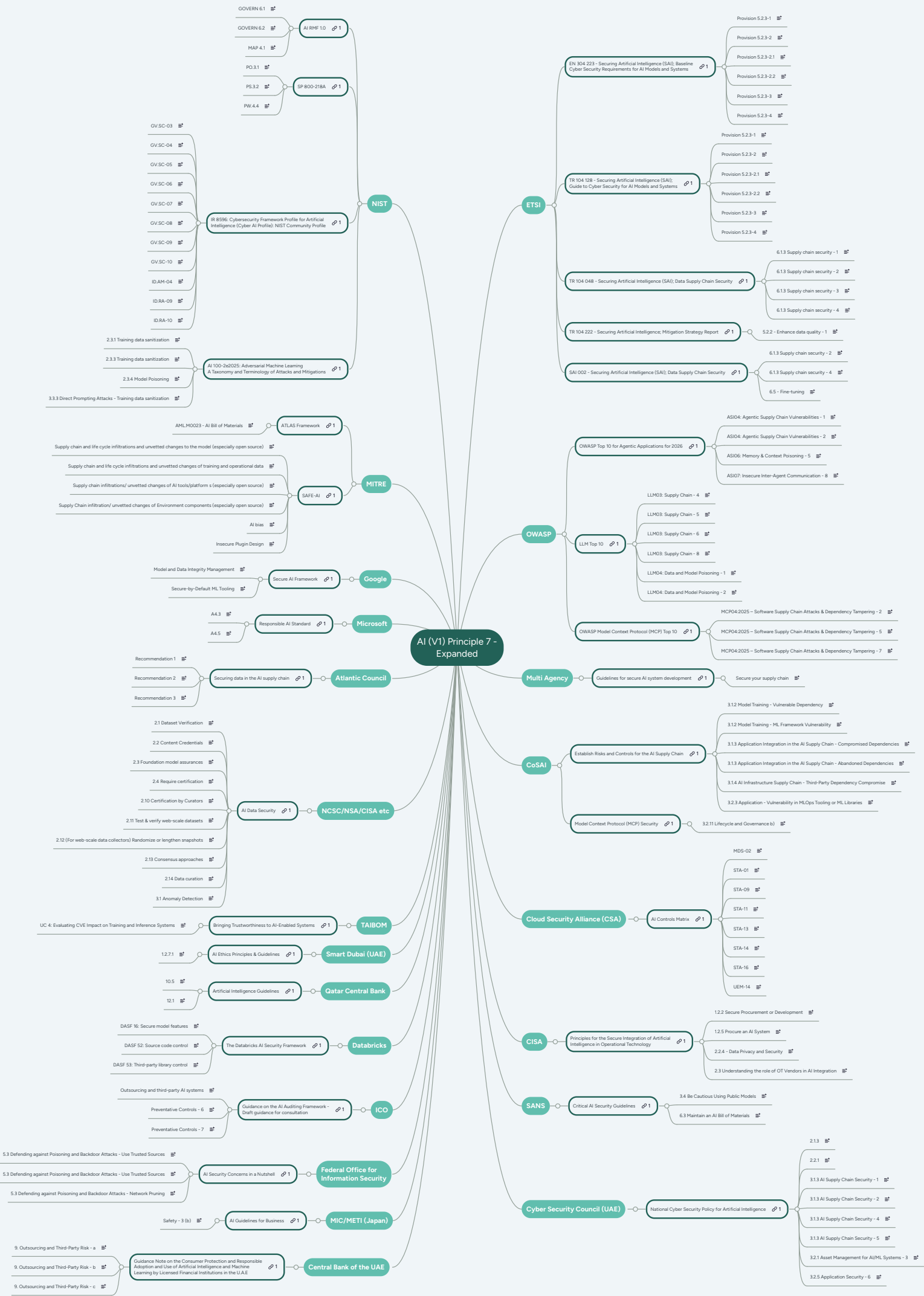




AI (V1) Principle 7 - Expanded

1. ETSI

1.1. 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

1.1.1. Provision 5.2.3-1

Developers and System Operators shall follow secure software supply chain processes for their AI model and system development.

1.1.2. Provision 5.2.3-2

System Operators that choose to use or adapt any models, or components, which are not well-documented or secured shall be able to justify their decision to use such models or components through documentation (for example if there was no other supplier for said component).

1.1.3. Provision 5.2.3-2.1

In this case, Developers and System Operators shall have mitigating controls and undertake a risk assessment linked to such models or components.

1.1.4. Provision 5.2.3-2.2

System Operators shall share this documentation with End-users in an accessible way.

1.1.5. Provision 5.2.3-3

Developers and System Operators shall re-run evaluations on released models that they intend on using.

1.1.6. Provision 5.2.3-4

System Operators shall communicate their intention to update models to End-users in an accessible way prior to models being updated.

1.2. 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

1.2.1. Provision 5.2.3-1

"Developers and System Operators shall follow secure software supply chain processes for their AI model and system development." (ETSI TS 104 223 [i.1])

Related threats/risks:

Without secure software supply chain processes, AI systems are vulnerable to risks like supply chain attacks, insertion of malicious components, and dependency issues, including models and datasets, which can compromise AI integrity and security. Lack of accountability across the supply-chain for jurisdiction-specific regulations can lead to data breaches and lack of regulatory compliance.

