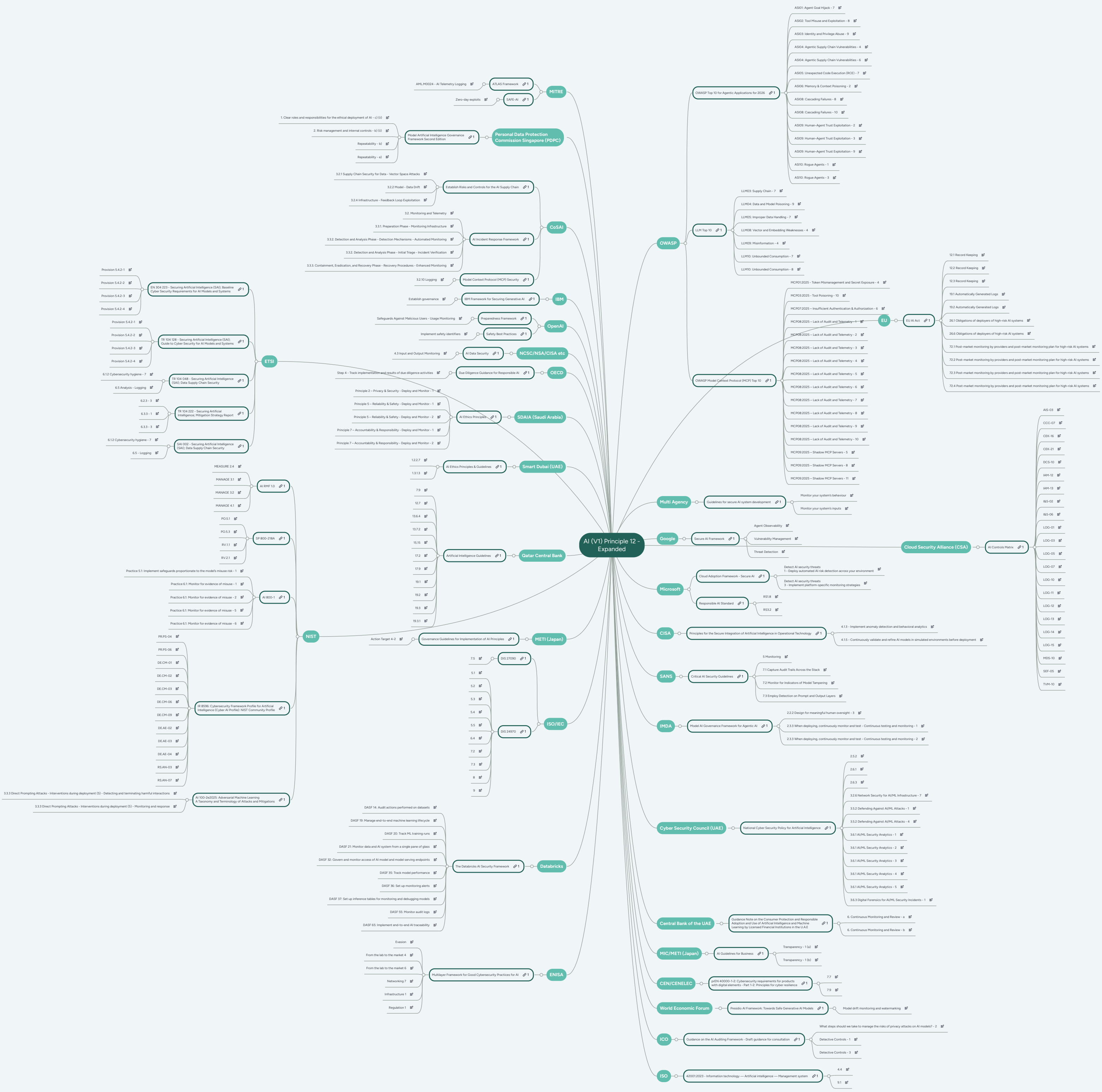

AI (V1) Principle 12 - Expanded
MITRE
ATLAS Framework
AML.M0024 - AI Telemetry Logging
SAIF-AI
Zero-day exploits
Personal Data Protection Commission Singapore (PDPC)
Model Artificial Intelligence Governance Framework Second Edition
1. Clear roles and responsibilities for the ethical deployment of AI - c) (ii)
2. Risk management and internal controls - b) (ii)
Repeatability - iii)
Repeatability - vi)
CoSAI
Establish Risks and Controls for the AI Supply Chain
3.2.1 Supply Chain Security for Data - Vector Space Attacks
3.2.2 Model - Data Drift
3.2.4 Infrastructure - Feedback Loop Exploitation
AI Incident Response Framework
3.2. Monitoring and Telemetry
3.3.1. Preparation Phase - Monitoring Infrastructure
3.3.2. Detection and Analysis Phase - Detection Mechanisms - Automated Monitoring
3.3.2. Detection and Analysis Phase - Initial Triage - Incident Verification
3.3.5. Containment, Eradication, and Recovery Phase - Recovery Procedures - Enhanced Monitoring
Model Context Protocol (MCP) Security
3.2.10 Logging
IBM
IBM Framework for Securing Generative AI
Establish governance
OpenAI
Preparedness Framework
Safeguards Against Malicious Users - Usage Monitoring
Safety Best Practices
Implement safety identifiers
NCSC/NSA/CISA etc
AI Data Security
4.3 Input and Output Monitoring
OECD
Due Diligence Guidance for Responsible AI
Step 4 - Track implementation and results of due diligence activities
SDAIA (Saudi Arabia)
AI Ethics Principles
Principle 2 – Privacy & Security - Deploy and Monitor - 1
Principle 5 – Reliability & Safety - Deploy and Monitor - 1
Principle 5 – Reliability & Safety - Deploy and Monitor - 2
Principle 7 – Accountability & Responsibility - Deploy and Monitor - 1
Principle 7 – Accountability & Responsibility - Deploy and Monitor - 2
Smart Dubai (UAE)
AI Ethics Principles & Guidelines
1.2.2.7
1.3.1.3
Qatar Central Bank
Artificial Intelligence Guidelines
7.9
12.7
13.6.4
13.7.2
15.15
17.2
17.9
19.1
19.2
19.3
19.3.1
METI (Japan)
Governance Guidelines for Implementation of AI Principles
Action Target 4-2
ISO/IEC
DIS 27090
7.5
DIS 24970
5.1
5.2
5.3
5.4
5.5
6.4
7.2
7.5
8
9
Databricks
The Databricks AI Security Framework
DASF 14: Audit actions performed on datasets
DASF 19: Manage end-to-end machine learning lifecycle
DASF 20: Track ML training runs
DASF 21: Monitor data and AI system from a single pane of glass
DASF 32: Govern and monitor access of AI model and model serving endpoints
DASF 35: Track model performance
DASF 36: Set up monitoring alerts
DASF 37: Set up inference tables for monitoring and debugging models
DASF 55: Monitor audit logs
DASF 65: Implement end-to-end AI traceability
ENISA
Multilayer Framework for Good Cybersecurity Practices for AI
Evasion
From the lab to the market 4
From the lab to the market 6
Networking 7
Infrastructure 1
Regulation 1
ETSI
EN 304 223 - Securing Artificial Intelligence (SAI); Baseline Cyber Security Requirements for AI Models and Systems
Provision 5.4.2-1
Provision 5.4.2-2
Provision 5.4.2-3
Provision 5.4.2-4
TR 104 128 - Securing Artificial Intelligence (SAI); Guide to Cyber Security for AI Models and Systems
Provision 5.4.2-1
Provision 5.4.2-2
Provision 5.4.2-3
Provision 5.4.2-4
TR 104 048 - Securing Artificial Intelligence (SAI); Data Supply Chain Security
6.1.2 Cybersecurity hygiene - 7
6.5 Analysis - Logging
TR 104 222 - Securing Artificial Intelligence; Mitigation Strategy Report
6.2.3 - 3
6.3.3 - 1
6.3.3 - 3
SAI 002 - Securing Artificial Intelligence (SAI); Data Supply Chain Security
6.1.2 Cybersecurity hygiene - 7
6.5 - Logging
NIST
AI RMF 1.0
MEASURE 2.4
MANAGE 3.1
MANAGE 3.2
MANAGE 4.1
SP 800-218A
PO.5.1
PO.5.3
RV.1.1
RV.2.1
AI 800-1
Practice 5.1: Implement safeguards proportionate to the model's misuse risk - 1
Practice 6.1: Monitor for evidence of misuse - 1
Practice 6.1: Monitor for evidence of misuse - 2
Practice 6.1: Monitor for evidence of misuse - 5
Practice 6.1: Monitor for evidence of misuse - 6
IR 8596: Cybersecurity Framework Profile for Artificial Intelligence (Cyber AI Profile): NIST Community Profile
PR.PS-04
PR.PS-06
DE.CM-01
DE.CM-02
DE.CM-03
DE.CM-06
DE.CM-09
DE.AE-02
DE.AE-03
DE.AE-04
RS.AN-03
RS.AN-07
AI 100-2e2025: Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations
3.3.3 Direct Prompting Attacks - Interventions during deployment (5) - Detecting and terminating harmful interactions
3.3.3 Direct Prompting Attacks - Interventions during deployment (5) - Monitoring and response
OWASP
OWASP Top 10 for Agentic Applications for 2026
ASI01: Agent Goal Hijack - 7
ASI02: Tool Misuse and Exploitation - 8
ASI03: Identity and Privilege Abuse - 9
ASI04: Agentic Supply Chain Vulnerabilities - 4
ASI04: Agentic Supply Chain Vulnerabilities - 6
ASI05: Unexpected Code Execution (RCE) - 7
ASI06: Memory & Context Poisoning - 2
ASI08: Cascading Failures - 8
ASI08: Cascading Failures - 10
ASI09: Human-Agent Trust Exploitation - 2
ASI09: Human-Agent Trust Exploitation - 8
ASI09: Human-Agent Trust Exploitation - 11
ASI10: Rogue Agents - 1
ASI10: Rogue Agents - 3
LLM Top 10
LLM03: Supply Chain - 7
LLM04: Data and Model Poisoning - 9
LLM05: Improper Output Handling - 7
LLM08: Vector and Embedding Weaknesses - 4
LLM09: Misinformation - 4
LLM10: Unbounded Consumption - 7
LLM10: Unbounded Consumption - 8
OWASP Model Context Protocol (MCP) Top 10
MCP01:2025 - Token Mismanagement and Secret Exposure - 4
MCP03:2025 - Tool Poisoning - 10
MCP07:2025 - Insufficient Authentication & Authorization - 8
MCP08:2025 – Lack of Audit and Telemetry - 1
MCP08:2025 – Lack of Audit and Telemetry - 2
MCP08:2025 – Lack of Audit and Telemetry - 3
MCP08:2025 – Lack of Audit and Telemetry - 4
MCP08:2025 – Lack of Audit and Telemetry - 5
MCP08:2025 – Lack of Audit and Telemetry - 6
MCP08:2025 – Lack of Audit and Telemetry - 7
MCP08:2025 – Lack of Audit and Telemetry - 8
MCP08:2025 – Lack of Audit and Telemetry - 9
MCP08:2025 – Lack of Audit and Telemetry - 10
MCP09:2025 – Shadow MCP Servers - 5
MCP09:2025 – Shadow MCP Servers - 8
MCP09:2025 – Shadow MCP Servers - 11
EU
EU AI Act
12.1 Record Keeping
12.2 Record Keeping
12.3 Record Keeping
19.1 Automatically Generated Logs
19.2 Automatically Generated Logs
26.1 Obligations of deployers of high-risk AI systems
26.6 Obligations of deployers of high-risk AI systems
72.1 Post-market monitoring by providers and post-market monitoring plan for high-risk AI systems
72.2 Post-market monitoring by providers and post-market monitoring plan for high-risk AI systems
72.3 Post-market monitoring by providers and post-market monitoring plan for high-risk AI systems
72.4 Post-market monitoring by providers and post-market monitoring plan for high-risk AI systems
Multi Agency
Guidelines for secure AI system development
Monitor your system's behaviour
Monitor your system's inputs
Google
Secure AI Framework
Agent Observability
Vulnerability Management
Threat Detection
Cloud Security Alliance (CSA)
AI Controls Matrix
AIS-03
CCC-07
CEK-16
CEK-21
DCS-10
IAM-12
IAM-13
IES-02
IES-06
LOG-01
LOG-03
LOG-05
LOG-07
LOG-10
LOG-11
LOG-12
LOG-13
LOG-14
LOG-15
MDS-10
SEF-05
TVM-10
Microsoft
Cloud Adoption Framework - Secure AI
Detect AI security threats
1 - Deploy automated AI risk detection across your environment
Detect AI security threats
3 - Implement platform-specific monitoring strategies
Responsible AI Standard
RS1.8
RS3.2
CISA
Principles for the Secure Integration of Artificial Intelligence in Operational Technology
4.1.3 - Implement anomaly detection and behavioral analytics
4.1.5 - Continuously validate and refine AI models in simulated environments before deployment
SANS
Critical AI Security Guidelines
5 Monitoring
7.1 Capture Audit Trails Across the Stack
7.2 Monitor for Indicators of Model Tampering
7.3 Employ Detection on Prompt and Output Layers
IMDA
Model AI Governance Framework for Agentic AI
2.2.2 Design for meaningful human oversight - 3
2.3.3 When deploying, continuously monitor and test - Continuous testing and monitoring - 1
2.3.3 When deploying, continuously monitor and test - Continuous testing and monitoring - 2
Cyber Security Council (UAE)
National Cyber Security Policy for Artificial Intelligence
2.5.2
2.6.1
2.6.3
3.2.6 Network Security for AI/ML Infrastructure - 7
3.5.2 Defending Against AI/ML Attacks - 1
3.5.2 Defending Against AI/ML Attacks - 4
3.6.1 AI/ML Security Analytics - 1
3.6.1 AI/ML Security Analytics - 2
3.6.1 AI/ML Security Analytics - 3
3.6.1 AI/ML Security Analytics - 4
3.6.1 AI/ML Security Analytics - 5
3.6.3 Digital Forensics for AI/ML Security Incidents - 1
Central Bank of the UAE
Guidance Note on the Consumer Protection and Responsible Adoption and Use of Artificial Intelligence and Machine Learning by Licensed Financial Institutions in the U.A.E
6. Continuous Monitoring and Review - a
6. Continuous Monitoring and Review - b
MIC/METI (Japan)
AI Guidelines for Business
Transparency - 1 (ii)
Transparency - 1 (iii)
CEN/CENELEC
prEN 40000-1-2: Cybersecurity requirements for products with digital elements - Part 1-2: Principles for cyber resilience
7.7
7.9
World Economic Forum
Presidio AI Framework: Towards Safe Generative AI Models
Model drift monitoring and watermarking
ICO
Guidance on the AI Auditing Framework - Draft guidance for consultation
What steps should we take to manage the risks of privacy attacks on AI models? - 2
Detective Controls - 1
Detective Controls - 3
ISO
42001:2023 - Information technology — Artificial intelligence — Management system
4.4
9.1

