Messaging, Malware and Mobile Anti-Abuse Working Group

M³AAWG AI Model Lifecycle Security Best Common Practices

May 2025

The reference URL for this document:
http://www.m3aawg.org/AIModelLifecycleSecurityBCP

Terms, Definitions, Abbreviations and Acronyms

1. Scope **3**
1.1. Introduction and Overview 3
1.2. Purpose of Document 3
1.3. Terminology 3
2. Technical Guidance **4**
2.0.1. Training Data (TD) **4**
2.0.2. Model Training (MT) 6
2.0.3. Model Testing (MTEST) 8
2.0.4. Model Deployment (MD) **9**
Docker Security 9
Infrastructure Security 10
2.0.5. Data At Rest Protection (DRP) **10**
2.0.6. Data Exchange Encryption (DE) 11
2.0.7. Remote Management Interfaces (MI) 11
2.0.8. Logging (LOG) 11
2.0.9. Data Retention (DR) **12**
3. Conclusion **13**
4. Glossary of Terms **14**

1. Scope

1.1. Introduction and Overview

The AI lifecycle refers to the end-to-end process of developing, deploying and maintaining artificial intelligence (AI) systems. Current practices encompass various stages, including data collection, preprocessing, model development, training, evaluation, deployment, monitoring and retraining. Each stage is critical for ensuring the effectiveness, reliability and security of AI applications and services. Throughout the AI lifecycle, it is essential to adhere to best practices to mitigate risks, ensure transparency and maintain the integrity of the system. This includes implementing integrity and authenticity controls on training data sources, employing robust model development and deployment processes, and continuously monitoring and updating AI systems to adapt to changing environments and requirements. This document is not intended for legal advice; please consult your company's legal counsel.

1.2. Purpose of Document

This document specifies the best known common practices, as of publication, for evaluating the security of AI applications and services, whether they are purchased or developed in-house. It aims to ensure that all stages of the AI lifecycle, from data collection to deployment and monitoring, adhere to best practices to mitigate risks, ensure transparency and maintain system integrity. By implementing robust model development and deployment processes and continuously updating AI systems to adapt to changing environments, the guidelines aim to enhance the effectiveness, reliability and security of AI applications and services. This document is intended to offer specific best common practices with clear normative language for these practices aimed at the information technology sector practitioners. This document is intended to augment other current practices from relevant bodies, such as ISO/IEC, the National Institute of Standards and Technology (NIST), the European Telecommunications Standards Institute (ETSI), CEN/CENELEC, and the IEEE. It will be updated to reflect changing technology and aims to support implementers and practitioners, rather than stipulating requirements.

1.3. Terminology

Per RFC2119[1] and in line with ISO/IEC[2], this document uses the following terminology:

1. MUST This word, or the terms "REQUIRED" or "SHALL," mean that the definition is an absolute requirement of the specification.

2. MUST NOT This phrase, or the phrase "SHALL NOT", means that the definition is an absolute prohibition of the specification.

3. SHOULD This word, or the term "RECOMMENDED," means that there may be valid reasons in particular circumstances to ignore a particular item, but the full implications must be understood and carefully weighed before choosing a different course.

4. SHOULD NOT This phrase, or the phrase "NOT RECOMMENDED," means that there may be valid reasons in certain circumstances when the particular behavior is acceptable or even useful, but the full implications should be understood and the case carefully weighed before implementing any behavior described with this label.

2. Technical Guidance

Software systems that use AI[3] are the result of multi-step, often complex, processes that include the relevant technical steps necessary to acquire and prepare data, train and test models, deploy and operate AI systems, and properly decommission those systems when necessary. Throughout these steps, data need to be protected at rest and in use, the system must be managed and updated, and the appropriate auditability must be maintained by data retention and logging.

The deployer of an AI system must first define relevant security, privacy and other objectives, as well as appropriate policies and requirements for the AI system on a case-by-case basis, e.g., by using risk management approaches like threat modeling.

Creators, deployers and users of AI systems shall analyze their systems and models in the context of their use case, aiming to define relevant objectives (and their comparative importance), pertinent threats and risks, and the controls necessary to achieve those objectives within the given regulatory, technical and organizational context. Secure software development, patching, data labeling, change control, platform security, third-party management and other terms apply to AI systems as they would to any other dataset or code.