Example Measures/Controls 1:

Safeguard Provenance and Transparency: Mandate in internal standards that components can only be sourced by trusted and approved sources, documenting source, version, licencing, history, and other related artifacts (e.g. Model Card for models); use checksums to verify integrity. SBOMs can help automated creation of signed attestations of all components and their metadata. This can help safeguard transparency and provenance with non-repudiation and component tampering checks. SBOMs cover libraries and system components but not models or datasets. AI and ML BOMs are an emerging field. Monitor progress in standards such as Cyclone DX ML BOM [i.106] and introduce similar scans when tools emerge. In the meanwhile, rely on ML Ops to enforce model and dataset provenance and use file checksum to verify integrity and provenance.

Example Measures/Controls 2:

Apply Vulnerability Management: Implement a vulnerability management process that includes regular (e.g. daily) scanning of third-party components for known vulnerabilities and timely patching to mitigate risks.

Example Measures/Controls 3:

Adopt Secure Supply Chain Frameworks: Follow the organization's preferred secure supply chain guidance or standard for all stages of AI development, from sourcing and integrating components to testing and deployment.

1.2.2. Provision 5.2.3-2

"System Operators that choose to use or adapt any models, or components, which are not well-documented or secured shall be able to justify their decision to use such models or components through documentation (for example if there was no other supplier for said component)." (ETSI TS 104 223 [i.1])

Related threats/risks:

Utilizing poorly documented or untrusted AI components because of some of their unique features introduces potential vulnerabilities, increasing risks of data leaks or unexpected behaviour.

Example Measures/Controls:

Document Justification for Untrusted Components: When using poorly documented or untrusted components, document the justification, including alternative evaluations, and the absence of better suppliers.

1.2.3. Provision 5.2.3-2.1

"In this case, Developers and System Operators shall have mitigating controls and undertake a risk assessment linked to such models or components." (ETSI TS 104 223 [i.1])

Related threats/risks:

Utilizing poorly documented or untrusted AI components because of some of their unique features introduces potential vulnerabilities, increasing risks of data leaks or unexpected behaviour.

Example Measures/Controls:

Evaluate and Mitigate Risks for Untrusted Components: Perform a risk assessment to identify potential risks associated with unsecured AI components, specifying, and implementing mitigating controls to minimize vulnerabilities.

1.2.4. Provision 5.2.3-2.2

"System Operators shall share this documentation with End-users in an accessible way." (ETSI TS 104 223 [i.1])

Related threats/risks:

Utilizing poorly documented or untrusted AI components because of some of their unique features introduces potential vulnerabilities, increasing risks of data leaks or unexpected behaviour.

Example Measures/Controls:

Share Documentation with End-Users: Share documentation of non-compliant components with end-users, detailing risks, and justifications for transparency.

1.2.5. Provision 5.2.3-3

"Developers and System Operators shall re-run evaluations on released models that they intend on using." (ETSI TS 104 223 [i.1])

Related threats/risks:

Without reusable evaluations, periodic re-evaluation can be difficult, resulting into performance degradation, bias, inaccuracies, or security vulnerabilities go unnoticed in new releases.

Example Measures/Controls:

Create and Maintain Appropriate and Reusable Model Evaluation Suites: Create appropriate model evaluation suites to assess performance, accuracy, security, and potential drift when required.

1.2.6. Provision 5.2.3-4

"System Operators shall communicate the intention to update models to End-users in an accessible way prior to models being updated." (ETSI TS 104 223 [i.1])

Related threats/risks:

Lack of transparency around model updates can make it challenging for users to use a new model version, especially in evaluating data protection compliance. This creates availability and operational issues for end-users relying on certain model feature or behaviour.

Example Measures/Controls:

Provide Advance Notice of Model Updates: Notify end-users of model updates at least one month in advance, including any potential impacts on model performance or functionality and offer previews to test changes for at least a month. Ensure notices are accessible.

1.3. 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

1.3.1. 6.1.3 Supply chain security - 1

Understanding the risks associated with the supply chain, particularly for high-value components such as datasets. This includes understanding the security posture of the suppliers.

1.3.2. 6.1.3 Supply chain security - 2

Setting minimum security standards for the supply chain and communicating these to the suppliers.

1.3.3. 6.1.3 Supply chain security - 3

Building data and model security considerations into the contracting processes.

1.3.4. 6.1.3 Supply chain security - 4

Adopting a view of supply chain security as a continuous process.

1.4. 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

1.4.1. 5.2.2 - Enhance data quality - 1

One mitigation is to protect the data supply chain against manipulations to thwart attacks which tamper with or degrade the quality of training data. The risk for incorrect or manipulated data is much lower if data is sampled from a controlled environment.

1.5. 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

1.5.1. 6.1.3 Supply chain security - 2

Setting minimum security standards for the supply chain and communicating these to the suppliers.

1.5.2. 6.1.3 Supply chain security - 4

Adopting a view of supply chain security as a continuous process.

1.5.3. 6.5 - Fine-tuning

And/or regular retraining of models with locally-verified or otherwise trusted data, where possible.

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. ASI04: Agentic Supply Chain Vulnerabilities - 1

Provenance and SBOMs, AIBOMs: Sign and attest manifests, prompts, and tool definitions; require and operationalize SBOMs, AIBOMs with periodic attestations; maintain inventory of AI components; use curated registries and block untrusted sources.