AI (V1) Principle 12 - Expanded

1. EU

1.1. EU AI Act

Link: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689

Link: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689

1.1.1. 12.1 Record Keeping

> High-risk AI systems shall technically allow for the automatic recording of events (logs) over the lifetime of the system.

1.1.2. 12.2 Record Keeping

> In order to ensure a level of traceability of the functioning of a high-risk AI system that is appropriate to the intended purpose of the system, logging capabilities shall enable the recording of events relevant for:
>
> (a) identifying situations that may result in the high-risk AI system presenting a risk within the meaning of Article 79(1) or in a substantial modification;
> (b) facilitating the post-market monitoring referred to in Article 72; and
> (c) monitoring the operation of high-risk AI systems referred to in Article 26(5).

1.1.3. 12.3 Record Keeping

> For high-risk AI systems referred to in point 1 (a), of Annex III, the logging capabilities shall provide, at a minimum:
> (a) recording of the period of each use of the system (start date and time and end date and time of each use);
> (b) the reference database against which input data has been checked by the system;
> (c) the input data for which the search has led to a match;
> (d) the identification of the natural persons involved in the verification of the results, as referred to in Article 14(5).

1.1.4. 19.1 Automatically Generated Logs

> Providers of high-risk AI systems shall keep the logs referred to in Article 12(1), automatically generated by their high-risk AI systems, to the extent such logs are under their control. Without prejudice to applicable Union or national law, the logs shall be kept for a period appropriate to the intended purpose of the high-risk AI system, of at least six months, unless provided otherwise in the applicable Union or national law, in particular in Union law on the protection of personal data.

1.1.5. 19.2 Automatically Generated Logs

> Providers that are financial institutions subject to requirements regarding their internal governance, arrangements or processes under Union financial services law shall maintain the logs automatically generated by their high-risk AI systems as part of the documentation kept under the relevant financial services law.

1.1.6. 26.1 Obligations of deployers of high-risk AI systems

> Deployers of high-risk AI systems shall take appropriate technical and organisational measures to ensure they use such systems in accordance with the instructions for use accompanying the systems, pursuant to paragraphs 3 and 6.

1.1.7. 26.6 Obligations of deployers of high-risk AI systems

> Deployers of high-risk AI systems shall keep the logs automatically generated by that high-risk AI system to the extent such logs are under their control, for a period appropriate to the intended purpose of the high-risk AI system, of at least six months, unless provided otherwise in applicable Union or national law, in particular in Union law on the protection of personal data.
>
> Deployers that are financial institutions subject to requirements regarding their internal governance, arrangements or processes under Union financial services law shall maintain the logs as part of the documentation kept pursuant to the relevant Union financial service law.

1.1.8. 72.1 Post-market monitoring by providers and post-market monitoring plan for high-risk AI systems

> Providers shall establish and document a post-market monitoring system in a manner that is proportionate to the nature of the AI technologies and the risks of the high-risk AI system.

1.1.9. 72.2 Post-market monitoring by providers and post-market monitoring plan for high-risk AI systems

> The post-market monitoring system shall actively and systematically collect, document and analyse relevant data which may be provided by deployers or which may be collected through other sources on the performance of high-risk AI systems throughout their lifetime, and which allow the provider to evaluate the continuous compliance of AI systems with the requirements set out in Chapter III, Section 2. Where relevant, post-market monitoring shall include an analysis of the interaction with other AI systems. This obligation shall not cover sensitive operational data of deployers which are law-enforcement authorities.

1.1.10. 72.3 Post-market monitoring by providers and post-market monitoring plan for high-risk AI systems

The post-market monitoring system shall be based on a post-market monitoring plan. The post-market monitoring plan shall be part of the technical documentation referred to in Annex IV. The Commission shall adopt an implementing act laying down detailed provisions establishing a template for the post-market monitoring plan and the list of elements to be included in the plan by 2 February 2026. That implementing act shall be adopted in accordance with the examination procedure referred to in Article 98(2).

1.1.11. 72.4 Post-market monitoring by providers and post-market monitoring plan for high-risk AI systems

For high-risk AI systems covered by the Union harmonisation legislation listed in Section A of Annex I, where a post-market monitoring system and plan are already established under that legislation, in order to ensure consistency, avoid duplications and minimise additional burdens, providers shall have a choice of integrating, as appropriate, the necessary elements described in paragraphs 1, 2 and 3 using the template referred in paragraph 3 into systems and plans already existing under that legislation, provided that it achieves an equivalent level of protection.

The first subparagraph of this paragraph shall also apply to high-risk AI systems referred to in point 5 of Annex III placed on the market or put into service by financial institutions that are subject to requirements under Union financial services law regarding their internal governance, arrangements or processes.

2. OWASP

2.1. OWASP Top 10 for Agentic Applications for 2026

Link: https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/

Link: https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/

2.1.1. ASI01: Agent Goal Hijack - 7

Maintain comprehensive logging and continuous monitoring of agent activity, establishing a behavioral baseline that includes goal state, tool-use patterns, and invariant properties (e.g., schema, access patterns). Track a stable identifier for the active goal where feasible, and alert on Page 11genai.owasp.org any deviations-such as unexpected goal changes, anomalous tool sequences, or shifts from the established baseline-so that unauthorized goal drift is immediately visible in operations.

2.1.2. ASI02: Tool Misuse and Exploitation - 8

Logging, Monitoring, and Drift Detection. Maintain immutable logs of all tool invocations and parameter changes. Continuously monitor for anomalous execution rates, unusual tool-chaining patterns (e.g., DB read followed by external transfer), and policy violations.

2.1.3. ASI03: Identity and Privilege Abuse - 9

Detect abnormal cross-agent privilege elevation and device-code style phishing flows by monitoring when agents request new scopes or reuse tokens outside their original, signed intent.

2.1.4. ASI04: Agentic Supply Chain Vulnerabilities - 4

Secure prompts and memory: Put prompts, orchestration scripts, and memory schemas under version control with peer review; scan for anomalies.

2.1.5. ASI04: Agentic Supply Chain Vulnerabilities - 6

Continuous validation and monitoring: Re-check signatures, hashes, and SBOMs (incl. AIBOMs) at runtime; monitor behavior, privilege use, lineage, and inter-module telemetry for anomalies.

