





AI Data Security

Best Practices for Securing Data Used to Train & Operate AI Systems

Executive summary

This Cybersecurity Information Sheet (CSI) provides essential guidance on securing data used in artificial intelligence (AI) and machine learning (ML) systems. It also highlights the importance of data security in ensuring the accuracy and integrity of AI outcomes and outlines potential risks arising from data integrity issues in various stages of AI development and deployment.

This CSI provides a brief overview of the AI system lifecycle and general best practices to secure data used during the development, testing, and operation of AI-based systems. These best practices include the incorporation of techniques such as data encryption, digital signatures, data provenance tracking, secure storage, and trust infrastructure. This CSI also provides an in-depth examination of three significant areas of data security risks in AI systems: data supply chain, maliciously modified ("poisoned") data, and data drift. Each section provides a detailed description of the risks and the corresponding best practices to mitigate those risks.

This guidance is intended primarily for organizations using AI systems in their operations, with a focus on protecting sensitive, proprietary, or mission critical data. The principles outlined in this information sheet provide a robust foundation for securing AI data and ensuring the reliability and accuracy of AI-driven outcomes.

This document was authored by the National Security Agency's Artificial Intelligence Security Center (AISC), the Cybersecurity and Infrastructure Security Agency (CISA), the Federal Bureau of Investigation (FBI), the Australian Signals Directorate's Australian Cyber Security Centre (ASD's ACSC), the New Zealand's Government Communications

This information is marked TLP:CLEAR. TLP:CLEAR information may be distributed without restriction. For more information on the Traffic Light Protocol, see cisa.gov/tlp/.

Security Bureau's National Cyber Security Centre (NCSC-NZ), and the United Kingdom's National Cyber Security Centre (NCSC-UK).

The goals of this guidance are to:

1. Raise awareness of the potential risks related to data security in the development, testing, and deployment of AI systems;
2. Provide guidance and best practices for securing AI data across various stages of the AI lifecycle, with an in-depth description of the three aforementioned significant areas of data security risks; and
3. Establish a strong foundation for data security in AI systems by promoting the adoption of robust data security measures and encouraging proactive risk mitigation strategies.

Introduction

The data resources used during the development, testing, and operation of an **AI**[1] system are a critical component of the AI supply chain; therefore, the data resources must be protected and secured. In its Data Management Lexicon, [1] the Intelligence Community (IC) defines **Data Security** as "The ability to protect data resources from unauthorized discovery, access, use, modification, and/or destruction…. Data Security is a component of Data Protection."

Data security is paramount in the development and deployment of AI systems. Therefore, it is a key component of strategies developed to safeguard and manage the overall security of AI-based systems. Successful data management strategies must ensure that the data has not been tampered with at any point throughout the entire AI system lifecycle; is free from malicious, unwanted, and unauthorized content; and does not have unintentional duplicative or anomalous information. Note that AI data security depends on robust, fundamental cybersecurity protection for all datasets used in designing, developing, deploying, operating, and maintaining AI systems and the ML models that enable them.

[1] In this document, **Artificial Intelligence** (AI) has the meaning set forth in 15 U.S.C. 9401(3):
"… a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems use machine- and human-based inputs to:
(A) Perceive real and virtual environments;
(B) Take these perceptions and turn them into models through analysis in an automated manner; and
(C) Use model inference to formulate options for information or action."

Audience and scope

This CSI outlines potential risks in AI systems stemming from data security issues that arise during different phases of an AI deployment, and it introduces recommended protocols to mitigate these risks. This guidance builds upon the NSA's joint guidance on Deploying AI Systems Securely [2] and delves deeper into securing the data used to train and operate AI-based systems. This guidance is primarily developed for organizations that use AI systems in their day-to-day operations, including the Defense Industrial Base (DIB), National Security System (NSS) owners, Federal Civilian Executive Branch (FCEB) agencies, and critical infrastructure owners and operators. Implementing these mitigations can help secure AI-enabled systems and protect proprietary, sensitive, and/or mission critical data.

Securing data throughout the AI system lifecycle

Data security is a critical enabler that spans all phases of the AI system lifecycle. ML models learn their decision logic from data, so an attacker who can manipulate the data can also manipulate the logic of an AI-based system. In the AI Risk Management Framework (RMF) [3], the National Institute of Standards and Technology (NIST) defines six major stages in the lifecycle of AI systems, starting from Plan & Design and progressing all the way to Operate & Monitor. The following table highlights relevant data security factors for each stage of the AI lifecycle:

Table 1: The AI System Lifecycle with key dimensions, necessary ongoing assessments, focus areas for data security, and particular data security risks covered in this CSI. [3]

AI Lifecycle Stage	Key Dimensions	Test, Evaluation, Verification, & Validation (TEVV)	Potential Focus Areas for Data Security	Particular Data Security Risks Covered in this CSI
1) Plan & Design	Application Context	Audit & Impact Assessment	Incorporating data security measures from inception, designing robust security protocols, threat modeling, and including privacy by design	Data supply chain