2.1.2. ASI04: Agentic Supply Chain Vulnerabilities - 2

Dependency gatekeeping: Allowlist and pin; scan for typosquats (PyPI, npm, LangChain, LlamaIndex); verify provenance before install or activation; auto-reject unsigned or unverified.

2.1.3. ASI06: Memory & Context Poisoning - 5

Provenance and anomalies: Require source attribution and detect suspicious updates or frequencies.

2.1.4. ASI07: Insecure Inter-Agent Communication - 8

Attested registry and agent verification: Use registries or marketplaces that provide digital attestation of agent identity, provenance, and descriptor integrity. Require signed agent cards and continuous verification before accepting discovery or coordination messages. Leverage the PKI trusted root certificate registries to enable robust agent verification and attestation of critical attributes.

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 - 4

Maintain an up-to-date inventory of components using a Software Bill of Materials (SBOM) to ensure you have an up-to-date, accurate, and signed inventory, preventing tampering with deployed packages. SBOMs can be used to detect and alert for new, zero-date vulnerabilities quickly. AI BOMs and ML SBOMs are an emerging area and you should evaluate options starting with OWASP CycloneDX

2.2.2. LLM03: Supply Chain - 5

To mitigate AI licensing risks, create an inventory of all types of licenses involved using BOMs and conduct regular audits of all software, tools, and datasets, ensuring compliance and transparency through BOMs. Use automated license management tools for real-time monitoring and train teams on licensing models. Maintain detailed licensing documentation in BOMs.

2.2.3. LLM03: Supply Chain - 6

Only use models from verifiable sources and use third-party model integrity checks with signing and file hashes to compensate for the lack of strong model provenance. Similarly, use code signing for externally supplied code.

2.2.4. LLM03: Supply Chain - 8

AAnomaly detection and adversarial robustness tests on supplied models and data can help detect tampering and poisoning as discussed in "LLM04 Data and Model Poisoning; ideally, this should be part of MLOps and LLM pipelines; however, these are emerging techniques and may be easier to implement as part of red teaming exercises.

2.2.5. LLM04: Data and Model Poisoning - 1

Track data origins and transformations using tools like OWASP CycloneDX or ML-BOM. Verify data legitimacy during all model development stages.

2.2.6. LLM04: Data and Model Poisoning - 2

Vet data vendors rigorously, and validate model outputs against trusted sources to detect signs of poisoning.

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. MCP04:2025 – Software Supply Chain Attacks & Dependency Tampering - 2

Build SBOM / CBOM Visibility Generate SBOM (software bill of materials) and CBOM (cryptographic bill of materials) snapshots for each MCP server + plugin package Store SBOM alongside deployments for auditing + incident response

2.3.2. MCP04:2025 – Software Supply Chain Attacks & Dependency Tampering - 5

Dependency Scanning

- Apply SCA (software composition analysis) + code scanning tools to detect:
- Known CVEs
- Malicious indicators
- Poisoned transitive dependencies

2.3.3. MCP04:2025 – Software Supply Chain Attacks & Dependency Tampering - 7

Supply-Chain Governance

- Maintain vendor risk profiles
- Require suppliers to provide signed attestations
- Review open-source maintainers' security maturity

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. Secure your supply chain

You assess and monitor the security of your AI supply chains across a system's life cycle, and require suppliers to adhere to the same standards your own organisation applies to other software. If suppliers cannot adhere to your organisation's standards, you act in accordance with your existing risk management policies.

Where not produced in-house, you acquire and maintain well-secured and well-documented hardware and software components (for example, models, data, software libraries, modules, middleware, frameworks, and external APIs) from verified commercial, open source, and other third-party developers to ensure robust security in your systems.

You are ready to failover to alternate solutions for mission-critical systems, if security criteria are not met. You use resources like the NCSC's Supply Chain Guidance and frameworks such as Supply Chain Levels for Software Artifacts (SLSA) for tracking attestations of the supply chain and software development life cycles.

4. CoSAI

4.1. 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>

4.1.1. 3.1.2 Model Training - Vulnerable Dependency

Training pipeline dependencies containing security vulnerabilities:
Comprehensive supply chain visibility, regular dependency updates, dependency documentation for each training run

4.1.2. 3.1.2 Model Training - ML Framework Vulnerability

Security flaws in the core ML framework used for training or inference:
Regular framework updates, version documentation in supply chain records

4.1.3. 3.1.3 Application Integration in the AI Supply Chain - Compromised Dependencies

Malicious code insertion or vulnerability exploitation in application libraries:
Comprehensive dependency validation, SBOMs, automated scanning, verified builds

4.1.4. 3.1.3 Application Integration in the AI Supply Chain - Abandoned Dependencies

Security vulnerabilities in unmaintained application components:
Dependency health monitoring, lifecycle policies, replacement strategies

4.1.5. 3.1.4 AI Infrastructure Supply Chain - Third-Party Dependency Compromise

Supply chain attacks targeting dependencies of the hosting infrastructure:
Review dependency management practices, request SBOM for infrastructure components

4.1.6. 3.2.3 Application - Vulnerability in MLOps Tooling or ML Libraries

Security defects present in development, deployment, or monitoring components within the AI infrastructure stack that can be exploited to compromise model integrity or operational security. These vulnerabilities often exist in underlying dependencies rather than primary components:
Implement systematic update management for all libraries and MLOps tooling with comprehensive vulnerability scanning. Document all tools and libraries in standardized supply chain artifacts for each training run to enable rapid impact assessment when vulnerabilities are discovered.