If the requirements for any model or AI system exceed the technical specifications outlined in this document, it will be necessary to implement appropriate managerial and procedural safeguards, establish abuse and

[1] https://www.rfc-editor.org/rfc/rfc2119
[2] https://www.iso.org/foreword-supplementary-information.html
[3] As the definitions and understandings of terms such as “AI system” and “model” are still being developed by relevant standards organizations, this document approaches the inclusion of AI in a wide-ranging manner, including the use of AI components in larger software systems that are not AI-focused.

incident handling structures, and apply various technical controls related to the development, hosting, and overall technology stack to ensure the proper operation of the AI system.

2.0.1. Training Data (TD)

Training data serves as the foundation upon which AI algorithms learn to make predictions and decisions. Therefore, any inaccuracies, biases or malicious alterations within the data can significantly compromise the performance and fairness of the resulting AI models. The following recommendations pertain to the training data. Where and how that training data was obtained is beyond the scope of this document.

As an overarching objective, organizations involved in developing AI solutions shall establish a functional data governance regime, giving the organization visibility into what data are being used and how, as well as ensuring that these data and associated processes are kept secure in terms of confidentiality and integrity. This data governance regime should also be able to trace what took place in the event of an investigation, and demonstrate the authenticity and integrity of the data, technical measures and management processes.

The integrity of training data through rigorous validation processes and implementation of digital signatures or cryptographic techniques ensures that organizations can establish a level of trust in the data's origin, authenticity and integrity. These measures not only safeguard against data manipulation and tampering but also contribute to transparency, accountability and ethical use of AI technologies, which can cycle back into higher level management, compliance and governance objectives and requirements.

Recommendation	Description	Comments
TD-001	Training data SHALL include an integrity check using SHA256 or higher.	Publishers of training data should include this hash on their website alongside the download.
TD-002	Training data SHOULD include a signature across each segment of data from a unique source.	
TD-003	Signatures SHALL chain up to a recognized PKI.	
TD-004	Training data sources SHOULD be transparent and listed.	
TD-005	The training process SHOULD be transparent.	
TD-006	Training data SHALL avoid including	Name, DOB, IP address, employee ID, any information

	personally identifiable information.	that is not fully anonymized and can identify and relate to an individual. De-identified data can still be considered PII (*Guidance on the Protection of Personal Identifiable Information* [4]).
TD-007	Training data SHOULD be labeled, for example: public, internal, confidential, sensitive, or highly restricted.	• NIST Special Publication 800-60 (Stine et al., 2008) • ISO/IEC 27001 (ISO/IEC, 2022) • ISO/IEC 5962:2021 Information technology SPDX (Linux Foundation, 2021) • NIST Special Publication 800-161 (Boyens et al., 2022)
TD-008	Training data classifications SHOULD be included in meta-data and signed along with the training data.	
TD-009	Training data SHOULD state if it contains known intellectual property.	
TD-010	Training data SHOULD include a list of possible biases that might affect the model inference.	
TD-011	Training data SHOULD list steps taken to remove personally identifiable data.	
TD-012	Fine-tuned or transfer learning models SHOULD attribute to the data of the foundation model.	

2.0.2. Model Training (MT)

In the context of model training, implementing robust mechanisms is imperative to ensure that the integrity of data and training processes is maintained.

[4] https://gdpr-info.eu/recitals/no-26/

As an overarching objective, organizations involved in developing AI solutions shall ensure that model training and testing maintain integrity throughout the training and testing process. They should be able to verify what data was used, when and how it was used, and verify that no accidental or unintended errors have been introduced.

For example, by checking signatures and hashes across each node involved in the training process, organizations can ensure the integrity and authenticity of both the training data and the model parameters exchanged between nodes. This involves employing cryptographic techniques to generate unique identifiers for datasets and model weights, which are then securely distributed across the training nodes. As the training progresses, each node independently verifies the signatures and hashes of the received data and parameters, thereby detecting any unauthorized alterations or tampering attempts.

Recommendation	Description	Comments
MT-001	Training data hash SHALL be verified using a SHA256 or higher hashing algorithm.	
MT-002	Training data signature SHOULD be verified at time of training, retraining or fine-tuning.	
MT-003	Training data signature expiration SHOULD be respected and checked at time of training.	
MT-004	Training data signature revocation SHOULD be respected and checked at time of training.	
MT–005	Trained model weights SHOULD inherit the classification of the most sensitive classification in the training data.	Model that is trained on confidential data inherits this classification from the training data.
MT-006	Model weights and hyper-parameters SHALL include an integrity check using SHA256 or higher.	
MT-007	Model weights and hyper-parameters SHOULD include a signature.	