AI Lifecycle Stage	Key Dimensions	Test, Evaluation, Verification, & Validation (TEVV)	Potential Focus Areas for Data Security	Particular Data Security Risks Covered in this CSI
2) Collect & Process Data	Data & Input	Internal & External Validation	Ensuring data integrity, authenticity, encryption, access controls, data minimization, anonymization, and secure data transfer	Data supply chain, maliciously modified data
3) Build & Use Model	AI Model	Model Testing	Protecting data from tampering, ensuring data quality and privacy (including differential privacy and secure multi-party computation when appropriate and possible), securing model training, and operating environments	Data supply chain, maliciously modified data
4) Verify & Validate	AI Model	Model Testing	Performing comprehensive security testing, identifying and mitigating risks, validating data integrity, adversarial testing, and formal verification when appropriate and possible	Data supply chain, maliciously modified data
5) Deploy & Use	Task & Output	Integration, Compliance Testing, Validation	Implementing strict access controls, zero-trust infrastructure, secure data transmission and storage, secure API endpoints, and monitoring for anomalous behavior	Data supply chain, maliciously modified data, data drift

AI Lifecycle Stage	Key Dimensions	Test, Evaluation, Verification, & Validation (TEVV)	Potential Focus Areas for Data Security	Particular Data Security Risks Covered in this CSI
6) Operate & Monitor	Application Context	Audit & Impact Assessment	Conducting continuous risk assessments, monitoring for data breaches, deleting data securely, complying with regulations, incident response planning, and regular security auditing	Data supply chain, maliciously modified data, data drift

Throughout the AI system lifecycle, securing data is paramount to maintaining information integrity and system reliability. Starting with the initial ***Plan & Design phase***, carefully plan data protection measures to provide proactive mitigations of potential risks. In the ***Collect & Process Data phase***, data must be carefully analyzed, labeled, sanitized, and protected from breaches and tampering. Securing data in the ***Build & Use Model phase*** helps ensure models are trained on reliably sourced, accurate, and representative information. In the ***Verify & Validate phase***, comprehensive and thorough testing of AI models, derived from training data, can identify security flaws and enable their mitigation.

Note that ***Verification & Validation*** is necessary each time new data or user feedback is introduced into the model; therefore, that data also needs to be handled with the same security standards as AI training data. Implementing strict access controls protects data from unauthorized access, especially in the ***Deploy & Use phase***. Lastly, continuous data risk assessments in the ***Operate & Monitor phase*** are necessary to adapt to evolving threats. Neglecting these practices can lead to data corruption, compromised models, data leaks, and non-compliance, emphasizing the critical importance of robust data security at every phase.

Best practices to secure data for AI-based systems

The following list contains recommended practical steps that system owners can take to better protect the data used to build and operate their AI-based systems, whether running on premises or in the cloud. For more details on general cybersecurity best

practices, see also NIST SP 800-53, “Security and Privacy Controls for Information Systems and Organizations.” [33]

1. Source reliable data and track data provenance

Verify data sources use trusted, reliable, and accurate data for training and operating AI systems. To the extent possible, only use data from authoritative sources. Implement provenance tracking to enable the tracing of data origins, and log the path that data follows through an AI system. [7] [8] [9] Incorporate a secure provenance database that is cryptographically signed and maintains an immutable, append-only ledger of data changes. This facilitates data provenance tracking, helps identify sources of maliciously modified data, and helps ensure that no single entity can undetectably manipulate the data.

2. Verify and maintain data integrity during storage and transport

Maintaining data integrity[2] is an essential component to preserve the accuracy, reliability, and trustworthiness of AI data. [4] Use checksums and cryptographic hashes to verify that data has not been altered or tampered with during storage or transmission. Generating such unique codes for AI datasets enables the detection of unauthorized changes or corruption, safeguarding the information’s authenticity.

3. Employ digital signatures to authenticate trusted data revisions

Digital signatures help ensure data integrity and prevent tampering by third parties. Adopt quantum-resistant digital signature standards [5] [6] to authenticate and verify datasets used during AI model training, fine tuning, alignment, reinforcement learning from human feedback (RLHF), and/or other post-training processes that affect model parameters. Original versions of the data should be cryptographically signed, and any subsequent data revisions should be signed by the person who made the change. Organizations are encouraged to use trusted certificate authorities to verify this process.