4.2. 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>

4.2.1. 3.2.11 Lifecycle and Governance b)

Supply chain security requires:

- implementing software bill of materials (SBOM) tracking for all MCP components,
- using dependency pinning with hash-based verification rather than version ranges,
- deploying automated vulnerability scanning for MCP servers and dependencies,
- implementing secure software development lifecycle (SSDLC) practices for internal MCP servers,
- using reproducible builds to verify package authenticity,
- and monitoring security advisories and CVE databases for known vulnerabilities in dependencies.

5. Cloud Security Alliance (CSA)

5.1. AI Controls Matrix

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

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

5.1.1. MDS-02

Define, implement, and evaluate policies, procedures, and technical measures for the scanning of model artifacts for vulnerabilities and attacks, at each step of the service lifecycle and at each hand over point. Regularly review and update policies, procedures and technical measures to address model artifact scanning.

5.1.2. STA-01

Establish, document, approve, communicate, apply, evaluate, and maintain policies and procedures for supply chain risk management. Review and update the policies and procedures at least annually or upon significant changes.

5.1.3. STA-09

Periodically review risk factors associated with supply chain relationships.

5.1.4. STA-11

Review supply chain agreements at least annually, or upon significant changes.

5.1.5. STA-13

Implement policies requiring all service providers throughout the supply chain to comply with information security, confidentiality, access control, privacy, audit, personnel policy and service level requirements and standards.

5.1.6. STA-14

Periodically review the organization's supply chain partners' IT governance policies and procedures.

5.1.7. STA-16

Define, implement, and enforce a process for establishing a Bill of Material for the service supply chain. Review and update the Bill of Material at least annually or upon significant changes.

5.1.8. UEM-14

Define, implement and evaluate processes, procedures and technical and/or contractual measures to maintain proper security of third-party endpoints with access to organizational assets.

6. CISA

6.1. 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>

6.1.1. 1.2.2 Secure Procurement or Development

Select vendors who adhere to secure practices and develop AI systems using secure methodologies and tools.

6.1.2. 1.2.5 Procure an AI System

Critical infrastructure owners and operators should also carefully evaluate the trade-offs between different methods for sourcing an AI system

6.1.3. 2.2.4 - Data Privacy and Security

Protect proprietary and personal information within OT datasets, including by instituting protection from access abuse, intentional or inadvertent data poisoning, or dependency on synthetic, generated data.

6.1.4. 2.3 Understanding the role of OT Vendors in AI Integration

Critical infrastructure owners and operators should demand transparency and security considerations from OT vendors regarding how AI technologies are embedded into their products

7. SANS

7.1. Critical AI Security Guidelines

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

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

7.1.1. 3.4 Be Cautious Using Public Models

Some of the mechanisms used to share models can be leveraged by bad actors to introduce malicious code into the packaging used to deploy the model. In other words, the model itself has not been tampered with, but the package the model is inside of has been built with malicious actions that will be executed when the model is unpacked or called. Models also may be created by bad actors with architectural backdoors in them. The idea of a backdoor like this would be to invoke a specific behavior in response to a specific input. Once a backdoor is created inside of a model, it can be difficult, if not impossible, to remove it via fine-tuning. This becomes an issue if a user inadvertently triggers the backdoor or, if it is exposed outside the organization, a bad actor attempts to exploit it across a wide range of models.

7.1.2. 6.3 Maintain an AI Bill of Materials

LLM applications depend upon a complex underlying ecosystem for their functionality. Modeled after software bill of materials (SBOM), creation and maintenance of an AIBOM can provide better visibility into relevant aspects of the AI supply chain, including considerations of dataset and model provenance. AIBOMs contain technical details that are useful to adversaries in attacking LLM applications. Care should be taken to limit the disclosure of AIBOMs.

8. Cyber Security Council (UAE)

8.1. 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

8.1.1. 2.1.3

The entity shall identify, assess, and manage cyber security risks associated with the supply chain of AI/ML systems, ensuring appropriate controls are implemented to mitigate threats arising from third-party components, services, and dependencies.

8.1.2. 2.2.1

The entity shall maintain an up-to-date inventory of AI/ML assets and apply cyber security controls commensurate with the sensitivity, criticality, and value of each asset.

8.1.3. 3.1.3 AI Supply Chain Security - 1

The entity should establish a comprehensive AI supply chain security strategy to manage cyber security risks associated with third-party AI/ML components and services, from sourcing to decommissioning.

8.1.4. 3.1.3 AI Supply Chain Security - 2

The entity should ensure a clear understanding of its AI/ML supply chain, including the source of components, models, training data, and libraries used in its AI/ML systems.

8.1.5. 3.1.3 AI Supply Chain Security - 4

The entity should conduct regular security assessments of its AI/ML supply chain, including vendors, developers, and other third parties contributing to the AI/ML system. The entity may alternatively leverage independent third-party assurance reports (e.g., SOC 2, ISO/IEC 27001) or recognized certifications to inform its assessment.

