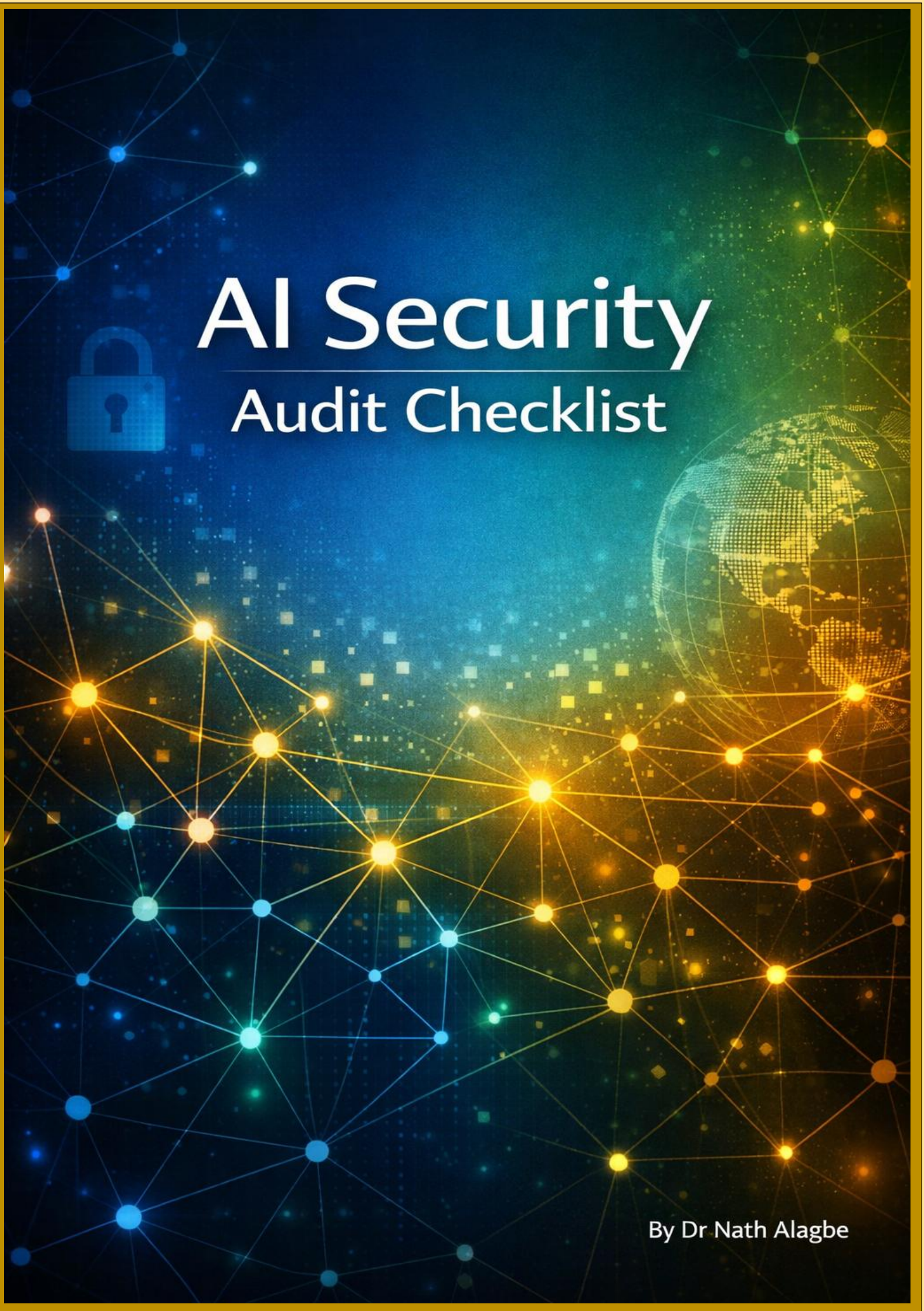

AI Security
Audit Checklist
By Dr Nath Alagbe

AI Security Audit Checklist: Practitioner-Grade Controls

This checklist is structured into ten distinct audit domains, each containing actionable, testable audit questions, validation methods, required evidence, and explicit references to authoritative standards and frameworks.