4. Leverage trusted infrastructure

Use a trusted computing environment that leverages Zero Trust architecture. [10] Provide secure enclaves for data processing and keep sensitive information protected and unaltered during computations. This approach fosters a secure foundation for data privacy and security in AI data workflows by isolating sensitive operations and mitigating

[2] **Data integrity** is defined by the IC Data Management Lexicon [1] as “The degree to which data can be trusted due to its provenance, pedigree, lineage and conformance with all business rules regarding its relationship with other data. In the context of data movement, this is the degree to which data has verifiably not been changed unexpectedly by a person or NPE.”

risks of tampering. Trusted computing infrastructure supports the integrity of data processes, reduces risks associated with unverified or altered data, and ultimately creates a more robust and transparent AI ecosystem. Trusted environments are essential for AI applications where data accuracy directly impacts their decision-making processes.

5. Classify data and use access controls

Categorize data using a classification system based on sensitivity and required protection measures. [11] This process enables organizations to apply appropriate security controls to different data types. Classifying data enables the enforcement of robust protection measures like stringent encryption and access controls. [33] In general, the output of AI systems should be classified at the same level as the input data (rather than creating a separate set of guardrails).

6. Encrypt data

Adopt advanced encryption protocols proportional to the organizational data protection level. This includes securing data at rest, in transit, and during processing. AES-256 encryption is the de facto industry standard and is considered resistant to quantum computing threats. [12] [13][13] Use protocols, such as TLS with AES-256 or post-quantum encryption, for data in transit. Refer to NIST SP 800-52r2, “Guidelines for the Selection, Configuration, and Use of Transport Layer Security (TLS) Implementations” [14] for more details.

7. Store data securely

Store data in certified storage devices that enforce NIST FIPS 140-3 [15] compliance, ensuring that the cryptographic modules used to encrypt the data provide high-level security against advanced intrusion attempts. Note that Security Level 3 (defined in NIST FIPS 140-2 [16]) provides robust data protection; however, evaluate and determine the appropriate level of security based on organizational needs and risk assessments.

8. Leverage privacy-preserving techniques

There are several privacy-preserving techniques [17] that can be leveraged for increased data security. Note that there may be practical limitations to their implementation due to computational cost.

- **Data depersonalization techniques** (e.g., data masking [18]) involve replacing sensitive data with inauthentic but realistic information that maintains the

distributions of values throughout the dataset. This enables AI systems to utilize datasets without exposing sensitive information, reducing the impact of data breaches and supporting secure data sharing and collaboration. When possible, use data masking to facilitate AI model training and development without compromising sensitive information (e.g., personally identifiable information [PII]).

- **Differential privacy** is a framework that provides a mathematical guarantee quantifying the level of privacy of a dataset or query. It requires a pre-specified privacy budget for the level of noise added to the data, but there are tradeoffs between protecting the training data from membership inference techniques and target task accuracy. Refer to [17] for further details.
- **Decentralized learning techniques** (e.g., federated learning [19]) permit AI system training over multiple local datasets with limited sharing of data among local instances. An aggregator model incorporates the results of the distributed models, limiting access on the local instance to the larger training dataset. Secure multi-party computation is recommended for training and inferencing processes.

9. Delete data securely

Prior to repurposing or decommissioning any functional drives used for AI data storage and processing, erase them using a secure deletion method such as cryptographic erase, block erase, or data overwrite. Refer to NIST SP 800-88, “Guidelines for Media Sanitization,” [20] for guidance on appropriate deletion methods.

10. Conduct ongoing data security risk assessments

Conduct ongoing risk assessments using industry-standard frameworks, such as the NIST SP 800-3r2, Risk Management Framework (RMF) [4] [21], and the NIST AI 100-1, Artificial Intelligence RMF [3]. These assessments should evaluate the AI data security landscape, identify risks, and prioritize actions to minimize security incidents. Continuously improve data security measures to keep pace with evolving threats and vulnerabilities, learn from security incidents, stay up to date with emerging technologies, and maintain a robust security posture.

Data supply chain – risks and mitigations

Relevant AI Lifecycle stages: 1) Plan & Design; 2) Collect & Process Data; 3) Build & Use Model; 4) Verify & Validate; 5) Deploy & Use; 6) Operate & Monitor

Developing and deploying secure and reliable AI systems requires understanding potential risks and methods of introducing inaccurate or maliciously modified (a.k.a. “poisoned”) data into the system. In short, the security of AI systems depends on

thorough verification of training data and proactive measures to detect and prevent the introduction of inaccurate material.

Threats can stem from large-scale data collected and curated by third parties, as well as from data that is not sufficiently protected after ingestion. Data collected and/or curated by a third party may contain inaccurate information, either unintentionally or through malicious intent. Inaccurate material can compromise not only models trained using that data, but also any additional models that rely on compromised models as a foundation.

It is crucial, therefore, to verify the integrity of the training data used when building an AI system. Organizations that utilize third-party data must take appropriate measures to ensure that: 1) the data is not compromised upon ingestion; and 2) the data cannot be compromised after it has been incorporated into the AI system. As such, both data curators and data consumers should follow the best practices for digital signatures, data integrity, and data provenance that are described in detail above.