2.1.6. ASI05: Unexpected Code Execution (RCE) - 7

Code analysis and monitoring: Do static scans before execution; enable runtime monitoring; watch for prompt-injection patterns; log and audit all generation and runs.

2.1.7. ASI06: Memory & Context Poisoning - 2

Content validation: Scan all new memory writes and model outputs (rules + AI) for malicious or sensitive content before commit.

2.1.8. ASI08: Cascading Failures - 8

Behavioral and governance drift detection: Track decisions vs baselines and alignment; flag gradual degradation.

2.1.9. ASI08: Cascading Failures - 10

Logging and non-repudiation. Record all inter-agent messages, policy decisions, and execution outcomes in tamper-evident, time-stamped logs bound to cryptographic agent identities. Maintain lineage metadata for every propagated action to support forensic traceability, rollback validation, and accountability during cascades.

2.1.10. ASI09: Human-Agent Trust Exploitation - 2

Immutable logs: Keep tamper-proof records of user queries and agent actions for audit and forensics.

2.1.11. ASI09: Human-Agent Trust Exploitation - 3

Behavioral detection: Monitor sensitive data being exposed in either conversations or Agentic connections, as well as risky action executions over time.

2.1.12. ASI09: Human-Agent Trust Exploitation - 9

Plan-divergence detection: Compare agent action sequences against approved workflow baselines and alert when unusual detours, skipped validation steps, or novel tool combinations indicate possible deception or drift.

2.1.13. ASI10: Rogue Agents - 1

Governance & Logging: Maintain comprehensive, immutable and signed audit logs of all agent actions, tool calls, and inter-agent communication to review for stealth infiltration or unapproved delegation.

2.1.14. ASI10: Rogue Agents - 3

Monitoring & Detection: Deploy behavioral detection, such as watchdog agents to validate peer behavior and outputs, focusing on detecting collusion patterns and coordinated false signals. Monitor for anomalies such as excessive or abnormal actions executions.

2.2. LLM Top 10

Link: https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/

Link: https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/

2.2.1. LLM03: Supply Chain - 7

Implement strict monitoring and auditing practices for collaborative model development environments to prevent and quickly detect any abuse. "HuggingFace SF_Convertbot Scanner" is an example of automated scripts to use.

2.2.2. LLM04: Data and Model Poisoning - 9

Monitor training loss and analyze model behavior for signs of poisoning. Use thresholds to detect anomalous outputs.

2.2.3. LLM05: Improper Data Handling - 7

Implement robust logging and monitoring systems to detect unusual patterns in LLM outputs that might indicate exploitation attempts.

2.2.4. LLM08: Vector and Embedding Weaknesses - 4

Maintain detailed immutable logs of retrieval activities to detect and respond promptly to suspicious behavior.

2.2.5. LLM09: Misinformation - 4

Implement tools and processes to automatically validate key outputs, especially output from high-stakes environments.

2.2.6. LLM10: Unbounded Consumption - 7

Continuously monitor resource usage and implement logging to detect and respond to unusual patterns of resource consumption.

2.2.7. LLM10: Unbounded Consumption - 8

Implement watermarking frameworks to embed and detect unauthorized use of LLM outputs.

2.3. OWASP Model Context Protocol (MCP) Top 10

Link: https://owasp.org/www-project-mcp-top-10/

Link: https://owasp.org/www-project-mcp-top-10/

2.3.1. MCP01:2025 - Token Mismanagement and Secret Exposure - 4

Secure Context & Log Management
- Redact or mask secrets before writing to logs or telemetry.
- Store diagnostic traces in protected locations with strict access control.
- Rotate and invalidate all tokens immediately upon suspected exposure.

2.3.2. MCP03:2025 - Tool Poisoning - 10

Conduct forensic analysis: which agents used the poisoned schema, what actions executed, which data changed or was removed.

2.3.3. MCP07:2025 – Insufficient Authentication & Authorization - 6

Logging, Monitoring & Auditing
- Log every authentication attempt and authorization decision.
- Detects repeated failed logins, invalid tokens, or cross-tenant token reuse.
- Feed these logs into a SIEM/XDR for anomaly detection and alerting.

2.3.4. MCP08:2025 – Lack of Audit and Telemetry - 1

Implement Structured, Tamper-Evident Logging Log all agent actions, tool invocations, schema versions, and context snapshots in a structured format (JSON, CEF, OTEL). Apply cryptographic hashing (HMAC, SHA-256) to log files for integrity. Store logs in append-only or write-once media (e.g., AWS S3 Object Lock, WORM storage).

2.3.5. MCP08:2025 – Lack of Audit and Telemetry - 2

Include essential fields:
timestamp
agent_id
session_id
tool_invoked
parameters_used
response_summary
user_identity (if applicable)

2.3.6. MCP08:2025 – Lack of Audit and Telemetry - 3

Integrate with SIEM, XDR, or Centralized Monitoring
- Forward MCP logs to enterprise SIEM systems (Splunk, ELK, Sentinel, Chronicle, etc.) for correlation.
- Establish automated alert rules for high-risk activities (e.g., tool execution involving sensitive data).
- Use Extended Detection and Response (XDR) systems to correlate agent behaviors with network or endpoint signals.

2.3.7. MCP08:2025 – Lack of Audit and Telemetry - 4

Protect Sensitive Data in Logs
- Implement PII-safe logging: tokenize or mask user identifiers and redact sensitive fields before storage.
- Use field-level encryption for secrets, tokens, or confidential context entries.
- Apply data classification labels to log streams to govern retention and access.

2.3.8. MCP08:2025 – Lack of Audit and Telemetry - 5

Establish Behavioral Baselines
- Collect telemetry to build a behavioral profile of normal agent operations.
- Use anomaly detection or ML-based behavioral analytics to flag deviations (e.g., unexpected API calls, unusual output patterns).
- Regularly review and update baseline thresholds.

2.3.9. MCP08:2025 – Lack of Audit and Telemetry - 6

Enforce Access Control & Segregation of Duties
- Restrict who can access logs — separate operational monitoring from security investigations.
- Require dual authorization for log deletion or retention changes.
- Apply least privilege and auditing on logging subsystems themselves.

2.3.10. MCP08:2025 – Lack of Audit and Telemetry - 7

Implement Real-Time Observability
- Use OpenTelemetry or equivalent frameworks to trace requests across the MCP pipeline — from prompt creation to tool invocation.
- Tag every trace with session and schema identifiers to enable end-to-end correlation.
- Display agent performance and behavior dashboards for operational visibility.

2.3.11. MCP08:2025 – Lack of Audit and Telemetry - 8

Retention & Compliance Policies
- Align log retention with applicable frameworks (e.g., PCI DSS: 1 year minimum).
- Automatically archive or purge logs per retention schedule.
- Periodically verify that retention, encryption, and deletion processes function as intended.

2.3.12. MCP08:2025 – Lack of Audit and Telemetry - 9

Continuous Audit & Verification
- Conduct periodic audit drills to ensure investigators can reconstruct events from logs.
- Test integrity checks — attempt to tamper with logs and validate detection alerts.
- Implement audit trail self-verification, where logs cross-reference session data for consistency.

2.3.13. MCP08:2025 – Lack of Audit and Telemetry - 10

Re-enable detailed logging at all MCP layers (agent, tool, and network). Deploy forwarders to send logs to central SIEM/XDR with retention guarantees. Implement masking and pseudonymization to balance privacy and audit needs. Reconstruct minimal timeline from external system logs (firewalls, proxies). Perform root-cause review and enforce mandatory logging for all MCP agents.

2.3.14. MCP09:2025 – Shadow MCP Servers - 5

Monitor for Anomalous or Unauthorized Behavior
- Correlate telemetry to identify new MCP-related API traffic or agent activity from unknown hosts.
- Set up alerts for endpoints responding on MCP-standard routes (/mcp, /agent/tools, /context).
- Track configuration drift and endpoint proliferation over time.

2.3.15. MCP09:2025 – Shadow MCP Servers - 8

Detection and Response Integration
- Include shadow MCP detection in threat-hunting playbooks.
- Upon detection, trigger an incident response workflow to contain, image, and analyze the rogue server.
- Track remediation metrics (mean time to discovery and closure).

2.3.16. MCP09:2025 – Shadow MCP Servers - 11

Review logs and assess data exposure or leakage.

3. Multi Agency

3.1. Guidelines for secure AI system development

Link: https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf

Link: https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf

3.1.1. Monitor your system's behaviour

You measure the outputs and performance of your model and system such that you can observe sudden and gradual changes in behaviour affecting security. You can account for and identify potential intrusions and compromises, as well as natural data drift.

3.1.2. Monitor your system's inputs

In line with privacy and data protection requirements, you monitor and log inputs to your system (such as inference requests, queries or prompts) to enable compliance obligations, audit, investigation and remediation in the case of compromise or misuse. This could include explicit detection of out-of- distribution and/or adversarial inputs, including those that aim to exploit data preparation steps (such as cropping and resizing for images).

4. Cloud Security Alliance (CSA)

4.1. AI Controls Matrix

Link: https://cloudsecurityalliance.org/artifacts/ai-controls-matrix

Link: https://cloudsecurityalliance.org/artifacts/ai-controls-matrix

4.1.1. AIS-03

Define and implement technical and operational metrics in alignment with business objectives, security requirements, and compliance obligations.

4.1.2. CCC-07

Implement detection measures with proactive notification in case of changes deviating from the established baseline.

4.1.3. CEK-16

Define, implement and evaluate processes, procedures and technical measures to monitor, review and approve key transitions from any state to/from suspension, which include provisions for legal and regulatory requirements.

4.1.4. CEK-21

Define, implement and evaluate processes, procedures and technical measures in order for the key management system to track and report all cryptographic materials and changes in status, which include provisions for legal and regulatory requirements.

4.1.5. DCS-10

Implement, maintain, and operate datacenter surveillance systems at the external perimeter and at all the ingress and egress points to detect unauthorized ingress and egress attempts.

4.1.6. IAM-12

Define, implement and evaluate processes, procedures and technical measures to ensure the logging infrastructure is read-only for all with write access, including privileged access roles, and that the ability to disable it is controlled through a procedure that ensures the segregation of duties and break glass procedures.

4.1.7. IAM-13

Define, implement and evaluate processes, procedures and technical measures, that ensure identities' activities are identifiable through uniquely associated IDs.

4.1.8. I&S-02

Plan and monitor the availability, quality, and adequate capacity of resources in order to deliver the required system performance as determined by the business.

4.1.9. I&S-06

Design, develop, deploy and configure applications and infrastructures such that tenant access is appropriately segmented and segregated, monitored and restricted.

4.1.10. LOG-01

Establish, document, approve, communicate, apply, evaluate and maintain policies and procedures for logging and monitoring. Review and update the policies and procedures at least annually, or upon significant changes.