MT-008	Model signatures SHOULD include metadata about the training data classification.	• NIST FIPS PUB 186-5: Digital Signature Standard (DSS) (NIST, 2023) • NIST FIPS PUB 180-4: Secure Hash Standard (SHS) (NIST, 2015)
MT-009	Model metadata SHOULD include the signatures and hashes of the training data.	

2.0.3. Model Testing (MTEST)

Effective model testing is a critical component of securing A.I. systems, as it enables the identification and mitigation of potential vulnerabilities that can compromise the integrity and reliability of models. Unlike traditional software development, where testing is often performed after coding, model testing is typically done during the development phase to catch issues early on.

Recommendation	**Description**	Comments
MTEST-001	Models SHOULD be red-team tested.	Red-team testing refers to a practice where a group of experts, known as the red team, intentionally tries to challenge and find vulnerabilities within an AI system. This is analogous to red-team testing in cybersecurity, where the goal is to attack the system from the inside in order to better understand its weaknesses and fortify its defenses. • NIST Special Publication 1270 (2022) Towards a Standard for Identifying and Managing Bias in Artificial Intelligence

		(Schwartz et al., 2022)
MTEST-002	Model testing SHOULD be done on models and systems that mimic production as closely as possible.	This includes, where possible, development, test, stage and production environments.
MTEST-003	The model SHOULD be scanned for malware (inserted into the model parameters, or weights) using modern tooling.	Models are subject to deserialization attacks where code can be injected into the model, and which will run upon the model being loaded.

2.0.4. Model Deployment (MD)

Model deployment is the stage where the model is loaded into memory for inference tasks. With different deployment options for AI models – such as self-hosted; SaaS or AIaaS; the use of hyperscalers and various types of AI systems that differ in their user and system interfaces, configuration approaches and use cases – deployment and hosting of AI systems is extremely specific.

Therefore, AI users, creators, deployers and relevant infrastructure providers must create appropriate security architectures and controls that meet their business, security and privacy objectives for the given systems and well as their importance and risk, use cases and technical deployment. Organizations must also ensure that appropriate process controls exist to verify secure deployment, management and oversight, maintenance (e.g. updates and patching) and deprecation of (parts of) the AI system and its underlying infrastructure.

Recommendation	**Description**	Comments
MD-001	Model weights and hyperparameters signatures SHALL be verified at each node that runs the model.	

In addition to verifying model signatures at each node, there are other security recommendations and common practices relating to secure deployment of AI models. This list is not exhaustive.

Docker Security

Models are often deployed in application containers that are small, isolated run-time environments for specific services and applications. There are several best practices for secure container deployment[5] [6] [7]:

Infrastructure Security

Infrastructure security for AI services, when focused specifically on hardware, involves securing the physical and computational assets that power AI algorithms and models. This includes the servers, GPUs, TPUs and storage devices essential for handling the intensive computational tasks AI requires. Guidelines and frameworks include NIST Special Publication 800-147: BIOS Protection Guideline[8], NIST Cybersecurity Framework[9].

For many users of AI systems, managing such requirements will rely on third-party management of external providers and ensuring that they implement necessary controls. Nevertheless, technical requirements such as ensuring encrypted transport and securing interfaces will remain relevant.

Relevant standards to consider for third-party risk management include, but are not limited to, ISO42001 (AI Governance), ISO27001/2, and possibly ISO27017/27018 (for cloud deployments).

2.0.5. Data At Rest Protection (DRP)

As training data are a key element of models and resulting AI systems, protecting these data is important. Data may be confidential and proprietary in nature, contain personal data or may be targeted by attackers to introduce random or non-random errors into future model iterations.

REQ#	Description	Status/Comment
DRP-001	Training data SHOULD be stored encrypted on disk.	
DRP-002	Model weights, if not in use, SHOULD be stored encrypted on disk.	

[5] *Boyens, J., Smith, A., Bartol, N., Winkler, K., Holbrook, A., & Fallon, M. (2022, May 5). NIST Special Publication (SP) 800-161 Rev. 1, Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations. NIST Computer Security Resource Center. Retrieved August 18, 2024, from https://csrc.nist.gov/publications/detail/sp/800-161/rev-1/final*

[6] The Docker Bench for Security. (2024). GitHub. Retrieved July 31, 2024, from https://github.com/docker/docker-bench-security Docker security. (2024). Docker Docs. Retrieved July 31, 2024, from https://docs.docker.com/engine/security/

[7] Docker security. (2024). Docker Docs. Retrieved July 31, 2024, from https://docs.docker.com/engine/security/)