General risks for data consumers[3]

The use of web-scale databases includes all of the risks outlined earlier, and one cannot simply assume that these datasets are clean, accurate, and free of malicious content. Third-party models trained on web-scraped data used to train a model for downstream tasks could also affect the model's learning process and result in behavior that was unintended by the AI system designer.

From the moment data is ingested for use with AI systems, the data acquirer must secure it against insider threats and malicious network activity to prevent unauthorized modification.

Mitigation strategies:

- **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.
- **Content credentials**: Use content credentials to track the provenance of media and other data. Content credentials are "metadata that are secured cryptographically and

[3] The term **data consumers** is defined as technical personnel (*e.g.* data scientists, engineers) who make use of data that they themselves did not produce or annotate to build and/or operate AI systems.

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]

- **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.
- **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.
- **Secure storage:** After ingest, data needs to be stored in a database that adheres to the best practices for digital signatures, data integrity, and data provenance that are described in detail above. Note that an append-only cryptographically signed database should be used where feasible, but there may be a need to delete older material that is no longer relevant. Each time a data element is updated (e.g., resized, cropped, flipped, etc.) for augmentation purposes in a non-temporary fashion, then the updated data should be stored as a new entry with documented changes. The database's certificate should be verified at the time the database is accessed for a training run. If the database does not pass the certificate check, abort the training and conduct a comprehensive database audit to discover any data modifications.

2023 investigations by various industry professionals explored low-resource methods for introducing malicious or inaccurate material into web-scale datasets, and potential strategies to mitigate this risk. [29] These vulnerabilities depend on the fact that curators or collectors do not have control over the data, as seen in cases of datasets curated by third parties (e.g., LAION) or datasets that are continually updated and released (e.g., Wikipedia).

Risk: Curated web-scale datasets

Curated AI datasets (e.g., LAION-2B or COYO-700M) are vulnerable to a type of technique known as **split-view poisoning**. This risk arises because these datasets often contain data hosted on domains that may have expired or are no longer actively maintained by their original owners. In such cases, anyone who purchases these expired domains gains control over the content hosted on them. This situation creates an opportunity for malicious actors to modify or replace the data that the curated list points to, potentially introducing inaccurate or misleading information into the dataset. In many instances, it is possible to purchase enough control of a dataset to conduct effective poisoning for roughly $1,000 USD. In some cases, effective techniques can cost as little as $60 USD (e.g., COYO-700M), making this a viable threat from low-resource threat actors.

Mitigation strategies:

- **Raw data hashes:** Data curators should attach a cryptographic hash to all raw data referenced in the dataset. This will enable follow-on data consumers to verify that the data has not changed since it was added to the list.
- **Hash verification:** Data consumers should incorporate a hash check at time of download in order to detect any changes made to it, and the downloader should discard any data that does not pass the hash check.
- **Periodic checks:** Curators should periodically scrape the data themselves to verify that the data has not been modified. If any changes are detected, the curator should take appropriate steps to ensure the data's integrity.
- **Verifying data:** Curators should verify that any changed data is clean and free from inaccurate or malicious material. If the content of the data has been altered in any way, the curator should either remove it from their list or flag it for further review.
- **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.

Risk: Collected web-scale datasets

Collected web-scale datasets (e.g., Wikipedia) are vulnerable to **frontrunning poisoning techniques**. Frontrunning poisoning occurs when an actor injects malicious examples in a short time window before websites with crowd-sourced content collect a

snapshot of their data. Wikipedia in particular conducts twice-monthly snapshots of their data and publishes these snapshots for people to download. Since the snapshots happen at known times, it is possible for malicious actors to edit pages close enough to the snapshot time so that malicious edits will be captured and published before they can be discovered and corrected. Industry analysis demonstrated potential malicious actors would be able to successfully poison as much as 6.5% of Wikipedia. [29]

Mitigation strategies:

- **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.
- **(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.

    These mitigations focus on increasing the amount of time a malicious actor must maintain control of the data for it to be included in the published snapshot. Any reasonable methods that increase the time a malicious actor must control the data are also recommended.

    Note that these mitigations are limited since they rely on trusted curators who can detect malicious edits. It is more difficult to defend against subtle edits (e.g., attempts to insert hidden watermarks) that appear valid to human reviewers but impact machine understanding.

Risk: Web-crawled datasets

Web-crawled datasets present a unique intersection of the risks discussed above. Since web-crawled datasets are substantially less curated than other web-scale datasets, they bring increased risk. There are no trusted curators to detect malicious edits. There are no original curated views to which cryptographic hashes can be attached. The unfortunate reality is that "updates to a web page have no realistic bound on the delta between versions which might act as a signal for attaching trust." [29]

Mitigation strategies:

- **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.
- **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.