4.1.11. LOG-03

Identify and monitor security-related events within applications, the underlying infrastructure, supply chain, and consider logging other events based on risk evaluation. Define and implement a system to generate alerts to responsible stakeholders based on such events and corresponding metrics.

4.1.12. LOG-05

Monitor security audit logs to detect activity outside of typical or expected patterns. Establish and follow a defined process to review and take appropriate and timely actions on detected anomalies.

4.1.13. LOG-07

Establish, document and implement which information meta/data system events should be logged. Review and update the scope at least annually or whenever there is a change in the threat environment.

4.1.14. LOG-10

Establish and maintain a monitoring and internal reporting capability over the operations of cryptographic, encryption and key management policies, processes, procedures, and controls.

4.1.15. LOG-11

Log and monitor key lifecycle management events to enable auditing and reporting on usage of cryptographic keys.

4.1.16. LOG-12

Monitor and log physical access using an auditable access control system.

4.1.17. LOG-13

Define, implement and evaluate processes, procedures and technical measures for the reporting of anomalies and failures of the monitoring system and provide immediate notification to the accountable party.

4.1.18. LOG-14

Log and monitor all input events (content and metadata) to enable auditing and reporting on the usage of AI models.

4.1.19. LOG-15

Log and monitor all output events (content and metadata) to enable auditing and reporting on usage of AI models.

4.1.20. MDS-10

Define, implement, and evaluate processes, procedures, and technical measures for continuous monitoring of model performance metrics over time to identify sudden shifts or unexpected changes in predictions that could degrade model performance.

4.1.21. SEF-05

Establish, monitor and report information security incident metrics.

4.1.22. TVM-10

> Establish, monitor and report metrics for vulnerability identification and remediation at defined intervals.

5. Google

5.1. Secure AI Framework

Link: https://www.saif.google/secure-ai-framework

Link: https://www.saif.google/secure-ai-framework

5.1.1. Agent Observability

> Ensure an agent's actions, tool use, and reasoning are transparent and auditable through logging, allowing for debugging, security oversight, and user insights into agent activity.

5.1.2. Vulnerability Management

> Proactively and continually test and monitor production infrastructure and products for security and privacy regressions.

5.1.3. Threat Detection

> Detect and alert on internal or external attacks on AI assets, infrastructure, and products.

6. Microsoft

6.1. Cloud Adoption Framework - Secure AI

Link: https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/scenarios/ai/secure

Link: https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/scenarios/ai/secure

6.1.1. Detect AI security threats 1 - Deploy automated AI risk detection across your environment

> AI workloads introduce dynamic threats that manual monitoring can't detect quickly enough to prevent damage. Automated systems provide real-time visibility into emerging risks and enable rapid response to security incidents. Use AI security posture management in Microsoft Defender for Cloud to automate detection and remediation of generative AI risks across your Azure environment.

6.1.2. Detect AI security threats 3 - Implement platform-specific monitoring strategies

> AI workloads deployed on different platforms face distinct security challenges that require tailored monitoring approaches. Platform-specific monitoring ensures comprehensive coverage of all potential attack vectors. Apply monitoring guidance based on your deployment architecture:
> AI monitoring on Azure platforms (PaaS)
> AI monitoring on Azure infrastructure (IaaS)

6.2. Responsible AI Standard

Link: https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft-Responsible-AI-Standard-General-Requirements.pdf?culture=en-us&country=us

Link: https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft-Responsible-AI-Standard-General-Requirements.pdf?culture=en-us&country=us

6.2.1. RS1.8

> In the event of failure cases within operational factors and defined ranges, work to resolve the issues. If the Responsible Release Criteria established in requirements RS1.1, RS1.3, RS1.4, and RS1.5 cannot be met, a reassessment of intended uses and updated documentation is required.

6.2.2. RS3.2

> Define and document a standard operating procedure and system health monitoring action plan for each monitoring channel for the system, to include:
> 1) processes for reproducing system failures to support troubleshooting and prevention of future failures,
> 2) which events will be monitored,
> 3) how events will be prioritized for review,
> 4) the expected frequency of those reviews,
> 5) how events will be prioritized for response and timing to resolution,
> 6) how high priority issues related to supporting the Standard and its goals will be escalated to the Office of Responsible AI, and
> 7) engaging customer service to ensure that they are aware of how to respond to issues for the system.

7. CISA

7.1. Principles for the Secure Integration of Artificial Intelligence in Operational Technology

Link: https://www.cisa.gov/sites/default/files/2026-01/joint-guidance-principles-for-the-secure-integration-of-artificial-intelligence-in-operational-technology-508cV2.pdf

Link: https://www.cisa.gov/sites/default/files/2026-01/joint-guidance-principles-for-the-secure-integration-of-artificial-intelligence-in-operational-technology-508cV2.pdf

7.1.1. 4.1.3 - Implement anomaly detection and behavioral analytics

> Establish safe operating bounds for OT devices that detect AI drift, model changes that impact safety and performance, or security risks. As operator processes mature, software safety thresholds can shift from setpoints to anomaly detection of increasingly sophisticated faults. Configure logging so AI decisions can be tracked for compliance and forensic analysis, and so the logged AI identity is distinct from any typical machine or user identifiers

7.1.2. 4.1.5 - Continuously validate and refine AI models in simulated environments before deployment

> Regularly update threat models with AI-specific attack vectors (such as adversarial inputs or data poisoning) and monitor AI system performance for anomalies or manipulation attempts. Continuously update and refine AI models with new OT data to improve precision and reduce false positives/negatives.

8. SANS

8.1. Critical AI Security Guidelines

Link: https://sansorg.egnyte.com/dl/bvkYQxrW8QMj

Link: https://sansorg.egnyte.com/dl/bvkYQxrW8QMj

8.1.1. 5 Monitoring

> Effective monitoring is essential to maintaining AI security over time. AI models and systems must be continuously observed for performance degradation, adversarial attacks, and unauthorized access. Implementing logging, anomaly detection, and drift monitoring ensures AI applications remain reliable and aligned with intended behaviors.

8.1.2. 7.1 Capture Audit Trails Across the Stack

> Ensure logs include prompt inputs, augmentation sources (like vectorDB queries), model outputs, function calls, and tool invocations.

8.1.3. 7.2 Monitor for Indicators of Model Tampering

> Watch for sudden changes in inference behavior, drift in outputs, or increased refusal rates, which may indicate adversarial manipulation or unauthorized updates.

8.1.4. 7.3 Employ Detection on Prompt and Output Layers

> Include pattern-based and behavioral monitoring to identify jailbreak attempts, abuse of multilingual prompts, or bypasses via encoding/compression.

9. IMDA

9.1. Model AI Governance Framework for Agentic AI

Link: https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf

Link: https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf

9.1.1. 2.2.2 Design for meaningful human oversight - 3

> Finally, human oversight should be complemented with automated real-time monitoring to escalate any unexpected or anomalous behaviour. This can be done by implementing alerts for certain logged events (e.g. attempted unauthorised access or multiple failed attempts to call a tool), using data science techniques to identify anomalous agent trajectories, or using agents to monitor other agents. For more information, see Continuous testing and monitoring below.

9.1.2. 2.3.3 When deploying, continuously monitor and test - Continuous testing and monitoring - 1

> Organisations should continuously monitor and log agent behaviour post-deployment, and establish reporting and failsafe mechanisms for agent failures or unexpected behaviours. This allows the organisation to:
> • Intervene in real-time: When potential failures are detected, stop agent workflow and escalate to a human supervisor e.g. if agent attempts unauthorised access
> • Debug when incidents happen: Logging and tracing each step of an agent workflow and agent-to-agent interactions help to identify points of failure
> • Audit at regular intervals: This ensures that the system is performing as expected.

9.1.3. 2.3.3 When deploying, continuously monitor and test - Continuous testing and monitoring - 2

Monitoring and observability are not new concepts, but agents introduce some challenges. As agents execute multiple actions at machine speed, organisations face the issue of extracting meaningful insights from the voluminous logs generated by monitoring systems. This becomes more difficult when high-risk anomalies are expected to be detected in real-time and surfaced as early as possible.
Key considerations when setting up a monitoring system include:
• What to log: Organisations should determine their objectives for monitoring (e.g. real-time intervention, debugging, integration between components) to identify what to log. In doing so, prioritise monitoring for high-risk activities such as updating database records or financial transactions.
• How to effectively monitor logs: Organisations can consider approaches such as:
o Defining alert thresholds:
▪ Programmatic, threshold-based: Define alerts when agents trigger thresholds e.g. agent attempts unauthorised access or makes too many repeated tool calls within a specified timeframe.
▪ Outlier / anomaly detection: Use data science or deep learning techniques to process agent signals and identify anomalous behaviour that may indicate malfunctions.
▪ Agents monitoring other agents: Design agents to monitor other agents in real-time, flagging any anomalies or inconsistencies.
o Defining specific interventions: For each alert type, consider what the level of intervention should be. Some degree of human review should be incorporated, proportionate to the risk level. For example, lower-priority alerts can be flagged for review at a scheduled time, whereas higher-priority ones might require temporarily halting agent execution until a human reviewer can assess. In the event of catastrophic agentic malfunction or compromise, commensurate measures such as termination and fallback solutions should be considered. Finally, continuously test the agentic system even post-deployment to ensure that it works as expected and is not affected by model drift or other changes in the environment.

10. Cyber Security Council (UAE)

10.1. National Cyber Security Policy for Artificial Intelligence

Link: https://csc.gov.ae/documents/38662/0/National+Cyber+Security+Policy+for+Artificial+Intelligence_v1.1.pdf/4e02b32e-9f62-948d-4bc8-b580d596451b?t=1766994254544

Link: https://csc.gov.ae/documents/38662/0/National+Cyber+Security+Policy+for+Artificial+Intelligence_v1.1.pdf/4e02b32e-9f62-948d-4bc8-b580d596451b?t=1766994254544

10.1.1. 2.5.2

The entity shall deploy comprehensive defense strategies for AI/ML systems, including proactive security measures, continuous monitoring, regular testing, and robust incident response capabilities.

10.1.2. 2.6.1

The entity shall establish an AI/ML security analytics framework that utilizes advanced technologies and diverse data sources to enable real-time threat detection, historical analysis, and predictive security insights.

10.1.3. 2.6.3

The entity shall establish a robust and comprehensive digital forensics framework tailored to AI/ML cyber security incidents, enabling effective investigation and analysis.

10.1.4. 3.2.6 Network Security for AI/ML Infrastructure - 7

The entity should ensure that network security logs for AI/ML infrastructure are stored, protected, and analyzed in accordance with the entity's log management policy.