Author: Dr. Nath Alagbe; CISA, CISM, CISSP, CRISC, CCAK, AAIA, CCEP, CFE, FCA, FCTI, FFAR, MBA, PhD

Domain 1: AI Governance and Accountability

This domain assesses the organizational structure, policies, and processes that ensure responsible and accountable development and deployment of AI systems.

Audit Area	Audit Question	Validation Method	Evidence Required	Auditor Comments with references to standards and frameworks
Organizational Structure	Is there a formally approved AI Governance Policy that defines the organization's risk appetite, ethical principles, and accountability structure for AI systems?	Review of the AI Governance Policy document; Interview with the Chief AI Officer (or equivalent).	Approved AI Governance Policy (version-controlled); Organizational chart showing the AI governance body (e.g., AI Steering Committee); Charter/Terms of Reference for the governance body.	**ISO/IEC 42001:2023** (Clause 5.2, 5.3 - Policy and Roles); **NIST AI RMF** (Govern function).
Roles & Responsibilities	Are roles, responsibilities, and authorities for AI system ownership, risk management, and compliance clearly defined, documented, and communicated across the organization?	Review of job descriptions and RACI matrix for AI projects; Interview with project leads and compliance officers.	Documented roles and responsibilities matrix for AI lifecycle; Training records confirming communication of responsibilities.	**ISO/IEC 27002:2022** (5.2 - Roles and responsibilities); **SOC 2** (Control Environment).
Ethical Principles	Has the organization established and documented a set of ethical principles for AI use, and are these principles integrated into the AI system design and review process?	Review of the AI Ethics Policy; Examination of AI System Impact Assessment (AIA) templates.	AI Ethics Policy document; Records of ethical review sign-offs for high-risk AI systems.	**NIST AI RMF** (Govern function, specifically addressing societal and ethical risks); **ISO/IEC 42001:2023** (6.1.2 - AI system impact assessment).

Domain 2: AI Risk Management and Oversight

This domain focuses on the systematic identification, analysis, evaluation, and treatment of risks specific to AI systems.

Audit Area	Audit Question	Validation Method	Evidence Required	Auditor Comments with references to standards and frameworks
Risk Assessment	Is a formal, documented AI-specific risk assessment performed for all new and significantly modified AI systems, covering technical, ethical, legal, and operational risks?	Review of the AI Risk Assessment methodology and recent assessment reports.	AI Risk Assessment reports (including residual risk); Risk register entries for AI systems; Sign-off by risk owners.	**ISO/IEC 23894:2023** (AI risk management principles); **NIST AI RMF** (Map and Measure functions).
Risk Treatment	Are identified AI risks treated with appropriate controls, and is the effectiveness of these controls periodically reviewed and documented?	Examination of risk treatment plans and control effectiveness testing results.	Documented risk treatment plans; Control effectiveness testing reports; Evidence of management review of residual risk.	**ISO/IEC 27001:2022** (6.1.3 - Statement of Applicability); **NIST SP 800-37** (Risk Management Framework).
Impact Assessment	Does the organization conduct an AI System Impact Assessment (AIA) to evaluate potential negative impacts on individuals, groups, and fundamental rights before deployment?	Review of AIA documentation for a sample of AI systems.	Completed AIA forms, including mitigation strategies for identified negative impacts.	**GDPR** (Article 35 - Data Protection Impact Assessment, extended for AI); **ISO/IEC 42001:2023** (6.1.2 - AI system impact assessment).

Domain 3: Model Lifecycle Security and Change Management

This domain covers the security controls applied throughout the entire AI/ML model development lifecycle (MDLC) and the process for managing changes to models and their environments.

Audit Area	Audit Question	Validation Method	Evidence Required	Auditor Comments with references to standards and frameworks
Secure MDLC	Are security requirements (e.g., threat modeling, secure coding practices) integrated into each phase of the Model Development Lifecycle (MDLC), from design to deployment?	Review of MDLC documentation and security gate checklists; Interview with MLOps and development teams.	MDLC documentation with embedded security checkpoints; Threat modeling reports for high-risk models.	**NIST SP 800-218** (Secure Software Development Framework); **ISO/IEC 27002:2022** (8.28 - Secure coding).
Model Versioning	Is a robust model versioning and artifact management system in place to ensure traceability, reproducibility, and integrity of all model	Inspection of the Model Registry/Artifact Store and review of the CI/CD pipeline configuration.	Model registry logs showing version history and metadata; Hashing/checksum validation records for model artifacts.	**ISO/IEC 42001:2023** (8.3.3 - Traceability of AI systems); **SOC 2** (System Operations).