8.1.6. 3.1.3 AI Supply Chain Security - 5

The entity should ensure that contractual agreements with vendors and other third parties address cyber security requirements for AI/ML systems.

8.1.7. 3.2.1 Asset Management for AI/ML Systems - 3

The entity should classify all AI/ML assets according to factors such as their criticality to business operations, associated risk, and sensitivity and assign appropriate cyber security controls based on this classification.

8.1.8. 3.2.5 Application Security - 6

The entity should maintain an up-to-date inventory of AI/ML applications, including their dependencies, and monitor them for any changes that could introduce cyber security risks.

9. MIC/METI (Japan)

9.1. 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

9.1.1. Safety - 3 (b)

Properly take actions such as the securement of transparency of data used for training, compliance with the legal framework, and update of AI models, within reasonable extent.

10. ICO

10.1. 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>

10.1.1. Outsourcing and third-party AI systems

When you either buy an AI solution from a third party, or outsource it altogether, you need to conduct an independent evaluation of any trade-offs as part of your due diligence process. You are also required to specify your requirements at the procurement stage, rather than addressing trade-offs ex post.

10.1.2. Preventative Controls - 6

Document contracts with third parties are clear about the role and responsibilities of third parties.

10.1.3. Preventative Controls - 7

Document policies / processes for dealing with third parties and evidence of the due diligence of information security completed.

11. Databricks

11.1. 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>

11.1.1. DASF 16: Secure model features

Databricks Feature Store is a centralized repository that enables data scientists to find and share features and also ensures that the same code used to compute the feature values is used for model training and inference. Unity Catalog's capabilities, such as security, lineage, table history, tagging and cross-workspace access, are automatically available to the feature table to reduce the risk of malicious actors manipulating the features that feed into ML training.

11.1.2. DASF 52: Source code control

Databricks' Git Repository integration supports effective code and third-party libraries management, enhancing customer control over their development environment.

11.1.3. DASF 53: Third-party library control

Databricks' library management system allows administrators to manage the installation and usage of third-party libraries effectively. This feature enhances the security and efficiency of systems, pipelines and data by giving administrators precise control over their development environment.

12. Federal Office for Information Security

12.1. AI Security Concerns in a Nutshell

Link: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Practical_AI-Security_Guide_2023.pdf?__blob=publicationFile&v=5

Link: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Practical_AI-Security_Guide_2023.pdf?__blob=publicationFile&v=5

12.1.1. 5.3 Defending against Poisoning and Backdoor Attacks - Use Trusted Sources

Depending on the security requirements for the use case, it is essential to make adequate efforts to ensure that the supply chain and the sources of training data, models, and code are known and trustworthy. Publicly available models could contain backdoors.

12.1.2. 5.3 Defending against Poisoning and Backdoor Attacks - Use Trusted Sources

Retraining a model with benign training samples, if available, reduces the probability of backdoors being successful [18]. The degree of success depends on the size and quality of the clean dataset [22]. Research suggests that even with a small retraining dataset, the vulnerability of a model to backdoor attacks significantly drops, while its accuracy may be slightly reduced [22].

12.1.3. 5.3 Defending against Poisoning and Backdoor Attacks - Network Pruning

An autoencoder is trained with a benign dataset whose feature distribution is close to the training dataset. As a result, the trained autoencoder may be able to detect manipulated data samples that lie outside of the learned distribution [22].

13. Central Bank of the UAE

13.1. 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>

13.1.1. 9. Outsourcing and Third-Party Risk - a

Where LFIs rely on third-party vendors or cloud service providers for AI and ML models, products or solutions, due diligence should be conducted on the provider's reputation in the field of AI, governance, security and data-protection practices. Contracts should include provisions that ensure access to relevant information, audit rights and compliance with CBUAE requirements.

13.1.2. 9. Outsourcing and Third-Party Risk - b

The outsourcing of any AI system and/or engagement with third party providers of AI should only be carried out with appropriate diligence, assessment and it should be considered if only to be done by way of Board of Director and/or Senior Management approval, with such processes being appropriately documented for audit purposes.

13.1.3. 9. Outsourcing and Third-Party Risk - c

The procurement, choice and justification for selection of a third-party AI provider and deployment of their AI models/systems/technology at an LFI should be appropriate and documented, which should include annual cybersecurity reviews by independent and suitably qualified third parties and pre-deployment tests and checks to ensure appropriateness and taking into accounts needs, risks and the options available.

14. Qatar Central Bank

14.1. 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

14.1.1. 10.5

AI Systems must highlight the Provider.

14.1.2. 12.1

An Entity that outsources any activity to support its operations when developing, providing, using, or implementing an AI System should ensure regular due diligence on the outsourcing service provider is carried out (identity, legal status, activities, financial position, etc.) and obtain prior consent from QCB.

15. Smart Dubai (UAE)

15.1. 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>

15.1.1. 1.2.7.1

In the case of critical decisions, AI operator organisations should avoid using AI systems that cannot be subjected to meaningful standards of accountability and transparency.

16. TAIBOM

16.1. Bringing Trustworthiness to AI-Enabled Systems

Link: <https://arxiv.org/pdf/2510.02169>

Link: <https://arxiv.org/pdf/2510.02169>

16.1.1. UC 4: Evaluating CVE Impact on Training and Inference Systems

New CVEs may affect software libraries used in AI training or inference. Identifying affected models requires understanding which code components were used and how they relate to model artifacts.