10.1.5. 3.5.2 Defending Against AI/ML Attacks - 1

The entity should implement robust defensive techniques to protect AI/ML systems from attacks, including, but not limited to, anomaly detection, adversarial training, defensive distillation, and feature squeezing.

10.1.6. 3.5.2 Defending Against AI/ML Attacks - 4

The entity should continuously monitor the performance and behavior of AI/ML systems to detect and respond to potential anomalies indicative of an attack.

10.1.7. 3.6.1 AI/ML Security Analytics - 1

The entity should implement advanced security analytics capabilities commensurate with the criticality of the AI/ ML system, to facilitate real-time monitoring, detection, and response to threats targeting AI/ML systems.

10.1.8. 3.6.1 AI/ML Security Analytics - 2

Security analytics should encompass various data sources, including system logs, network traffic, and user activity to provide a holistic view of the security posture of AI/ML systems.

10.1.9. 3.6.1 AI/ML Security Analytics - 3

The entity should employ statistical and behavioral analysis techniques to identify anomalies and potential security incidents in AI/ML systems such as abnormal input patterns that may indicate adversarial attacks, unexpected model behavior, data poisoning, or model extraction attempts.

10.1.10. 3.6.1 AI/ML Security Analytics - 4

The entity should maintain historical security data for AI/ML systems to aid in trend analysis, incident investigation, and the development of predictive security analytics.

10.1.11. 3.6.1 AI/ML Security Analytics - 5

The entity should integrate external threat intelligence feeds with internal data sources to enhance threat detection capabilities.

10.1.12. 3.6.3 Digital Forensics for AI/ML Security Incidents - 1

The entity should develop a framework for conducting digital forensics in the event of AI/ML security incidents, to gather evidence, determine the cause and impact, and provide insight into how similar incidents can be prevented in the future.

11. Central Bank of the UAE

11.1. Guidance Note on the Consumer Protection and Responsible Adoption and Use of Artificial Intelligence and Machine Learning by Licensed Financial Institutions in the U.A.E

Link: https://rulebook.centralbank.ae/en/rulebook/guidance-note-consumer-protection-and-responsible-adoption-and-use-artificial-intelligence

Link: https://rulebook.centralbank.ae/en/rulebook/guidance-note-consumer-protection-and-responsible-adoption-and-use-artificial-intelligence

11.1.1. 6. Continuous Monitoring and Review - a

AI should be subject to continuous monitoring to ensure ongoing understanding, reliability, relevance and alignment with consumer protection objectives.

11.1.2. 6. Continuous Monitoring and Review - b

LFIs are expected to consistently monitor and review and, where appropriate, update or cease using AI and ML models, taking into account changes in data, market conditions and customer behaviors. Independent third-party providers of AI, independent experts and third parties (willing to challenge and question the use of AI in an LFI) should be engaged periodically to assess the development of and use of AI including third-party AI providers.

12. MIC/METI (Japan)

12.1. AI Guidelines for Business

Link: https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20240419_9.pdf

Link: https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20240419_9.pdf

12.1.1. Transparency - 1 (a)

In order to ensure verifiability relating to decisions made by AI, record or store logs of AI training processes, inference processes, rationales of decisions made by AI, and the like (for example, input/output generated when developing and using the AI system or service) to the extent possible based on the data amount or contents.

12.1.2. Transparency - 1 (b)

Discuss method, frequency, maintenance period and so on recordings of data logs, taking into account the importance for identifying causes of accidents, devising preventive measures, or proving requirements for responsibilities for damages, in accordance with characteristic of used technology as well as purposes.

13. CEN/CENELEC

13.1. prEN 40000-1-2: Cybersecurity requirements for products with digital elements - Part 1-2: Principles for cyber resilience

Link: https://genorma.com/en/standards/pren-40000-1-2

Link: https://genorma.com/en/standards/pren-40000-1-2

13.1.1. 7.7

Secure production and distribution

13.1.2. 7.9

Product monitoring

14. World Economic Forum

14.1. Presidio AI Framework: Towards Safe Generative AI Models

Link: https://www3.weforum.org/docs/WEF_Presidio_AI%20Framework_2024.pdf

Link: https://www3.weforum.org/docs/WEF_Presidio_AI%20Framework_2024.pdf

14.1.1. Model drift monitoring and watermarking

> A critical goal of the adaptation phase is to ensure that the adapted model remains effective and aligned with the selected use case. Model drift monitoring involves regularly comparing postdeployment metrics to maintain performance in the face of evolving data, adversarial inputs, noise and external factors. The goal is to mitigate the risk of model drift, where the model's output deviates from expectations over time. Best practices include systematically using data, algorithms, and tools for tracking data drift, and defining response protocols and adaptation techniques to sustain model performance and customer trust.

15. ICO

15.1. Guidance on the AI Auditing Framework - Draft guidance for consultation

Link: https://ico.org.uk/media2/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

Link: https://ico.org.uk/media2/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

15.1.1. What steps should we take to manage the risks of privacy attacks on AI models? - 2

> If you are going to provide a whole model to others via an Application Programming Interface (API), you would not be subject to white-box attacks in this way, because the API's users would not have direct access to the model itself. However, you might still be subjected to black box attacks.
> To mitigate this risk, you could monitor queries from the API's users, in order to detect whether it is being used suspiciously. This may indicate a privacy attack and would require prompt investigation, and potential suspension or blocking of a particular user account. Such measures may become part of common real-time monitoring techniques used to protect against other security threats, such as 'rate-limiting' (reducing the number of queries that can be performed by a particular user in a given time limit).

15.1.2. Detective Controls - 1

> Monitor API requests to detect suspicious requests and take action as a result.

15.1.3. Detective Controls - 3

> Monitor complaints monitoring and take action as a result, including broader analysis to identify other individuals who may be impacted.

16. ISO

16.1. 42001:2023 - Information technology — Artificial intelligence — Management system

Link: https://www.iso.org/standard/42001

Link: https://www.iso.org/standard/42001

16.1.1. 4.4

> AI management system

16.1.2. 9.1

> Monitoring, measurement, analysis and evaluation

17. ENISA

17.1. Multilayer Framework for Good Cybersecurity Practices for AI

Link:
https://www.enisa.europa.eu/sites/default/files/publications/Multilayer%20Framework%20for%20Good%20Cybersecurity%20Practices%20for%20AI.pdf

Link:
https://www.enisa.europa.eu/sites/default/files/publications/Multilayer%20Framework%20for%20Good%20Cybersecurity%20Practices%20for%20AI.pdf

17.1.1. Evasion

> For evasion, tools can be implemented to detect whether a given input is an adversarial example, adversarial training can be used to make the model more robust, and models that are less easily transferable can be used to significantly decrease the ability of a given attacker to properly study the algorithm that works underneath the system.

17.1.2. From the lab to the market 4

> Do you have specific measurements/KPIs/metrics that the AI stakeholders are imposed to use?

17.1.3. From the lab to the market 6

> How do you monitor if such requirements have been met?

17.1.4. Networking 7

How do you monitor/audit the level of the cybersecurity of the AI systems throughout their life cycle?

17.1.5. Infrastructure 1

How do you monitor/audit the appropriateness of the controls undertaken by the AI stakeholders (developers, integrators, critical infrastructures, e.g. telecom operators) to adequately secure the underlying ICT infrastructure?

17.1.6. Regulation 1

How do you monitor the integrity and quality of data sets used for the development of AI systems?

18. Databricks

18.1. The Databricks AI Security Framework

Link: https://www.databricks.com/sites/default/files/2025-02/databricks-ebook-dasf-2.pdf

Link: https://www.databricks.com/sites/default/files/2025-02/databricks-ebook-dasf-2.pdf

18.1.1. DASF 14: Audit actions performed on datasets

Databricks auditing, enhanced by Unity Catalog's events, delivers fine-grained visibility into data access and user activities. This is vital for robust data governance and security, especially in regulated industries. It enables organizations to proactively identify and manage over entitled users, enhancing data security and ensuring compliance.

18.1.2. DASF 19: Manage end-to-end machine learning lifecycle

Databricks includes a managed version of MLflow featuring enterprise security controls and high availability. It supports functionalities like experiments, run management and notebook revision capture. MLflow on Databricks allows tracking and measuring machine learning model training runs, logging model training artifacts and securing machine learning projects.

18.1.3. DASF 20: Track ML training runs

MLflow tracking facilitates the automated recording and retrieval of experiment details, including algorithms, code, datasets, parameters, configurations, signatures and artifacts.

18.1.4. DASF 21: Monitor data and AI system from a single pane of glass

Databricks Lakehouse Monitoring offers a single pane of glass to centrally track tables' data quality and statistical properties and automatically classifies data. It can also track the performance of machine learning models and model serving endpoints by monitoring inference tables containing model inputs and predictions through a single pane of glass.

18.1.5. DASF 32: Govern and monitor access of AI model and model serving endpoints

Mosaic AI Gateway is designed to streamline the usage and management of generative AI models and agents within an organization. It is a centralized service that brings governance, monitoring, and production readiness to model serving endpoints. It also allows you to run, secure, and govern AI traffic to democratize and accelerate AI adoption for your organization. Supported by Model Serving AI Gateway, Databricks external models via the AI Gateway allow you to streamline the usage and management of various large language model (LLM) providers, such as OpenAI and Anthropic, within an organization. You can also use Mosaic AI Model Serving as a provider to serve predictive ML models, which offers rate limits for those endpoints. As part of this support, Model Serving offers a high-level interface that simplifies the interaction with these services by providing a unified endpoint to handle specific LLM-related requests. In addition, Databricks support for external models provides centralized credential management. By storing API keys in one secure location, organizations can enhance their security posture by minimizing the exposure of sensitive API keys throughout the system. It also helps to prevent exposing these keys within code or requiring end users to manage keys safely.

18.1.6. DASF 35: Track model performance

Databricks Lakehouse Monitoring provides performance metrics and data quality statistics across all account tables. It tracks the performance of machine learning models and model serving endpoints by observing inference tables with model inputs and predictions.