	components (code, data, configuration)?			
Change Management	Is there a formal change management process for model updates, retraining, and deployment that includes security review, testing, and rollback capabilities?	Review of change tickets for recent model deployments; Examination of pre-deployment testing reports.	Change Management records (e.g., JIRA tickets) showing security sign-off; Rollback procedure documentation.	**ISO/IEC 27002:2022** (8.32 - Change management); **NIST SP 800-53** (CM-3 - Configuration Change Control).

Domain 4: Training, Validation, and Inference Data Security

This domain focuses on the security, privacy, and integrity of the data used to train, validate, and operate AI models.

Audit Area	Audit Question	Validation Method	Evidence Required	Auditor Comments with references to standards and frameworks
Data Provenance	Is the provenance (source, collection method, licensing) of all training and validation data documented, and are controls in place to prevent the use of unauthorized or poisoned data?	Review of data catalog and data acquisition records; Examination of data ingestion pipeline security controls.	Data Provenance documentation; Data licensing agreements; Data quality and integrity checks in the pipeline.	**ISO/IEC 42001:2023** (8.3.2 - Data for AI systems); **OWASP Top 10 for LLMs** (TDP - Training Data Poisoning).
Data Privacy	Are appropriate privacy-enhancing technologies (PETs) or anonymization techniques applied to sensitive data before it is used for model training, and is access strictly controlled?	Review of data masking/anonymization scripts; Inspection of access control lists (ACLs) for data stores.	Data anonymization/masking policy; Evidence of differential privacy or k-anonymity implementation; Data access logs.	**GDPR** (Principle of Data Minimisation and Pseudonymisation); **ISO/IEC 27002:2022** (8.10 - Data masking).
Inference Data	Are input and output data at the inference stage validated for integrity and sanitized to prevent injection attacks or data leakage?	Review of API gateway and model input validation code; Penetration testing reports focused on inference endpoints.	Input validation and sanitization code snippets; Logs showing rejection of malicious inputs.	**OWASP Top 10 for LLMs** (PI - Prompt Injection; IO - Insecure Output Handling); **NIST SP 800-53** (SI-10 - Information Input Validation).

Domain 5: Model Integrity, Robustness, and Adversarial Resistance

This domain assesses the AI model's technical security, focusing on its resilience against malicious attacks and its ability to maintain its intended performance.

Audit Area	Audit Question	Validation Method	Evidence Required	Auditor Comments with references to standards and frameworks
Adversarial Testing	Are AI systems subjected to dedicated adversarial robustness testing (e.g., evasion, poisoning, model inversion attacks) before deployment, and are mitigation strategies documented?	Review of adversarial testing methodology and reports.	Adversarial testing reports (e.g., using tools like Adversarial Robustness Toolbox); Documentation of model hardening techniques.	**OWASP Top 10 for LLMs** (MA - Model Abuse; MI - Model Denial of Service); **NIST AI RMF** (Measure function - Robustness).
Model Explainability	Are model explainability (XAI) techniques implemented and validated to ensure that model decisions can be understood, debugged, and audited for bias or security flaws?	Review of XAI implementation (e.g., LIME, SHAP); Interview with data scientists on interpretability.	XAI documentation and reports (e.g., feature importance scores); Logs of model explanations for a sample of decisions.	**ISO/IEC 42001:2023** (8.3.4 - Transparency and explainability); **NIST AI RMF** (Measure function - Interpretability).
Bias and Fairness	Are systematic bias and fairness assessments conducted on the model and its training data, and are documented efforts made to mitigate unfair outcomes?	Review of fairness metrics (e.g., disparate impact, equal opportunity difference) and mitigation strategies.	Fairness assessment reports; Documentation of bias mitigation techniques (e.g., re-weighting, adversarial debiasing).	**NIST AI RMF** (Measure function - Fairness); **Regulatory Compliance** (Emerging AI-specific regulations).

Domain 6: Access Control, Identity, and Privilege Management

This domain examines the controls governing access to AI systems, data, and infrastructure components.