[8] Cooper, D., Polk, W., Regenscheid, A., & Souppaya, M. (2011, April 29). SP 800-147, BIOS Protection Guidelines | CSRC. NIST Computer Security Resource Center. Retrieved August 18, 2024, from https://csrc.nist.gov/publications/detail/sp/800-147/final

[9] The NIST Cybersecurity Framework (CSF) 2.0, https://csrc.nist.gov/pubs/cswp/29/the-nist-cybersecurity-framework-csf-20/final

DRP-003	Input interaction history SHALL be stored encrypted on disk.	
DRP-004	Inference output SHOULD be stored encrypted on disk.	
DRP-005	Custom instructions, pre-prompts, and tone SHOULD be stored encrypted on disk.	
DRP-006	Documents for embedding SHOULD be stored encrypted.	
DRP-007	Vector and graph databases SHOULD be encrypted.	

It is good practice to maximize the time and data states for which encryption applies but there are logical and usability limitations in doing so.

Utilize hardware-based encryption for data at rest and in transit. Deploy Trusted Platform Modules (TPMs) for secure key storage and execution of cryptographic operations. Consider the use of Hardware Security Modules (HSMs) for robust cryptographic key management [10] [11] [12].

2.0.6. Data Exchange Encryption (DE)

Data exchange encryption is essential for ensuring the confidentiality and integrity of data as it moves across different systems, networks or services, particularly in environments where sensitive information is transmitted [13] [14] [15] [16].

2.0.7. Remote Management Interfaces (MI)

Remote interface management is a critical aspect of network and systems administration, involving the configuration, monitoring and maintenance of interfaces that allow remote access to devices and services. It

[10] NIST. (2019, March 22). *FIPS 140-3, Security Requirements for Cryptographic Modules | CSRC*. NIST Computer Security Resource Center. Retrieved August 20, 2024, from https://csrc.nist.gov/publications/detail/fips/140/3/final
[11] NIST. (2019, March 22). *FIPS 140-3, Security Requirements for Cryptographic Modules | CSRC*. NIST Computer Security Resource Center. Retrieved August 20, 2024, from https://csrc.nist.gov/publications/detail/fips/140/3/final
[12] Chen, L. (2022, August 1). *Recommendation for Key Derivation Using Pseudorandom Functions*. NIST Technical Series Publications. Retrieved August 20, 2024, from https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-108r1-upd1.pdf
[13] Chen, L. (2022, August 1). *Recommendation for Key Derivation Using Pseudorandom Functions*. NIST Technical Series Publications. Retrieved August 20, 2024, from https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-108r1-upd1.pdf
[14] Chen, L. (2022, August 1). *Recommendation for Key Derivation Using Pseudorandom Functions*. NIST Technical Series Publications. Retrieved August 20, 2024, from https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-108r1-upd1.pdf
[15] *RFC 5246 (TLS 1.2)*. (2008, August). tools.ietf. Retrieved August 20, 2024, from https://tools.ietf.org/html/rfc5246
[16] *RFC 5246 (TLS 1.2)*. (2008, August). tools.ietf. Retrieved August 20, 2024, from https://tools.ietf.org/html/rfc5246

ensures that reliable and secure access is provided to administrators and authorized users, while preventing unauthorized access and potential threats [17] [18] [19].

2.0.8. Logging (LOG)

Logging is extremely important to ensure higher-level objectives can be met, as they allow for quicker and better incident response, auditability and troubleshooting, among others. However, organizations should consider possible implications, for example, privacy implications when setting policies around logs, access and retention.

Recommendation	Description	Status/Comment
LOG-001	An AI service SHALL support logging of all administrative events.	
LOG-002	An AI service SHOULD support remote logging via syslog or equivalent standard protocol.	
LOG-003	An AI service SHOULD allow two or more remote logging endpoints to be defined.	
LOG-004	An AI service SHALL use a secure channel for sending all remote logging traffic.	
LOG-005	An AI service SHALL NOT log any sensitive information, including but not limited to passwords, encryption keys, API keys, usernames for unsuccessful login attempts, session keys or other forms of credentials.	Logging[20]
LOG-006	An AI service SHOULD have alerting capabilities for anomalous logged activity.	
LOG-007	An AI service SHOULD log when data is purged or deleted, added or amended.	