18.1.7. DASF 36: Set up monitoring alerts

Databricks SQL alerts can monitor the metrics table for security-based conditions, ensuring data integrity and timely response to potential issues:
- Statistic range alert: Triggers when a specific statistic, such as the fraction of missing values, exceeds a predetermined threshold
- Data distribution shift alert: Activates upon shifts in data distribution, as indicated by the drift metrics table
- Baseline divergence alert: Alerts if data significantly diverges from a baseline, suggesting potential needs for data analysis or model retraining, particularly in InferenceLog analysis

18.1.8. DASF 37: Set up inference tables for monitoring and debugging models

Databricks inference tables automatically record incoming requests and outgoing responses to model serving endpoints, storing them as a Unity Catalog Delta table. This table can be used to monitor, debug and enhance ML models. By coupling inference tables with Lakehouse Monitoring, customers can also set up automated monitoring jobs and alerts on inference tables, such as monitoring text quality or toxicity from endpoints serving LLMs, etc.
Critical applications of an inference table include:
- Retraining dataset creation: Building datasets for the next iteration of your models
- Quality monitoring: Keeping track of production data and model performance
- Diagnostics and debugging: Investigating and resolving issues with suspicious inferences
- Mislabeled data identification: Compiling data that needs relabeling

18.1.9. DASF 55: Monitor audit logs

Audit logs and system tables serve as a centralized operational data store, backed by Delta Lake and governed by Unity Catalog. Audit logs and system tables can be used for a variety of purposes, from user activity, model serving events, and cost monitoring to audit logging. Databricks recommends that customers configure system tables and set up automated monitoring and alerting to meet their needs. The blog post Improve Lakehouse Security Monitoring Using System Tables in Databricks Unity Catalog is a good starting point to help customers get started.

Customers that are using Enhanced Security Monitoring or the Compliance Security Profile can monitor and alert on suspicious activity detected by the behavior-based malware and file integrity monitoring agents.

18.1.10. DASF 65: Implement end-to-end AI traceability

MLflow Tracing is a powerful feature that provides end-to-end observability for gen AI applications, including complex agent-based systems. It records inputs, outputs, intermediate steps, and metadata to give you a complete picture of how your app behaves.

Tracing allows you to:
• Debug and understand your application
• Monitor performance and optimize cost
• Evaluate and enhance application quality
• Ensure auditability and compliance
• Integrate tracing with many popular third-party frameworks

19. ISO/IEC

19.1. DIS 27090

Link: https://www.iso.org/obp/ui/en/#iso:std:iso-iec:27090:dis:ed-1:v1:en

Link: https://www.iso.org/obp/ui/en/#iso:std:iso-iec:27090:dis:ed-1:v1:en

19.1.1. 7.5

Logging and monitoring

19.2. DIS 24970

Link: https://www.iso.org/obp/ui/en/#iso:std:iso-iec:24970:dis:ed-1:v1:en

Link: https://www.iso.org/obp/ui/en/#iso:std:iso-iec:24970:dis:ed-1:v1:en

19.2.1. 5.1

AI system logs

19.2.2. 5.2

Logging components

19.2.3. 5.3

Logging in context

19.2.4. 5.4

Log entries

19.2.5. 5.5

AI system logging

19.2.6. 6.4

Anomoly monitoring of the logging

19.2.7. 7.2

Triggers from operation

19.2.8. 7.3

Triggers from automated monitoring

19.2.9. 8

Information to log

19.2.10. 9

Storing and access to logs

20. METI (Japan)

20.1. Governance Guidelines for Implementation of AI Principles

Link: https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20220128_2.pdf

Link: https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20220128_2.pdf

20.1.1. Action Target 4-2

Companies that develop and operate AI systems should, under the leadership of top management, monitor and record the status of preliminary and full-scale operations so that gap analysis for individual AI systems in preliminary and full-scale operations can be continuously implemented. Companies that develop AI systems should assist the monitoring conducted by companies that operate AI systems.

21. Qatar Central Bank

21.1. Artificial Intelligence Guidelines

Link: https://www.qcb.gov.qa/Services/Financial%20Technology/QCB_Artificial_Intelligence_Guideline.pdf

Link: https://www.qcb.gov.qa/Services/Financial%20Technology/QCB_Artificial_Intelligence_Guideline.pdf

21.1.1. 7.9

An Entity may develop monitoring controls to measure the fairness of an Entity's AI Model and policies when and how to initiate remedial actions.

21.1.2. 12.7

An Entity must put in place proper reporting and monitoring mechanisms to ensure that the integrity and quality of work conducted by the outsourcing service provider is maintained.

21.1.3. 13.6.4

Entities must set in-built limits and link to warning levels or auto-close routines.

21.1.4. 13.7.2

Human monitoring must provide information that a Supervisor can respond to in a manageable time frame to adjust parameters during the operation of the Algorithm.

21.1.5. 15.15

To the extent that it is strictly necessary for the purposes of ensuring Bias monitoring, detection, and correction in relation to the High-Risk AI Systems, the Providers of such systems may process personal data, subject to compliance with the Qatar Law No. (13) of 2016 on Personal Data Privacy Protection. To the extent feasible, using of security and privacy-preserving measures, such as pseudonymization, or encryption where anonymization is used.

21.1.6. 17.2

An Entity must develop a program for AI Trust, Risk and Security Management (AI TRiSM). The program must include:
- Toolsets for content anomaly detection.
- Toolsets for data protection from Providers and other third-parties.
- Toolsets for third-party system security.
- Policy for the acceptable use of AI.
- Processes for assessment of privacy, fairness and bias.

21.1.7. 17.9

An Entity that utilizes AI must deploy tools for content anomaly detection.

21.1.8. 19.1

An Entity must support the Monitoring Systems and plans for High-Risk AI Systems. This means for its own AI Systems but also accessing the information provided of all activities carried out by Providers of AI Systems to collect and review experience gained from the use of AI Systems.

21.1.9. 19.2

The Entity must ensure the compliance of the Provider to its obligations to evaluate the conformance to the predicted model outcomes of AI Systems throughout their Life Cycle, the Monitoring System must collect, document, and analyze relevant data, which may be provided by Users, or which may be collected through other sources on the performance of High-Risk AI Systems.

21.1.10. 19.3

The Entity should maintain records of its experience with AI Systems which are auditable and where relevant and considering the type of system used, should maintain on-going and up-to-date information.

21.1.11. 19.3.1

Establish audit logs and maintaining traceability of decisions and outcomes of the AI System.

22. Smart Dubai (UAE)

22.1. AI Ethics Principles & Guidelines

Link: https://www.digitaldubai.ae/docs/default-source/ai-principles-resources/ai-ethics.pdf

Link: https://www.digitaldubai.ae/docs/default-source/ai-principles-resources/ai-ethics.pdf

22.1.1. 1.2.2.7

AI operator organisations should consider working with their vendors (AI developer organisations) to continually monitor performance.

22.1.2. 1.3.1.3

Where possible given the model design, AI developer organisations should consider building in a means by which the “decision journey” of a specific outcome (i.e. the component decisions leading to it) can be logged.

23. SDAIA (Saudi Arabia)

23.1. AI Ethics Principles

Link: https://sdaia.gov.sa/en/SDAIA/about/Documents/ai-principles.pdf

Link: https://sdaia.gov.sa/en/SDAIA/about/Documents/ai-principles.pdf

23.1.1. Principle 2 – Privacy & Security - Deploy and Monitor - 1

After the deployment of the AI system, when its outcomes are realized, there must be continuous monitoring to ensure that the AI system is privacy-preserving, safe and secure. The privacy impact assessment and risk management assessment should be continuously revisited to ensure that societal and ethical considerations are regularly evaluated.

23.1.2. Principle 5 – Reliability & Safety - Deploy and Monitor - 1

Monitoring the robustness of the AI system should be adopted and undertaken in a periodic and continuous manner to measure and assess any risks related to the technicalities of the AI system (an inward perspective) as well as the magnitude of the risk posed by the system and its capabilities (an outward perspective).

23.1.3. Principle 5 – Reliability & Safety - Deploy and Monitor - 2

The model must also be monitored in a periodic and continuous manner to verify whether its operations and functions are compatible with the designed structure and frameworks. The AI system must also be safe to prevent destructive use to exploit its data and results to harm entities, individuals, or groups. It is necessary to continuously work on implementation and development to ensure system reliability.

23.1.4. Principle 7 – Accountability & Responsibility - Deploy and Monitor - 1

The responsibility and associated liability in the Deploy and Monitor step should be set clearly. The outcomes and decisions set in the build and validate step should be monitored continuously and should result in periodic performance reports.

23.1.5. Principle 7 – Accountability & Responsibility - Deploy and Monitor - 2

Predefined triggers/alerts should be defined for this step on the data and performance metrics. Setting these triggers is a rigorous process and each trigger should be assigned to the appropriate stakeholder. These triggers/alerts can be defined as part of the risk mitigation or disaster recovery procedure and may need human oversight.

24. OECD

24.1. Due Diligence Guidance for Responsible AI

Link: https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/02/oecd-due-diligence-guidance-for-responsible-ai_7831bb49/41671712-en.pdf

Link: https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/02/oecd-due-diligence-guidance-for-responsible-ai_7831bb49/41671712-en.pdf

24.1.1. Step 4 - Track implementation and results of due diligence activities

> Track the implementation and effectiveness of the enterprise's due diligence activities, i.e., its measures to identify, prevent, mitigate and, where appropriate, support remediation of adverse impacts.

25. NCSC/NSA/CISA etc

25.1. AI Data Security

Link: https://media.defense.gov/2025/May/22/2003720601/-1/-1/0/CSI_AI_DATA_SECURITY.PDF

Link: https://media.defense.gov/2025/May/22/2003720601/-1/-1/0/CSI_AI_DATA_SECURITY.PDF

25.1.1. 4.3 Input and Output Monitoring

> Monitor the AI system inputs and outputs to verify the model is performing as expected. [9] Regularly update your model using current data. Utilize meaningful statistical methods that measure expected dataset metrics and compare the distribution of the training data to the test data to help determine if data drift is occurring. [7]

26. OpenAI

26.1. Preparedness Framework

Link: https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf

Link: https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf

26.1.1. Safeguards Against Malicious Users - Usage Monitoring

> If a model does not refuse and provides assistance to harmful tasks, monitors can stop or catch malicious users before they have achieved an unacceptable scale of harm, through a combination of automated and human detection and enforcement within an acceptable time frame.

26.2. Safety Best Practices

Link: https://platform.openai.com/docs/guides/safety-best-practices

Link: https://platform.openai.com/docs/guides/safety-best-practices

26.2.1. Implement safety identifiers

> Sending safety identifiers in your requests can be a useful tool to help OpenAI monitor and detect abuse. This allows OpenAI to provide your team with more actionable feedback in the event that we detect any policy violations in your application.
>
> A safety identifier should be a string that uniquely identifies each user. Hash the username or email address in order to avoid sending us any identifying information. If you offer a preview of your product to non-logged in users, you can send a session ID instead.
>
> Include safety identifiers in your API requests with the safety_identifier parameter

27. IBM

27.1. IBM Framework for Securing Generative AI

Link: https://www.ibm.com/products/tutorials/ibm-framework-for-securing-generative-ai

Link: https://www.ibm.com/products/tutorials/ibm-framework-for-securing-generative-ai

27.1.1. Establish governance

> IBM provides not only security for AI but also operational governance for AI. IBM is leading the industry in AI governance to reach trustworthy AI models. As organizations offload operational business processes to AI, they need to make sure the AI system is not drifting and is acting as expected. This makes operational guardrails central to an effective AI strategy. A model that operationally strays from what it was designed to do can introduce the same level of risk as an adversary that's compromised your infrastructure.