Audit Area	Audit Question	Validation Method	Evidence Required	Auditor Comments with references to standards and frameworks
Least Privilege	Is the principle of least privilege strictly enforced for all human and service accounts accessing AI training environments, model repositories, and production endpoints?	Review of access control lists (ACLs) and role-based access control (RBAC) policies in MLOps platforms and cloud environments.	RBAC matrix for AI infrastructure; Access review logs; Evidence of automated privilege revocation for terminated users.	**ISO/IEC 27002:2022** (5.15 - Access control); **NIST SP 800-53** (AC-6 - Least Privilege).
Service Account Security	Are service accounts and API keys used by AI pipelines (e.g., for data access, model deployment) managed securely, including rotation, vaulting, and non-use of default credentials?	Inspection of secrets management solution (e.g., HashiCorp Vault, AWS Secrets Manager) configuration.	Secrets management policy; Audit logs of key rotation; Inventory of service accounts and their permissions.	**ISO/IEC 27002:2022** (5.18 - Access rights review); **SOC 2** (Security - Logical Access Controls).

Authentication	Is multi-factor authentication (MFA) enforced for all administrative and privileged access to AI development and production environments?	Review of Identity Provider (IdP) configuration and access logs.	IdP configuration screenshots showing MFA enforcement; Access logs confirming MFA usage.	**ISO/IEC 27002:2022** (5.17 - Authentication information); **NIST SP 800-63B** (Digital Identity Guidelines).

Domain 7: Infrastructure, Cloud, and MLOps Security

This domain assesses the security of the underlying infrastructure, cloud services, and the MLOps pipelines used to build, train, and deploy AI models.

Audit Area	Audit Question	Validation Method	Evidence Required	Auditor Comments with references to standards and frameworks
Network Segmentation	Are AI development, training, and production environments logically and physically separated (e.g., via VPCs, subnets, firewalls) from the corporate network and each other?	Review of network architecture diagrams and firewall rulesets.	Network segmentation diagrams; Firewall/Security Group configuration rules; Penetration test reports on network boundaries.	**ISO/IEC 27002:2022** (7.2 - Network security); **NIST SP 800-53** (SC-7 - Boundary Protection).
Configuration Hardening	Are all compute resources (e.g., VMs, containers, Kubernetes clusters) used for AI workloads hardened according to established security baselines (e.g., CIS benchmarks)?	Review of configuration management tools (e.g., Ansible, Terraform) and vulnerability scan reports.	Hardening standards documentation; Configuration compliance reports; Vulnerability scan results for AI infrastructure.	**ISO/IEC 27002:2022** (8.9 - Configuration management); **NIST SP 800-70** (Security Configuration Checklists).
MLOps Pipeline Security	Is the MLOps CI/CD pipeline secured against tampering, ensuring that only authorized, scanned, and tested code/models can be deployed to production?	Review of CI/CD pipeline definition files (e.g., Jenkinsfile, GitHub Actions); Examination of artifact signing and verification processes.	Pipeline definition showing security gates (SAST/DAST, model scanning); Evidence of digital signing for deployed artifacts.	**OWASP Top 10 for LLMs** (SCV - Supply Chain Vulnerabilities); **NIST SP 800-218** (Secure Software Development Framework).

Domain 8: Monitoring, Logging, and Incident Response

This domain assesses the capabilities for continuous monitoring of AI system performance and security events, and for responding to AI-specific incidents.

Audit Area	Audit Question	Validation Method	Evidence Required	Auditor Comments with references to standards and frameworks
Model Monitoring	Are AI models continuously monitored in production for	Review of model monitoring dashboards and	Model performance and drift monitoring reports; Alert logs for detected	**ISO/IEC 42001:2023** (8.3.5 - Monitoring and measuring of AI system