17. NCSC/NSA/CISA etc

17.1. 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

17.1.1. 2.1 Dataset Verification

Before ingest, the consumer or curator should verify, as much as possible, that the dataset to be ingested is free of malicious or inaccurate material. Any detected abnormalities should be addressed, and suspicious data should not be stored. The dataset verification process should include a digital signature of the dataset at time of ingestion.

17.1.2. 2.2 Content Credentials

Use content credentials to track the provenance of media and other data. Content credentials are “metadata that are secured cryptographically and allow creators the ability to add information about themselves or their creative process, or both, directly to media content.... Content Credentials securely bind essential metadata to a media file that can track its origin(s), any edits made, and/or what was used to create or modify the content.... This metadata alone does not allow a consumer to determine whether a piece of content is ‘true,’ but rather provides contextual information that assists in determining the authenticity of the content.” [24]

17.1.3. 2.3 Foundation model assurances

In the case where a consumer is not ingesting a dataset but a foundation model trained by another party, the developers of the foundation model need to be able to provide assurances regarding the data and sources used and certify that their training data did not contain any known compromised data. Take care to track the training data used in various model lineages. Exercise caution before using a model without such assurances.

17.1.4. 2.4 Require certification

Data consumers should strongly consider requiring a formal certification from dataset and model providers, attesting that their systems are free from known compromised data before using third-party data and/or foundation models.

17.1.5. 2.10 Certification by Curators

Since the data supply chain begins with the curators, the certification process must start there as well. To the best of their ability, curators should be able to certify that, at the time of publication, the dataset contains no malicious or inaccurate material.

17.1.6. 2.11 Test & verify web-scale datasets

Be cautious when using web-scale datasets that are vulnerable to frontrunning poisoning. Check that the data hasn't been manipulated, and only use snapshots verified by a trusted party.

17.1.7. 2.12 (For web-scale data collectors) Randomize or lengthen snapshots

Collectors such as Wikipedia should defend against actors making malicious edits ahead of a planned snapshot by: 1. Randomizing the snapshot order. 2. Freezing edits to content long enough for edits to go through review before releasing the snapshot.

17.1.8. 2.13 Consensus approaches

Data consumers using web-crawled datasets should rely on consensus-based approaches, since notional determinations of which domains to trust are ad-hoc and insufficient. For example, an AI developer could choose to only trust an image-caption pair when it appears on many different websites to reduce susceptibility to poisoning techniques, since a malicious actor would have to poison a sufficiently large number of websites to be successful.

17.1.9. 2.14 Data curation

Ultimately, it is incumbent on organizations to ensure malicious or inaccurate material is not present in the data they use. If an organization does not have resources to conduct the necessary due diligence, then the use of web-crawled datasets is not recommended until some sort of trust infrastructure can be implemented

17.1.10. 3.1 Anomaly Detection

Incorporate anomaly detection algorithms during data preprocessing to identify and remove malicious or suspicious data points before training. These algorithms can recognize statistically deviant patterns in the data, making it possible to isolate and eliminate poisoned inputs

18. Atlantic Council

18.1. Securing data in the AI supply chain

Link: <https://www.atlanticcouncil.org/in-depth-research-reports/issue-brief/securing-data-in-the-ai-supply-chain/>

Link: <https://www.atlanticcouncil.org/in-depth-research-reports/issue-brief/securing-data-in-the-ai-supply-chain/>

18.1.1. Recommendation 1

Developers, users, maintainers, governors, and securers of AI technologies should map the data components of the AI supply chain to existing cybersecurity best practices—and use that mapping to identify where existing best practices fall short for AI-specific risks to the data components of the AI supply chain.

18.1.2. Recommendation 2

Developers, users, maintainers, governors, and securers of AI technologies should “Know Your Supplier,” using the supply chain-focused approach to mitigate both AI-specific and non-AI-specific risks to the data components of the AI supply chain.

18.1.3. Recommendation 3

Policymakers should widen their lens on AI data to encompass all data components of the AI supply chain. This includes assessing whether sufficient attention is given to the diversity of data use cases that need protection (e.g., not just training data for chatbots but for transportation safety or drug discovery) and whether they have mapped existing security best practices to non-AI-specific and AI-specific risks.

19. Microsoft

19.1. 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>

19.1.1. A4.3

If you plan to use existing data sets to train the system, assess the quantity and suitability of available data sets that will be needed by the system in relation to the data requirements defined in A4.1. Document this assessment in the Impact Assessment.

19.1.2. A4.5

Evaluate all data sets using the methods defined in requirement A4.4. Document the results of the evaluation.

20. Google

20.1. Secure AI Framework

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

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

20.1.1. Model and Data Integrity Management

Ensure that all data, models, and code used to produce AI models are verifiably integrity-protected during development and deployment.

20.1.2. Secure-by-Default ML Tooling

Use secure-by-default frameworks, libraries, software systems, and hardware components for AI development or deployment to protect confidentiality and integrity of AI assets and outputs.

21. MITRE

21.1. ATLAS Framework

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

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

21.1.1. AML.M0023 - AI Bill of Materials

An AI Bill of Materials (AI BOM) contains a full listing of artifacts and resources that were used in building the AI. The AI BOM can help mitigate supply chain risks and enable rapid response to reported vulnerabilities.