28. NIST

28.1. AI RMF 1.0

Link: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

Link: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

28.1.1. MEASURE 2.4

The functionality and behavior of the AI system and its components – as identified in the MAP function – are monitored when in production.

28.1.2. MANAGE 3.1

AI risks and benefits from third-party resources are regularly monitored, and risk controls are applied and documented.

28.1.3. MANAGE 3.2

Pre-trained models which are used for development are monitored as part of AI system regular monitoring and maintenance.

28.1.4. MANAGE 4.1

Post-deployment AI system monitoring plans are implemented, including mechanisms for capturing and evaluating input from users and other relevant AI actors, appeal and override, decommissioning, incident response, recovery, and change management.

28.2. SP 800-218A

Link: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-218A.pdf

Link: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-218A.pdf

28.2.1. PO.5.1

Separate and protect each environment involved in software development.

Monitor, track, and limit resource usage and rates for AI model users during model development.

Only store sensitive data used during AI model development, including production data, within organization-approved environments and locations within those environments.

Protect all training pipelines, model registries, and other components within the environments according to the principle of least privilege.

Continuously monitor training-related activity in pipelines and model modifications in the model registry.

Follow recommended practices for securely configuring each environment.

Continuously monitor each environment for plaintext secrets.

28.2.2. PO.5.3

Continuously monitor software execution performance and behavior in software development environments to identify potential suspicious activity and other issues.

Perform continuous security monitoring for all development environment components that host an AI model or related resources (e.g., model APIs, weights, configuration parameters, training datasets).

Continuous monitoring and analysis tools should generate alerts when detected activity involving an AI model passes a risk threshold or otherwise merits additional investigation.

28.2.3. RV.1.1

Gather information from software acquirers, users, and public sources on potential vulnerabilities in the software and third-party components that the software uses, and investigate all credible reports.

Log, monitor, and analyze all inputs and outputs for AI models to detect possible security and performance issues (see PO.5.3).

Make the users of AI models aware of mechanisms for reporting potential security and performance issues.

Monitor vulnerability and incident databases for information on AI-related concerns, including the machine learning frameworks and libraries used to build AI models.

28.2.4. RV.2.1

Analyze each vulnerability to gather sufficient information about risk to plan its remediation or other risk response.

28.3. AI 800-1

Link: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd2.pdf

Link: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd2.pdf

28.3.1. Practice 5.1: Implement safeguards proportionate to the model's misuse risk - 1

Establish evidence of safeguards' effectiveness before relying on them to prevent misuse risks, such as by red-teaming (Practice 4.2) and monitoring the efficacy of safeguards for proxy models (Practice 1.1).

28.3.2. Practice 6.1: Monitor for evidence of misuse - 1

Monitor APIs, model hosting platforms, and other distribution channels for misuse while maintaining privacy of users.

28.3.3. Practice 6.1: Monitor for evidence of misuse - 2

Build or procure systems to enable automated detection of misuse.

28.3.4. Practice 6.1: Monitor for evidence of misuse - 5

Consider tiered methods of detection, such as scanning for misuse using less costly automated methods and then validating through direct human intervention, to help prioritize limited resources, improve privacy, and increase coverage.

28.3.5. Practice 6.1: Monitor for evidence of misuse - 6

Collaborate with other actors, such as content distribution platforms, downstream model adapters, cloud service providers, third-party researchers, and law enforcement officials, to track real-world incidents of misuse.

28.4. IR 8596: Cybersecurity Framework Profile for Artificial Intelligence (Cyber AI Profile): NIST Community Profile

Link: https://csrc.nist.gov/pubs/ir/8596/iprd

Link: https://csrc.nist.gov/pubs/ir/8596/iprd

28.4.1. PR.PS-04

Log records are generated and made available for continuous monitoring

28.4.2. PR.PS-06

Secure software development practices are integrated, and their performance is monitored throughout the software development life cycle

28.4.3. DE.CM-01

Networks and network services are monitored to find potentially adverse events

28.4.4. DE.CM-02

The physical environment is monitored to find potentially adverse events

28.4.5. DE.CM-03

Personnel activity and technology usage are monitored to find potentially adverse events

28.4.6. DE.CM-06

External service provider activities and services are monitored to find potentially adverse events

28.4.7. DE.CM-09

Computing hardware and software, runtime environments, and their data are monitored to find potentially adverse events

28.4.8. DE.AE-02

Potentially adverse events are analyzed to better understand associated activities

28.4.9. DE.AE-03

Information is correlated from multiple sources

28.4.10. DE.AE-04

The estimated impact and scope of adverse events are understood

28.4.11. RS.AN-03

Analysis is performed to establish what has taken place during an incident and the root cause of the incident

28.4.12. RS.AN-07

Incident data and metadata are collected, and their integrity and provenance are preserved

28.5. AI 100-2e2025: Adversarial Machine Learning A Taxonomy and Terminology of Attacks and Mitigations

Link: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf

Link: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf

28.5.1. 3.3.3 Direct Prompting Attacks - Interventions during deployment (5) - Detecting and terminating harmful interactions

Rather than preventing harmful model generations, AI systems may be able to detect these generations and terminate interactions. Several open [5, 6, 154] and closed [18, 204, 313] solutions have explored LLM-based detection systems with distinctly prompted and/or fine-tunedmodelsthat classify userinput and/ormodel output as harmful or undesirable. These may provide supplementary assurance through a defense-in-depth philosophy. However, these detection systems are also vulnerable to attacks [235] and may have correlated failures to the main models that they are monitoring. Some lines of research have investigated constraining the space of generationsto enable deterministic guardrails[306].Early work suggeststhatinterpretability-based techniques can also be used to detect anomalous input [31], as well as keyword- or perplexity-based defenses [9, 164].

28.5.2. 3.3.3 Direct Prompting Attacks - Interventions during deployment (5) - Monitoring and response

Following deployment, monitoring and logging of user activity may allow model deployersto identify and respond to instances of attempted and successful direct prompt injection attacks [266]. This response could include banning or otherwise acting against users if their intentions appear malicious, or remediating the prompt injection vulnerability in the event of a successful attack. Standard user- or organization-level vetting or identity verification procedures, as well as clear incentive mechanisms(such as a policy of restricting model access in response to violations) may enhance the efficacy of this mitigation.

29. CoSAI

29.1. Establish Risks and Controls for the AI Supply Chain

Link: https://github.com/cosai-oasis/ws1-supply-chain/blob/main/risks-and-controls-for-the-ai-supply-chain-v1.md

Link: https://github.com/cosai-oasis/ws1-supply-chain/blob/main/risks-and-controls-for-the-ai-supply-chain-v1.md

29.1.1. 3.2.1 Supply Chain Security for Data - Vector Space Attacks

Exploiting similarities in embedding space to manipulate system behavior:
• Monitor embedding space for anomalies
• Implement adversarial testing
• Apply dimensional security controls

29.1.2. 3.2.2 Model - Data Drift

Gradual changes in input data distribution over time that degrade model performance, or scenarios where external knowledge bases become outdated relative to real-world conditions:
Implement continuous distribution monitoring paired with scheduled model retraining protocols to adapt to evolving data patterns. Establish baseline performance metrics and automated drift detection thresholds.

29.1.3. 3.2.4 Infrastructure - Feedback Loop Exploitation

Adversarial manipulation of model training or reinforcement learning processes by introducing carefully crafted inputs designed to gradually shift model behavior in directions favorable to attackers:
Deploy anomaly detection for feedback patterns, implement robust validation gates for training data, and establish monitoring systems for unexpected model behavior drift.

29.2. AI Incident Response Framework

Link: https://github.com/cosai-oasis/ws2-defenders/blob/main/incident-response/AI%20Incident%20Response.md

Link: https://github.com/cosai-oasis/ws2-defenders/blob/main/incident-response/AI%20Incident%20Response.md

29.2.1. 3.2. Monitoring and Telemetry

To protect AI systems from evolving threats, telemetry must capture a comprehensive set of signals spanning model inference behavior, prompt and output risks, content safety, and model integrity. This includes monitoring for prompt injection, output manipulation, knowledge base poisoning, model drift, unauthorized tool or API usage, and other adversarial activities. Tracking agent workflows, context exchanges, and system lifecycles provides visibility into misuse, operational anomalies, and attacks. This data enables organizations to safeguard AI pipelines, ensure compliance, and strengthen incident detection and response with traceable, structured observability data.

29.2.2. 3.3.1. Preparation Phase - Monitoring Infrastructure

• Input/output logging
• System prompt versioning
• Configuration management
• Authentication tracking
• API usage analysis
• Resource utilization
• Data access patterns

29.2.3. 3.3.2. Detection and Analysis Phase - Detection Mechanisms - Automated Monitoring

• LLM-based classifiers for suspicious interactions
• Semantic similarity checks
• Token usage anomaly detection
• User feedback signal monitoring

29.2.4. 3.3.2. Detection and Analysis Phase - Initial Triage - Incident Verification

• Confirm security incident status
• Classify per AI attack categories (Section 2)
• Determine affected components

29.2.5. 3.3.3. Containment, Eradication, and Recovery Phase - Recovery Procedures - Enhanced Monitoring

• Heightened monitoring deployment
• Normal/abnormal metrics
• Additional logging
• Scheduled reviews

29.3. Model Context Protocol (MCP) Security

Link: https://github.com/cosai-oasis/ws4-secure-design-agentic-systems/blob/main/model-context-protocol-security.md

Link: https://github.com/cosai-oasis/ws4-secure-design-agentic-systems/blob/main/model-context-protocol-security.md

29.3.1. 3.2.10 Logging

Implement at all layers (MCP host, client and server) the capability to store a log of what tools have been decided to use, with which parameters, and as a result of which prompt. Having a log of the decisions made is crucial in order to troubleshoot or perform forensics in case of a security event.

Leverage the use of centralization tools like MCP gateways or proxies, for example, between the MCP clients and the MCP servers, to centralize there key functionality (e.g. logging) and avoid the need to implement it on each component.

30. Personal Data Protection Commission Singapore (PDPC)

30.1. Model Artificial Intelligence Governance Framework Second Edition

Link: https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf

Link: https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf

30.1.1. 1. Clear roles and responsibilities for the ethical deployment of AI - c) (ii)

Maintenance, monitoring, documentation and review of the AI models that have been deployed, with a view to taking remediation measures where needed.