Final note on web-scale datasets and data poisoning

Both split-view and frontrunning poisoning are reasonably straightforward for a malicious actor to execute, since they do not require particularly sophisticated methodology. These poisoning techniques should be considered viable threats by anyone looking to incorporate web-scale data into their AI systems. The danger here comes not only from directly using compromised data, but also from using models which may themselves have been trained on compromised data.

Ultimately, data poisoning must be addressed from a supply chain perspective by those who train and fine-tune AI models. Proper supply chain integrity and security management (i.e., selecting reliable model providers and verifying the legitimacy of the models used) can reduce the risk of data poisoning and system compromise. The most reliable providers are those who assure that they do everything possible to prevent the influence and distribution of poisoned data and models. [34]

Every effort must be made by those building foundation models to filter out malicious and inaccurate data. Foundation models are evolving rapidly, and filtering out inaccurate, unauthorized, and malicious training data is an active area of research, particularly at web-scale. As such, is currently impractical to prescribe precise methods for doing so; it is a best-effort endeavor. Ideally, data curators and foundation model providers should be able to attest to their filtering methods and provide evidence (e.g. test results) of their effectiveness. Likewise, if possible, downstream model consumers should include a review of the security claims as part of their security processes before accepting a foundation model for use.

Maliciously modified data – risks and mitigations

Relevant AI Lifecycle stages: 2) Collect & Process Data; 3) Build & Use Model; 4) Verify & Validate; 5) Deploy & Use; 6) Operate & Monitor

Maliciously modified data presents a significant threat to the accuracy and integrity of AI systems. Deliberate manipulation of data can result in inaccurate outcomes, poor decisions, and compromised security. Note that there are also risks associated with unintentional data errors and duplications that can affect the security and performance of AI systems. Challenges like adversarial machine learning threats, statistical bias, and inaccurate information can impact the overall security of AI-driven outcomes.

Risk: Adversarial Machine Learning threats

Adversarial Machine Learning (AML) threats involve intentional, malicious attempts to deceive, manipulate, or disrupt AI systems. [7] [17] [22] Malicious actors employ **data poisoning** to corrupt the learning process, compromising the integrity of training datasets and leading to unreliable or malicious model behavior. Additionally, malicious actors may introduce **adversarial examples** into datasets that, while subtle, can evade correct classification, thereby undermining the model's performance. Furthermore, **sensitive information** in training datasets can be indirectly extracted through techniques like model inversion[4], posing significant data security risks.

Mitigation Strategies:

- **Anomaly detection**: Incorporate anomaly detection algorithms during data pre-processing 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.
- **Data sanitization**: Sanitize the training data by applying techniques like data filtering, sampling, and normalization. This helps reduce the impact of outliers, noisy data, and other potentially poisoned inputs, ensuring that models learn from high-quality, representative datasets. Perform sanitization on a regular basis, especially prior to each and every training, fine-tuning, or any other process that adjusts model parameters.

[4] **Model inversion** refers to the process by which an attacker analyzes the output patterns of an AI system to reverse-engineer and uncover details about the training dataset, such as individual data points or patterns. This process can potentially expose confidential or proprietary information from the data that was used to train the AI models.

- **Secure training pipelines**: Secure data collection, pre-processing, and training pipelines to prevent malicious actors from tampering with datasets or model parameters.
- **Ensemble methods / collaborative learning**: Implement collaborative learning frameworks that combine an ensemble of multiple, distinct AI models to reach a consensus on output predictions. This approach can help counteract the impact of data poisoning, since malicious inputs may only affect a subset of the collaborative models, allowing the majority to maintain accuracy and reliability.
- **Data anonymization**: Implement anonymization techniques to protect sensitive data attributes, keeping them confidential while allowing AI models to learn patterns and generate accurate predictions.

Risk: Bad data statements

Bad data statements[5] [7] [23], such as missing metadata, can significantly influence AI data security by introducing data integrity issues that can lead to faulty model performance. Error-free metadata provides valuable contextual information about the data, including its structure, purpose, and collection methods. When metadata is missing, it becomes difficult to interpret data accurately and draw meaningful conclusions. This situation can result in incomplete or inaccurate data representation, compromising AI system performance and reliability. If metadata is modified by a malicious actor, then the security of the AI system is also at risk.

Mitigation strategies:

- **Metadata management**: Implement strong data governance practices to help ensure metadata is well-documented, complete, accurate, and secured.
- **Metadata validation**: Establish data validation processes to check the completeness and consistency of metadata before data is used for AI training.
- **Data enrichment**: Use available resources, such as reference data and trusted third-party data, to supplement missing metadata and improve the overall quality of the training data.

[5] "A **data statement** is a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software." [23]

Risk: Statistical bias[6]

Robust data security and collection practices are key to mitigating statistical bias. Executive Order (EO) 14179 mandates that U.S. government entities "develop AI systems that are free from ideological bias or engineered social agendas." [25] Note that "an AI system is said to be biased when it exhibits systematically inaccurate behavior." [26] Statistical bias in AI systems can arise from artifacts present in training data that can lead to artificially slanted or inaccurate outcomes. Sampling biases or biases in data collection can affect the overall outcomes and performance of AI. Left unaddressed, statistical bias can degrade the accuracy and effectiveness of AI systems.

