized checklists and templates for secure AI agent decommissioning, clearly specifying responsible roles and actions.
- Regularly audit decommissioning and retirement processes to confirm adherence to defined procedures and compliance obligations.
- Ensure comprehensive and centralized documentation repositories to retain knowledge and lessons learned for future reference and improvement.
- Maintain clear communication strategies and notification templates to effectively manage stakeholder expectations and compliance obligations during AI agent retirement.

ISO Control Mapping

Standard	Clause / Control	Description
ISO/IEC 27001:2022	• Annex A.8.10.1, A.5.37.1	• Secure disposal of assets, system decommissioning
ISO/IEC 42001:2023	• Clause 10, Annex A (Lifecycle Retirement)	• Retirement procedures for AI systems
ISO/IEC 42005:2025	• Closure & Documentation	• AI system closure, impacts, residual risk documentation

17 Continuous Improvement and Compliance Maintenance

Objective

Establish an ongoing, structured approach for continuous improvement, compliance assurance, and maturity enhancement of AI agent security, governance, and operational practices, ensuring sustained effectiveness, adaptability to evolving threats, technologies, and regulatory landscapes.

Requirements

17.1 Periodic, documented compliance assessments and performance evaluations of AI agent security controls and governance practices shall be systematically conducted to proactively identify gaps, nonconformities, emerging risks, and improvement opportunities.

17.2 Formal continuous improvement mechanisms shall systematically incorporate insights from audits, incidents, feedback loops, operational performance monitoring, regulatory updates, and ethical assessments to iteratively refine AI agent governance and controls.

17.3 The organization shall actively monitor the external environment—including emerging security threats, technological advancements, ethical developments, and regulatory changes—and integrate these insights promptly into AI governance and operational practices.

17.4 Structured processes must be implemented to systematically collect, analyze, and incorporate feedback from stakeholders, operational teams, compliance bodies, auditors, regulators, and users, enhancing trust and continuous responsiveness.

17.5 Training and awareness programs shall be continuously updated and regularly delivered to relevant personnel, incorporating emerging security practices, ethical developments, compliance requirements, operational lessons learned, and improvement insights.

17.6 Regular reviews of AI-related documentation—including policies, procedures, standards, assessments, incident reports, and compliance documentation—shall be conducted to ensure currency, clarity, accuracy, and relevance to evolving practices.

17.7 Systematic maturity assessments shall be conducted at defined intervals, objectively benchmarking organizational AI security and governance practices against industry best practices, recognized frameworks (e.g., NIST AI RMF, ISO standards), and peer organizations, establishing clear maturity targets and strategic improvement roadmaps.

17.8 All documentation related to continuous improvement activities—including compliance reviews, performance evaluations, stakeholder feedback, maturity assessments, and training records—shall be centrally maintained, accessible, and auditable.

Implementation Guidelines

- Conduct regular structured review cycles (e.g., quarterly, annually) of compliance and performance against defined AI security frameworks and standards.
- Utilize automated dashboards and compliance monitoring tools to track compliance status and improvement activities in real-time.
- Clearly define roles and responsibilities for monitoring external threats, technological developments, and regulatory updates, ensuring prompt integration into organizational practices.
- Engage diverse stakeholders through regular structured feedback sessions, surveys, or consultations to validate effectiveness and continuously refine AI governance practices.

ISO Control Mapping

Standard	Clause / Control	Description
ISO/IEC 27001:2022	• Clause 10.2, Annex A.5.35	• Continual improvement of ISMS, regular reviews
ISO/IEC 42001:2023	• Clause 10, Annex A.10.2.1	• Continuous improvement of AI management systems
ISO/IEC 42005:2025	• Continuous Risk Evaluation	• Ongoing AI impact and compliance evaluation, continuous improvement integration