	performance drift, data drift, concept drift, and security anomalies (e.g., sudden changes in input patterns)?	alert configurations in the MLOps platform.	anomalies; Threshold configuration documentation.	performance); **NIST AI RMF** (Maintain function).
Security Logging	Are comprehensive security logs (including input/output data, access attempts, and administrative actions) generated, protected from tampering, and retained according to policy?	Inspection of SIEM/log management system configuration; Review of log retention policy.	Log retention policy; Evidence of log integrity controls (e.g., hashing, WORM storage); Sample of security logs.	**ISO/IEC 27002:2022** (8.15 - Logging); **SOC 2** (Security - Monitoring Activities).
Incident Response	Does the incident response plan include specific playbooks and procedures for handling AI-specific incidents, such as model poisoning, adversarial attacks, or bias-related failures?	Review of the AI Incident Response Plan; Walkthrough or tabletop exercise documentation.	AI Incident Response Playbooks; Records of incident response training; post-incident review reports for AI-related events.	**ISO/IEC 27002:2022** (5.26 - Incident management); **NIST SP 800-61** (Computer Security Incident Handling Guide).

Domain 9: Third Party, Open Source, and Supply Chain Risk

This domain assesses the management of security risks introduced by external components, including third-party models, open-source libraries, and external data providers.

Audit Area	Audit Question	Validation Method	Evidence Required	Auditor Comments with references to standards and frameworks
Third-Party Models	Is a due diligence process performed on all third-party or pre-trained models (e.g., LLMs, foundation models) to assess their security, provenance, and compliance with organizational standards?	Review of third-party risk assessment reports for external AI services/models.	Third-party AI vendor risk assessment documentation; Contractual agreements including security clauses.	**ISO/IEC 27002:2022** (5.21 - Managing information security in the supply chain); **NIST AI RMF** (Govern function - Supply Chain).
Open-Source Security	Are automated tools used to scan open-source libraries and dependencies for known vulnerabilities (CVEs) and license compliance before they are integrated into the AI system code base?	Review of Software Composition Analysis (SCA) tool configuration and reports.	SCA tool reports showing vulnerability remediation; Policy on acceptable open-source licenses.	**OWASP Top 10 for LLMs** (SCV - Supply Chain Vulnerabilities); **NIST SP 800-161** (Supply Chain Risk Management).
Data Provider Risk	Are contractual and technical controls in place to ensure that external data providers maintain the security and integrity of the data they supply for AI training?	Review of Data Sharing Agreements (DSAs) and security audit reports of data providers.	Data Sharing Agreements with security and audit clauses; Evidence of data integrity checks upon ingestion.	**ISO/IEC 27002:2022** (5.23 - Information security for use of cloud services); **GDPR** (Data Processor requirements).

Domain 10: Regulatory Compliance, Ethics, and Responsible AI

This domain ensures that AI systems comply with applicable laws, regulations, and internal ethical guidelines, with particular focus on emerging AI-specific regulations.

Audit Area	Audit Question	Validation Method	Evidence Required	Auditor Comments with references to standards and frameworks
Regulatory Mapping	Is there a current inventory of all applicable AI-specific regulations (e.g., EU AI Act, state-level laws) and a documented mapping of these requirements to internal controls?	Review of the regulatory compliance matrix and legal opinions.	Regulatory compliance matrix for AI systems; Legal counsel sign-off on compliance status.	**GDPR** (Lawfulness, fairness, and transparency); **Emerging AI-specific regulations** (e.g., EU AI Act requirements for high-risk systems).
Responsible AI	Are processes in place to continuously monitor AI systems for unintended consequences, societal impacts, and deviations from responsible AI principles post-deployment?	Review of Responsible AI (RAI) monitoring reports and stakeholder feedback mechanisms.	RAI monitoring dashboards; Records of stakeholder engagement and feedback; Remediation plans for unintended consequences.	**NIST AI RMF** (Govern and Map functions - Societal Impact); **ISO/IEC 42001:2023** (6.1.2 - AI system impact assessment).
Audit Trail	Is a complete and immutable audit trail maintained for all significant decisions and actions related to the AI system (e.g., model selection, parameter tuning, data filtering)?	Inspection of the MLOps platform's experiment tracking and metadata store.	Experiment tracking logs; Metadata store records showing full lineage of model development; Audit trail policy.	**ISO/IEC 42001:2023** (8.3.3 - Traceability of AI systems); **SOC 2** (Control Activities).

Follow me on:

📌 Click 👉 ***here*** for my LinkedIn Profile.
📌 Click 👉 ***here*** for CyberVerge's LinkedIn Page
📌 Click 👉 ***here*** & 👉 ***here*** for X Accounts