Mitigation strategies:

- **Regular training data audits**: Regularly audit training data to detect, assess, and address potential issues that can result in systematically inaccurate AI systems.
- **Representative training data**: Ensure that training data is representative of the totality of the information relevant to any given topic to reduce the risk of statistical bias. Also ensure that AI data is properly divided into training, development, and evaluation sets without overlap to properly measure statistical bias and other measures of performance.
- **Edge cases**: Identify and mitigate edge cases that can cause models to malfunction.
- **Test and correct for statistical bias**: Create a repository with instances of observed model output bias. Leverage that information to improve training data audits and with reinforcement learning to "undo" some of the measured bias.

Risk: Data poisoning via inaccurate information

One form of data poisoning (sometimes referred to as "disinformation" [27]) involves the intentional insertion of inaccurate or misleading information in AI training datasets, which can negatively impact AI system performance, outcomes, and decision-making processes.

Mitigation strategies:

- **Remove inaccurate information from training data**: Identify and remove inaccurate or misleading information from AI datasets to the extent feasible.

[6] "In technical systems, **bias** is most commonly understood and treated as a statistical phenomenon. Bias is an effect that deprives a statistical result of representativeness by systematically distorting it, as distinct from random error, which may distort on any one occasion but balances out on the average." [26] [32]

- **Data provenance and verification**: Implement provenance verification mechanisms during data collection to help ensure that only accurate and reliable data is used. This process can include methods such as cross-verification, fact-checking, source analysis, data provenance tracking, and content credentials.
- **Add more training data**: Increasing the amount of non-malicious data makes training more robust against poisoned examples—provided that these poisoned examples are small in number. One way to do this is through data augmentation—the creation of artificial training set samples that are small variations of existing samples. The goal is to "outnumber" the poisoned samples so the model "forgets" them. Note that this mitigation can only be applied during training, and therefore does not apply to an already trained model. [28]
- **Data quality control**: Perform quality control on data including detecting poisoned samples through integrity checks, statistical deviation, or pattern recognition. Proactively implement data quality controls during the training phase to prevent issues before they arise in production.

Risk: Data duplications

Unintended duplicate data elements [7] in training datasets can skew model performance and cause overfitting, reducing the AI model's ability to generalize across a variety of real-world applications. Duplicates are not always exact; near-duplicates may contain minor differences like formatting, abbreviations, or errors, which makes detecting them more complex. Duplicate data often leads to inaccurate predictions, making the AI system less effective in real-world applications.

Mitigation strategies:

- **Data deduplication**: Implement deduplication techniques (such as fuzzy matching, hashing, clustering, etc.) to carefully identify and handle duplicates and near-duplicates in the data.

Data drift – risks and mitigations

Relevant AI Lifecycle stages: 5) Deploy & Use; 6) Operate & Monitor

Data drift, or distribution shift, refers to changes in the underlying statistical properties of the input data to an operational AI system. Over time, the input data can become significantly different from the data originally used to train the model. [7] [8] Degradation caused by data drift is a natural and expected occurrence, and AI system developers

and operators need to regularly update models to maintain accuracy and performance. Data drift ordinarily begins as small, seemingly insignificant degradations in model performance. Left unchecked, the degradation caused by data drift can snowball into substantial reductions in AI system accuracy and integrity that become increasingly difficult to correct.

It is crucial to distinguish between data drift and data poisoning attacks designed to affect an AI model. Continuous monitoring of system accuracy and performance provides important indicators based on the nature of the changes observed. If the changes are slow and gradual over time, it is more likely that the model is experiencing data drift. If the changes are abrupt and dramatic in one or more dimensions, it is more likely that an actor is trying to compromise the model. Cyber compromises often aim to manipulate the model's performance quickly and significantly, leading to abrupt changes in the input data or model outputs.

AI system operators and developers should employ a wide range of techniques for detecting and mitigating data drift, including data preprocessing, increasing dataset coverage of real-world scenarios, and adopting robust training and adaptation strategies. [30] Packages that automate dataset loading assist AI system developers in creating application-specific detection and mitigation techniques for data drift.