2.0.9. Data Retention (DR)

Data retention in terms of AI refers to the practice of storing and managing information collected and utilized during the development and deployment of artificial intelligence models. This involves retaining datasets, training data, model outputs, and related metadata for varying durations, depending on regulatory,

[17] https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r5.pdf
[18] https://www.cisecurity.org/cis-benchmarks/
[19] https://www.cisecurity.org/controls/
[20] Relevant guidance can be found in ISO24970 (AI System Logging) specifically and in ISO27002 8.15 and 8.16 at a more general level.

operational and strategic needs. Effective data retention practices ensure that data is available for future reference, audits, compliance with legal obligations and further model training or refinement while balancing the ethical considerations of privacy, security and data minimization to protect sensitive information and maintain trust with data subjects.

Recommendation	Description	Status/Comment
DR-001	Model weights SHOULD be clearly versioned.	Review data when the model was reviewed and verified.
DR-002	Outdated model weight versions SHOULD be erased.	Outdated model weights should be treated as outdated software that may contain vulnerabilities.
DR-003	Data retention policies of user interaction history SHALL be published.	
DR-004	The user interaction history duration that is stored SHOULD be configurable by the user.	
DR-005	Users SHALL be able to delete their interaction history.	
DR-006	Users SHALL be able to request deletion of their data in training data.	
DR-007	For trained models a differential privacy method SHOULD be implemented.	
DR-008	Organizations SHOULD store audit and event data required to use and develop an AI model.	
DR-009	Organizations SHALL define conditions and thresholds for incident response, troubleshooting and model re-evaluation.	

3. Conclusion

This document outlines best practices and technical guidance for evaluating the security of AI life cycle as it applies to applications and services. The guidelines cover various aspects, including training data integrity, model development and deployment, data at rest protection and data exchange encryption requirements. By adhering to these recommendations, organizations can better ensure the security, reliability, effectiveness and transparency of their AI systems. Furthermore, this document emphasizes the importance of continuous monitoring and updating AI systems to adapt to changing environments and requirements. By following these guidelines, organizations can minimize risks and maintain the integrity of their AI applications and services.

As noted throughout the document, the best practices provided here may be necessary but not sufficient to secure real-world AI systems end to end. To achieve end-to-end security, organizations should take into account higher level documents such as the NIST Cyber Security Framework 2.0 (CSF) that thoroughly cover organizational, procedural and general technical requirements and controls, and provide guidelines on how to manage systems, relevant threats and risks and interfaces in context.

4. Glossary of Terms

Personally Identifiable Information (PII)	Any data that relates to an identifiable individual or natural person that can be identified directly or indirectly.[21]
Artificial Intelligence (AI)	Systems that learn patterns in data to automatically improve performance on a task without being explicitly programmed.
Software as a Service (SaaS)	A delivery model in which software applications are hosted, managed and delivered over the internet, allowing users to access and use them on demand, without the need for local installation or maintenance.
Artificial Intelligence as a Service (AIaaS)	A delivery model that provides artificial intelligence (AI) capabilities, algorithms and predictive models to users on demand, allowing them to incorporate AI into their applications and systems without requiring significant expertise or infrastructure.

[21] Specific definitions of PII vary by country and applicable regulation. The definition used here is in accord with GDPR's definition, which has been adopted by various other laws and regulations.

Public Key Infrastructure (PKI)	A system that enables secure communication over the internet by using pairs of cryptographic keys, one public and one private, to verify the identity of entities and ensure the confidentiality, integrity and authenticity of data.
Differential Privacy (DP)	A mathematically rigorous framework for releasing statistical information about datasets while protecting the privacy of individual data subjects.
Secure Software Development	Secure software development is the practice of designing, building and maintaining software with a focus on protecting it from security threats and vulnerabilities throughout its lifecycle.
Patching	Patching is the process of updating software to fix security vulnerabilities, bugs and other issues to maintain its functionality and protect against potential threats.
Data Labeling	Data labeling is the process of annotating or tagging data with relevant labels or categories to help train machine learning models and improve their accuracy.
Change Control	Change control is the process of managing and tracking changes to a system, software or project to ensure that modifications are implemented smoothly, documented properly and do not introduce new risks or issues.
Platform Security	Platform security refers to the measures and practices implemented to protect the underlying hardware, software and network infrastructure of a computing platform from unauthorized access, attacks and other security threats.
Third-Party Management	Third-party management is the process of overseeing and controlling the risks associated with external vendors, contractors or service providers to ensure they meet security, compliance and performance standards.
Pseudonymization	The processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the data are not attributed to an identified or

	identifiable natural person.
AI Node	An AI node is the logical part of an AI system from which a model runs.

As with all documents that we publish, please check the M[3]AAWG website (www.m3aawg.org) for updates.

M3AAWG-151