30.1.2. 2. Risk management and internal controls - b) (ii)

Establishing monitoring and reporting systems as well as processes to ensure that the appropriate level of management is aware of the performance of and other issues relating to the deployed AI. Where appropriate, the monitoring can include autonomous monitoring to effectively scale human oversight. AI systems can be designed to report on the confidence level of their predictions, and explainability features could focus on why the AI model had a certain level of confidence.

30.1.3. Repeatability - b)

Assessing how exceptions can be identified and handled when decisions are not repeatable, e.g. when randomness has been introduced by design;

30.1.4. Repeatability - e)

Identifying and accounting for changes over time to ensure that models trained on time-sensitive data remain relevant.

31. MITRE

31.1. ATLAS Framework

Link: https://atlas.mitre.org/mitigations

Link: https://atlas.mitre.org/mitigations

31.1.1. AML.M0024 - AI Telemetry Logging

Implement logging of inputs and outputs of deployed AI models. When deploying AI agents, implement logging of the intermediate steps of agentic actions and decisions, data access and tool use, and identity of the agent. Monitoring logs can help to detect security threats and mitigate impacts.

Additionally, having logging enabled can discourage adversaries who want to remain undetected from utilizing AI resources.

31.2. SAFE-AI

Link: https://atlas.mitre.org/pdf-files/SAFEAI_Full_Report.pdf

Link: https://atlas.mitre.org/pdf-files/SAFEAI_Full_Report.pdf

31.2.1. Zero-day exploits

> AI-enabled systems typically have failure modes that can be difficult to characterize, and those modes and their causes can often be poorly understood (or even unknown). For this reason, it is critically important to continuously monitor performance once a system is deployed and proactively investigate reports of anomalous events (using, for example, red team exercises to discover and assess failure modes). Information sharing is a key component of this monitoring activity. Information about errors and other potential precursors to security abuses must be shared with incident databases, other organizations with similar systems, and system users and stakeholders.

32. ETSI

32.1. EN 304 223 - Securing Artificial Intelligence (SAI); Baseline Cyber Security Requirements for AI Models and Systems

Link: https://www.etsi.org/deliver/etsi_en/304200_304299/304223/02.01.01_60/en_304223v020101p.pdf

Link: https://www.etsi.org/deliver/etsi_en/304200_304299/304223/02.01.01_60/en_304223v020101p.pdf

32.1.1. Provision 5.4.2-1

> System Operators shall log system and user actions to support security compliance, incident investigations, and vulnerability remediation.

32.1.2. Provision 5.4.2-2

> System Operators should analyse their logs to ensure that AI models continue to produce desired outputs and to detect anomalies, security breaches, or unexpected behaviour over time (such as due to data drift or data poisoning).

32.1.3. Provision 5.4.2-3

> System Operators and Developers should monitor internal states of their AI systems where this could better enable them to address security threats, or to enable future security analytics.

32.1.4. Provision 5.4.2-4

> System Operators and Developers should monitor the performance of their models and system over time so that they can detect sudden or gradual changes in behaviour that could affect security.

32.2. TR 104 128 - Securing Artificial Intelligence (SAI); Guide to Cyber Security for AI Models and Systems

Link: https://www.etsi.org/deliver/etsi_tr/104100_104199/104128/01.01.01_60/tr_104128v010101p.pdf

Link: https://www.etsi.org/deliver/etsi_tr/104100_104199/104128/01.01.01_60/tr_104128v010101p.pdf

32.2.1. Provision 5.4.2-1

> "System Operators shall log system and user actions to support security compliance, incident investigations, and vulnerability remediation." (ETSI TS 104 223 [i.1])
>
> Related threats/risks:
> Insufficient logging limits incident investigation and compliance enforcement, weakening the ability to detect and respond to security incidents.
>
> Example Measures/Controls 1:
> Implement Comprehensive Logging for Security and Compliance: Establish a logging framework that captures key aspects of system behaviour, including user interactions, access events, data flows, and model outputs, ensuring logs support compliance and security standards.
>
> Example Measures/Controls 2:
> Ensure Secure Storage and Retention of Logs: Implement secure storage mechanisms, such as encryption (both at rest and in transit), to protect logs from unauthorized access. Define a retention policy that aligns with compliance obligations and supports security incident investigations. Avoid logging personal data unless strictly necessary; if personal data needs to be logged, ensure it is anonymized or obfuscated and managed in compliance with applicable privacy laws (e.g. UK GDPR, CCPA). Periodically review and update the retention policy to address evolving compliance and operational requirements.
>
> Example Measures/Controls 3:
> Establish Routine Log Analysis for Model Validation: Define a regular schedule for log analysis to assess model output consistency, detect anomalies, and verify that outputs align with desired outcomes.

32.2.2. Provision 5.4.2-2

"System Operators should analyse their logs to ensure that AI models continue to produce desired outputs and to detect anomalies, security breaches, or unexpected behaviour over time (such as due to data drift or data poisoning)." (ETSI TS 104 223 [i.1])

Related threats/risks:
Without regular log analysis, security breaches or model issues can go undetected, potentially leading to incorrect outputs or system vulnerabilities

Example Measures/Controls:
Implement Alerts for Anomalous Model Behaviour: Set up alerts to notify operators of unexpected behaviour, such as unusual outputs, abnormal input patterns, or significant deviations from historical performance.

32.2.3. Provision 5.4.2-3

"System Operators and Developers should monitor internal states of their AI systems where they feel this could better enable them to address security threats, or to enable future security analytics." (ETSI TS 104 223 [i.1])

Related threats/risks:
Lack of internal state monitoring can delay detection of security threats or model drift, increasing risks of unauthorized modifications and system failure.

Example Measures/Controls 1:
Implement Monitoring of Key Internal States: Identify and monitor critical internal states of the AI system, such as hidden layers, attention weights, or feature importance, which could provide early indicators of security threats.

Example Measures/Controls 2:
Use Secure Storage for Internal State Data: Store data from monitored internal states securely, ensuring that sensitive internal metrics are protected from unauthorized access.

Example Measures/Controls 3:
Track and Benchmark Model Performance Metrics: Define and monitor key performance metrics for the AI system, such as statistical accuracy, factual correctness, response time, and error rate establishing benchmarks to detect deviations.

32.2.4. Provision 5.4.2-4

"System Operators and Developers should monitor the performance of their models and system over time so that they can detect sudden or gradual changes in behaviour that could affect security." (ETSI TS 104 223 [i.1])

Related threats/risks:
Failing to monitor performance over time can hide behavioural shifts, making the system more vulnerable to degradation, attacks, and inconsistencies

Example Measures/Controls:
Implement Drift Detection to Identify Behavioural Shifts: Use drift detection tools to identify shifts in model behaviour due to changing data patterns or environmental factors, allowing proactive responses.

32.3. TR 104 048 - Securing Artificial Intelligence (SAI); Data Supply Chain Security

Link: https://www.etsi.org/deliver/etsi_tr/104000_104099/104048/01.01.01_60/tr_104048v010101p.pdf

Link: https://www.etsi.org/deliver/etsi_tr/104000_104099/104048/01.01.01_60/tr_104048v010101p.pdf

32.3.1. 6.1.2 Cybersecurity hygiene - 7

Following deployment of a service, auditing and logging enables the detection of possible anomalies. In an AI context, this could include a representation of the inputs to the ML model. Though significant research has been conducted on mapping established software security practices to AI environments, these practices remain less developed in the AI domain.

32.3.2. 6.5 Analysis - Logging

Logging at all stages of processing and deployment, including collecting model telemetry.

32.4. TR 104 222 - Securing Artificial Intelligence; Mitigation Strategy Report

Link: https://www.etsi.org/deliver/etsi_tr/104200_104299/104222/01.02.01_60/tr_104222v010201p.pdf

Link: https://www.etsi.org/deliver/etsi_tr/104200_104299/104222/01.02.01_60/tr_104222v010201p.pdf

32.4.1. 6.2.3 - 3

MagNet [i.74] is a framework of adversarial example detection and recovery. It integrates a set of detectors and a reformer. When an adversarial example has significant perturbation from the benign example, it can be detected by detectors. When the perturbation is less significant such that it cannot be detected by detectors, the reformer can restore the adversarial example to the benign example. One reformer example is to use autoencoder, which is an unsupervised learning algorithm. The autoencoder is trained by the training dataset without labels such that the output of autoencoder is close to the input. Using autoencoder as the reformer in MagNet does not change benign examples much but push adversarial examples to benign examples. Yet MagNet is less effective against the CW attacks [i.75] and has considerable computing overhead.

32.4.2. 6.3.3 - 1

Extraction warning [i.82]: a cloud-based extraction monitor to inform model owners about the status of model extraction by both individual and colluding adversaries using a decision tree. The monitor observes the query response pairs of each adversary to the deployed decision tree model, and incrementally learns a local decision tree based on these tuples. The detection can be done at fixed time intervals or after the deployed model has answered certain number of queries. Two novel metrics are proposed to measure the model learning rate of adversaries. A first metric is based on entropy and measures the information gain of a decision tree with respect to a validation set provided by the model owner. The second metric is based on maintaining a compact model summary corresponding to each adversary with increasing number of queries. The compact model summary represents the boundaries of regions within the feature space that the adversary may have learnt for each class. The monitor assesses the coverage of summaries within their respective classes to compute the overall learning rate of an adversary. The objective is to detect if adversaries can, either individually or jointly, reconstruct a model that yields an accuracy beyond a given threshold from the queries obtained from the addressed model.

32.4.3. 6.3.3 - 3

Extraction query distribution identification [i.84]: Protecting Against DNN Model Stealing Attacks (PRADA) proposed a method that detects suspicious queries to the addressed DNN model by analysing the distribution of consecutive API queries and identifying whether the distribution deviates from benign behaviour, i.e. from a normal (Gaussian) distribution. PRADA collects stateful information of queries in addressed model prediction APIs and can detect all prior model extraction attacks with no false positives. This defense does not require any knowledge about the addressed model, nor about the training dataset, and it is agnostic to the distribution of the sample dataset used in the queries. However, the method can be circumvented by mimicking benign query distributions.

32.5. SAI 002 - Securing Artificial Intelligence (SAI); Data Supply Chain Security

Link: https://www.etsi.org/deliver/etsi_gr/SAI/001_099/002/01.01.01_60/gr_SAI002v010101p.pdf

Link: https://www.etsi.org/deliver/etsi_gr/SAI/001_099/002/01.01.01_60/gr_SAI002v010101p.pdf

32.5.1. 6.1.2 Cybersecurity hygiene - 7

Following deployment of a service, auditing and logging enables the detection of possible anomalies. In an AI context, this could include a representation of the inputs to the ML model. Though significant research has been conducted on mapping established software security practices to AI environments, these practices remain less developed in the AI domain [i.14].

32.5.2. 6.5 - Logging

Logging at all stages of processing and deployment, including collecting model telemetry.