There are many potential causes of data drift, including:

1. A change in the upstream data pipeline not represented in the model training data (e.g., the units of a particular data element change from miles to kilometers)
2. The introduction of completely new data elements that the model had not previously seen (e.g., a new type of malware not recognized in the ML layer of an anti-virus product)
3. A change in the context of how inputs and outputs are related (e.g., a change in organizational structure due to a merger or acquisition could lead to new data access patterns that might be misinterpreted as security threats by an AI system)

The data associated with a given AI model should be regularly checked for any updates to help ensure the model still predicts as expected. [7] [8] [9] The interval for this update and check will depend on the particular AI system and application. For example, in high-stakes applications such as healthcare, early detection and mitigation of data drift are critical prior to patient impact. Thus, continuous monitoring of model performance with additional direct analysis of the input data is important in such applications. [30]

Mitigation strategies:

- **Data management:** Employ a data management strategy in keeping with the best practices in this CSI to help ensure that it is easy to add and track new data elements for model training and adaptation. This management strategy enables identification of data elements causing drift for appropriate mitigation or action.
- **Data-quality testing:** AI system developers should use data-quality assessment tools to assist in selecting and filtering data used for model training or adaptation. Understanding the current dataset and its impact on model behavior is critical to detecting data drift.
- **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]

Data management tools and methods are currently an active area of research. However, data drift can be mitigated by incorporating application-specific data management protocols that include: continuous monitoring, retraining (regularly incorporating the latest data into the models), data cleansing (correcting errors or inconsistencies in the data), and using ensemble models (combining predictions of multiple models). Incorporation of a data management framework into the design of AI systems from the beginning is essential for improving the overall integrity and security posture. [31]

Conclusion

Data security is of paramount importance when developing and operating AI systems. As organizations in various sectors rely more and more on AI-driven outcomes, data security becomes crucial for maintaining accuracy, reliability, and integrity. The guidance provided in this CSI outlines a robust approach to securing AI data and addressing the risks associated with the data supply chain, malicious data, and data drift.

Data security is an ever-evolving field, and continuous vigilance and adaptation are key to staying ahead of emerging threats and vulnerabilities. The best practices presented here encourage the highest standards of data security in AI while helping ensure the accuracy and integrity of AI-driven outcomes. By adopting these best practices and risk

mitigation strategies, organizations can fortify their AI systems against potential threats and safeguard sensitive, proprietary, and mission critical data used in the development and operation of their AI systems.

Works cited

[1] Office of the Director of National Intelligence. The Intelligence Community Data Management Lexicon. 2024. https://dni.gov/files/ODNI/documents/IC_Data_Management_Lexicon.pdf

[2] National Security Agency et al. Deploying AI Systems Securely: Best Practices for Deploying Secure and Resilient AI Systems. 2024. https://media.defense.gov/2024/Apr/15/2003439257/-1/-1/0/CSI-DEPLOYING-AI-SYSTEMS-SECURELY.PDF

[3] National Institute of Standards and Technology (NIST). NIST AI 100-1: Artificial Intelligence Risk Management Framework (AI RMF 1.0). 2023. https://doi.org/10.6028/NIST.AI.100-1

[4] NIST. NIST Special Publication 800-37 Rev. 2: Guide for Applying the Risk Management Framework to Federal Information Systems. 2018. https://doi.org/10.6028/NIST.SP.800-37r2

[5] NIST. Federal Information Processing Standards Publication (FIPS) 204: Module-Lattice-Based Digital Signature Standard. 2024. https://doi.org/10.6028/NIST.FIPS.204

[6] NIST. FIPS 205: Stateless Hash-Based Digital Signature Standard. 2024. https://doi.org/10.6028/NIST.FIPS.205

[7] Bommasani, R. et al. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258v3. 2022. https://arxiv.org/abs/2108.07258v3

[8] Securing Artificial Intelligence (SAI); Data Supply Chain Security. ESTI GR SAI 002 V1.1.1. 2021. https://etsi.org/deliver/etsi_gr/SAI/001_099/002/01.01.01_60/gr_SAI002v010101p.pdf

[9] National Cyber Security Centre et al. Guidelines for Secure AI System Development. 2023. https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf

[10] NIST. NIST Special Publication 800-207: Zero Trust Architecture. 2020. https://doi.org/10.6028/NIST.SP.800-207

[11] NIST. NIST IR 8496 ipd: Data Classification Concepts and Considerations for Improving Data Protection. 2023. https://doi.org/10.6028/NIST.IR.8496.ipd

[12] Cybersecurity and Infrastructure Security Agency (CISA), NSA, and NIST. Quantum-Readiness: Migration to Post-Quantum Cryptography. 2023. https://www.cisa.gov/resources-tools/resources/quantum-readiness-migration-post-quantum-cryptography

[13] NIST. FIPS 203: Module-Lattice-Based Key-Encapsulation Mechanism Standard. 2024. https://doi.org/10.6028/NIST.FIPS.203

[14] NIST. NIST SP 800-52 Rev. 2: Guidelines for the Selection, Configuration, and Use of Transport Layer Security (TLS) Implementations. 2019. https://doi.org/10.6028/NIST.SP.800-52r2

[15] NIST. FIPS 140-3, Security Requirements for Cryptographic Modules. 2019. https://doi.org/10.6028/NIST.FIPS.140-3