This can include maintaining dataset provenance, i.e. a detailed history of datasets used for AI applications. The history can include information about the dataset source as well as a complete record of any modifications.

21.2. SAFE-AI

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

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

21.2.1. Supply chain and life cycle infiltrations and unvetted changes to the model (especially open source)

AI-enabled systems often incorporate pre-trained models obtained from external sources. Failure to assure that these models are securely sourced from external suppliers will enable attacks that insert malicious code into the system that can compromise the AI system's security and integrity. It is also important to scrutinize any updates or changes to these models, since the impact of a revised model on system behavior may not be immediately obvious. This makes change management and configuration management critically important for models. Continual/continuous testing may be needed to detect unexpected changes over time. It may also be important to establish baselines and understanding of operational data and its drift.

21.2.2. Supply chain and life cycle infiltrations and unvetted changes of training and operational data

AI-enabled systems often rely on data obtained from external sources. Failure to assure that these data resources are securely sourced from external suppliers will enable attacks that insert malicious code into the system that can compromise the AI system's security and integrity. If the provenance of the data from external sources is not carefully documented (e.g., origins, transformations, dependencies, metadata, etc.), it may be challenging to recognize changes that render the data unsuitable for the current system. Mitigations include stringent vetting of AI model training, use of operational data, and thorough documentation of the data provenance.

21.2.3. Supply chain infiltrations/ unvetted changes of AI tools/platform s (especially open source)

AI-enabled systems are often built using tools and platforms obtained from external sources. If one of these resources has vulnerabilities because it is outdated, unpatched or compromised, it could provide a point of entry for attacks that poison data, perturb model inputs, or modify AI models to undermine their reliability, integrity and availability. Similar concerns arise if the external resources are not securely sourced from external suppliers. The vulnerabilities of the tools and platforms from external sources must be carefully understood and managed. Otherwise, it may be challenging to identify the underlying cause of any adverse outcomes in the AI system. Mitigations include meticulous attention to the preparation and maintenance of relevant risk assessment documents such as software bills of materials (SBOMs), AI system bills of materials (AIBOMs), data cards, and model cards.

21.2.4. Supply Chain infiltration/ unvetted changes of Environment components (especially open source)

Since AI is such a rapidly evolving technical area, AI-enabled systems tend to incorporate open-source components, or capabilities provided by external suppliers. Adversaries can sometimes gain initial access to a system by infiltrating and compromising targeted portions of the AI supply chain. This could include specialized hardware like GPUs, software stacks for AI code development, or pretrained AI models. Failure to assure that AI-related components are securely sourced from external suppliers will enable attacks that insert malicious code into the system that can compromise the AI system's security and integrity. Moreover, it is especially important to scrutinize all updates and patches needed for any third party or open-source capabilities that may be integrated into an AI component. This implies that relevant risk assessment documents must be carefully prepared, such as software bills of materials (SBOMs), AI system bills of materials (AIBOMs), data cards, and model cards.

21.2.5. AI bias

Biases in the data and models associated with an AI-enabled system can lead to inaccurate outcomes or discriminatory treatment of certain individuals or demographics, with a corresponding negative impact on the trustworthiness of the system. A key underlying issue is that the scale and complexity of many AI-enabled systems can result in high levels of statistical uncertainty and many potential sources of bias, making bias management a challenge. Unintended sources of bias like spurious correlations and unrepresentative data sources may be difficult to avoid. Malicious sources of bias caused by adversarial attacks on data and models may be challenging to detect. Mitigations include careful attention to the quality of data and its statistical properties, stringent access controls for data and models, and vigilant monitoring of system performance.

21.2.6. Insecure Plugin Design

AI software often extends its functionality by using plugins, extensions, or APIs to connect to other services or resources. Plugins may provide a variety of useful capabilities, such as integrations with other applications, access to public or private data sources, and the ability to execute code. If these plugins are not securely designed (e.g., plugins have insufficient access controls or inadequate input validation), adversaries may exploit their access to the AI software to compromise the plugins with attacks that have harmful consequences like data exfiltration, remote code execution, and privilege escalation. Developers must implement robust security measures for plugins, like strict parameterized inputs and secure access control guidelines, to mitigate this potential vulnerability.

22. NIST

22.1. 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>

22.1.1. GOVERN 6.1

Policies and procedures are in place that address AI risks associated with third-party entities, including risks of infringement of a third-party's intellectual property or other rights.

22.1.2. GOVERN 6.2

Contingency processes are in place to handle failures or incidents in third-party data or AI systems deemed to be high-risk.

22.1.3. MAP 4.1

Approaches for mapping AI technology and legal risks of its components – including the use of third-party data or software – are in place, followed, and documented, as are risks of infringement of a third party's intellectual property or other rights.

22.2. 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>

22.2.1. PO.3.1

Specify which tools or tool types must or should be included in each toolchain to mitigate identified risks, as well as how the toolchain components are to be integrated with each other.

Plan to develop and implement automated toolchains that secure AI model development and reduce human effort, especially at the scale often used by AI models.

22.2.2. PS.3.2

Collect, safeguard, maintain, and share provenance data for all components of each software release (e.g., in a software bill of materials [SBOM], through Supply-chain Levels for Software Artifacts [SLSA]).