[16] NIST. FIPS 140-2, Security Requirements for Cryptographic Modules. 2001. https://doi.org/10.6028/NIST.FIPS.140-2

[17] NIST. NIST AI 100-2e2023: Trustworthy and Responsible AI, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. 2024. https://doi.org/10.6028/NIST.AI.100-2e2023

[18] Adak, M. F., Kose, Z. N., & Akpinar, M. Dynamic Data Masking by Two-Step Encryption. In 2023 Innovations in Intelligent Systems and Applications Conference (ASYU) (pp. 1-5). IEEE. 2023https://doi.org/10.1109/ASYU58738.2023.10296545

[19] Kairouz, P. et al. Advances and Open Problems in Federated Learning. Foundations and Trends in Machine Learning 14 (1-2): 1-210. arXiv:1912.04977. 2021. https://arxiv.org/abs/1912.04977

[20] NIST. NIST SP 800-88 Rev. 1: Guidelines for Media Sanitization. 2014. https://doi.org/10.6028/NIST.SP.800-88r1

[21] NIST. NIST Special Publication 800-3 Rev. 2: Risk Management Framework for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy. 2018. https://doi.org/10.6028/NIST.SP.800-37r2
[22] U.S. Department of Homeland Security. Preparedness Series June 2023: Risks and Mitigation Strategies for Adversarial Artificial Intelligence Threats: A DHS S&T Study. 2023. https://www.dhs.gov/sites/default/files/2023-12/23_1222_st_risks_mitigation_strategies.pdf
[23] Bender, E. M., & Friedman, B. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Transactions of the Association for Computational Linguistics (TACL) 6, 587–604. 2018. https://doi.org/10.1162/tacl_a_00041
[24] NSA et al. Content Credentials: Strengthening Multimedia Integrity in the Generative AI Era. 2025. https://media.defense.gov/2025/Jan/29/2003634788/-1/-1/0/CSI-CONTENT-CREDENTIALS.PDF
[25] Executive Order (EO) 14179: "Removing Barriers to American Leadership in Artificial Intelligence" https://www.federalregister.gov/executive-order/14179
[26] NIST. NIST Special Publication 1270: Framework for Identifying and Managing Bias in Artificial Intelligence. 2023. https://doi.org/10.6028/NIST.SP.1270
[27] NIST. NIST AI 600-1: Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. 2023. https://doi.org/10.6028/NIST.AI.600-1
[28] Open Web Application Security Project (OWASP). AI Exchange. #Moretraindata. https://owaspai.org/goto/moretraindata/
[29] Carlini, N. et al. Poisoning Web-Scale Training Datasets is Practical. arXiv:2302.10149. 2023. https://arxiv.org/abs/2302.10149
[30] Kore, A., Abbasi Bavil, E., Subasri, V., Abdalla, M., Fine, B., Dolatabadi, E., & Abdalla, M. Empirical Data Drift Detection Experiments on Real-World Medical Image Data. Nature Communications 15, 1887. 2024. https://doi.org/10.1038/s41467-024-46142-w
[31] NIST. NIST Special Publication 800-208: Recommendation for Stateful Hash-Based Signature Schemes. 2020. https://doi.org/10.6028/NIST.SP.800-208
[32] The Organisation for Economic Cooperation and Development (OECD). Glossary of statistical terms. 2008. https://doi.org/10.1787/9789264055087-en
[33] NIST. NIST SP 800-53 Rev. 5: Security and Privacy Controls for Information Systems and Organizations. 2020. https://doi.org/10.6028/NIST.SP.800-53r5
[34] OWASP. AI Exchange. How to select relevant threats and controls? risk analysis. https://owaspai.org/goto/riskanalysis/

Disclaimer of Endorsement

The information and opinions contained in this document are provided "as is" and without any warranties or guarantees. Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the United States Government, and this guidance shall not be used for advertising or product endorsement purposes.

Purpose

This document was developed in furtherance of the authoring organizations' cybersecurity missions, including their responsibilities to identify and disseminate threats, and to develop and issue cybersecurity specifications and mitigations. This information may be shared broadly to reach all appropriate stakeholders.

Notice of Generative AI Use

Generative AI technology was carefully and responsibly used in the development of this document. The authors maintain ultimate responsibility for the accuracy of the information provided herein.

Contact

U.S. Organizations

- **National Security Agency**
  Cybersecurity Report Feedback: CybersecurityReports@nsa.gov
  Defense Industrial Base Inquiries and Cybersecurity Services: DIB_Defense@cyber.nsa.gov
  Media Inquiries / Press Desk: NSA Media Relations: 443-634-0721, MediaRelations@nsa.gov

Australian organizations

- Visit cyber.gov.au/report or call 1300 292 371 (1300 CYBER1) to report cybersecurity incidents and vulnerabilities.

New Zealand organizations

- For general enquiries, contact info@ncsc.govt.nz