Track the provenance of an AI model and its components and derivatives, including the training libraries, frameworks, and pipelines used to build the model.

Track AI models that were trained on sensitive data (e.g., payment card data, protected health information, other types of personally identifiable information), and determine if access to the models should be restricted to individuals who already have access to the sensitive data used for training.

22.2.3. PW.4.4

Verify that acquired commercial, open-source, and all other third-party software components comply with the requirements, as defined by the organization, throughout their life cycles.

Verify the integrity, provenance, and security of an existing AI model or any other acquired AI components — including training, testing, fine-tuning, and aligning datasets; reward models; adaptation layers; and configuration parameters — before using them.

Scan and thoroughly test acquired AI models and their components for vulnerabilities and malicious content before use.

22.3. 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>

22.3.1. GV.SC-03

Cybersecurity supply chain risk management is integrated into cybersecurity and enterprise risk management, risk assessment, and improvement processes

22.3.2. GV.SC-04

Suppliers are known and prioritized by criticality

22.3.3. GV.SC-05

Requirements to address cybersecurity risks in supply chains are established, prioritized, and integrated into contracts and other types of agreements with suppliers and other relevant third parties

22.3.4. GV.SC-06

Planning and due diligence are performed to reduce risks before entering into formal supplier or other third-party relationships

22.3.5. GV.SC-07

The risks posed by a supplier, their products and services, and other third parties are understood, recorded, prioritized, assessed, responded to, and monitored over the course of the relationship

22.3.6. GV.SC-08

Relevant suppliers and other third parties are included in incident planning, response, and recovery activities

22.3.7. GV.SC-09

Supply chain security practices are integrated into cybersecurity and enterprise risk management programs, and their performance is monitored throughout the technology product and service life cycle

22.3.8. GV.SC-10

Cybersecurity supply chain risk management plans include provisions for activities that occur after the conclusion of a partnership or service agreement

22.3.9. ID.AM-04

Inventories of services provided by suppliers are maintained

22.3.10. ID.RA-09

The authenticity and integrity of hardware and software are assessed prior to acquisition and use

22.3.11. ID.RA-10

Critical suppliers are assessed prior to acquisition

22.4. 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>

22.4.1. 2.3.1 Training data sanitization

These methods leverage the insight that poisoned samples are typically different than regular training samples that are not controlled by adversaries. As such, data sanitization techniques are designed to clean the training set and remove the poisoned samples before the ML training is performed. Cretu et al. [96] proposed the first sanitization procedure for unlabeled datasets that relies on majority voting of multiple models trained on subsets of the training set. They apply the method to anomaly detection on network packets. Nelson et al. [269] introduced the Region of Non-Interest (RONI) method, which examines each sample and excludes it from training if the accuracy of the model decreases when the sample is added. Subsequently proposed sanitization methods improved upon these early approaches by reducing their computational complexity and considering other applications. Paudice et al. [289] introduced a method for label cleaning that was specifically designed for label-flipping attacks. Steinhardt et al. [354] proposed the use of outlier detection methods for identifying poisoned samples. Clustering methods have also been used to detect poisoned samples [203, 363]. Other work has suggested that computing the variance of predictions made by an ensemble of multiple ML models is an effective data sanitization method for network intrusion detection [384]. Once sanitized, datasets may be protected by cybersecurity mechanisms for provenance and integrity attestation [267]

22.4.2. 2.3.3 Training data sanitization

Similar to poisoning availability attacks, training data sanitization can be applied to detecting backdoor poisoning attacks. For example, outlier detection in the latent feature space [157, 293, 378] has been effective for convolutional neural networks used for computer vision applications. Activation Clustering [76] clusters training data in representation space to isolate the backdoored samples in a separate cluster. Data sanitization achieves better results when the poisoning attack controls a relatively large fraction of training data but is not as effective against stealthy poisoning attacks. Overall, this leads to a trade-off between attack success and the detectability of malicious samples.

22.4.3. 2.3.4 Model Poisoning

A variety of Byzantine-resilient aggregation rules have been designed and evaluated to defend federated learning from model poisoning attacks. Most of them attempt to identify and exclude the malicious updates when performing the aggregation at the server [8, 43, 51, 149, 242–244, 359, 423]. However, motivated adversaries can bypass these defenses by adding constraints to the attack generation optimization problem [23, 123, 335]. Gradient clipping and differential privacy have the potential to mitigate model poisoning attacks to some extent [23, 271, 360], but they usually decrease accuracy and do not provide complete mitigation.

For specific model poisoning vulnerabilities, such as backdoor attacks, there are some techniques for model inspection and sanitization (see Sec. 2.3.3). However, mitigating supplychain attacks in which adversaries might control the source code of the training algorithm or the ML hyperparameters remains challenging. Program verification techniques used in other domains (e.g., cryptographic protocol verification [299]) might be adapted to this setting, but ML algorithms have intrinsic randomness and non-deterministic behavior, which enhances the difficulty of verification.

22.4.4. 3.3.3 Direct Prompting Attacks - Training data sanitization

Model training data can be sanitized to remove sensitive or toxic content and data that are largely or exclusively relevant for developing undesirable capabilities. Such sanitization may prevent harmful capabilities from being learned and reduce the potential harms from direct prompt injection, though they may harm generalization and harmful content detection abilities [224].