Microsoft Security

AI Security Risk Assessment

Best practices and guidance to secure AI systems

If you have questions, comments or feedback, please contact atml@microsoft.com

Authors and contributors

This is a living document and we welcome comments, feedback and questions. Please reach out to us: **atml@microsoft.com**.

Will Pearce, Senior Software Developer
Hyrum Anderson, Principal Architect
Ram Shankar Siva Kumar, Data Cowboy
Neil Coles, Principal Systems Architect
Andrew Paverd, Senior Researcher
Andrew Marshall, Principal Program Manager
Aashish Bhateja, Principal Program Manager
Alex Sutton, Principal Program Manager

Acknowledgements

We would like to thank Alishba Khawaja, Amanda Minnich, Lu Zhang, Matthieu Maitre, Raja Sekhar Rao Dheekonda, Seetharaman Harikrishnan, Sharon Xia, Sudipto Rakshit, Mario Goertzel, Imran Mohiuddin, John Montgomery, Eric Horvitz, Ann Johnson, Mark Russinovich and Ryan McGee for their feedback and support.

1.0 Executive summary

Despite the compelling reasons to secure ML systems, Microsoft's survey spanning 28 businesses found that most industry practitioners have yet to come to terms with adversarial machine learning (ML). Twenty-five out of the 28 businesses indicated that they don't have the right tools in place to secure their ML systems. What's more, they are explicitly looking for guidance. We found that the lack of preparation is not just limited to smaller organizations – they range from Fortune 500 companies, governments to non-profit organizations. Customers acknowledge the need to secure AI systems but simply do not know how.

This document is a first step for organizations to assess the security posture of their AI systems. But instead of adding yet another framework for organizations to follow, we attempted to provide the content in a manner that can be snapped to existing traditional security risk assessment frameworks.

There are three goals for this whitepaper:

- **Provide a comprehensive perspective to AI system security.** We looked at each element of the AI system lifecycle in a production setting: from data collection, data processing to model deployment. We also accounted for AI supply chain as well as the controls and policies with respect to backup, recovery and contingency planning related to AI systems.
- **Outline threats to critical AI assets and guidance to secure them.** To directly help engineers and security professionals, we enumerated the threat statement at each step of the AI system building process. Next, we provide a set of guidelines which overlay and reinforce existing practices in the context of AI systems.
- **Enable organizations to conduct AI security risk assessments.** The framework provides the ability to gather information about the current state of security of AI systems in an organization, perform gap analysis and track the progress of the security posture.

We formulated it conjunction with stakeholders across Microsoft, with representatives from Azure Security, Responsible AI Strategy in Engineering, Microsoft Security Response Center, Azure Security and AI, Ethics and Effects in Engineering and Research (Aether).

If you have questions or feedback, please contact us at atml@microsoft.com.

2.0 Introduction

We suggest using this document to start the discussion around securing AI systems aligned to on-going information security efforts and business objectives. The document focuses on AI systems, and inclusion of traditional controls because AI systems are built on traditional IT infrastructure.

We cover the following areas specifically related to AI systems.

Administrative controls

Machine learning security policies	Controls and policies relating to the documented policies that govern machine learning, artificial intelligence, and information security.

Technical controls

Data collection	Controls and policies related to the collection, storage, and classification of data that are used for machine learning and artificial intelligence.
Data processing	Controls and policies relating to the processing and engineering of data used for machine learning and artificial intelligence.
Model training	Controls and policies relating to the design, training, and validation of models.
Model deployment	Controls and policies relating to the deployment of models and supporting infrastructure.
System monitoring	Controls and policies relating to the ongoing monitoring of machine learning systems.
Incident management	Controls and policies relating to how incidents related to AI system are handled.
Business continuity and disaster recovery	Controls and policies relating to loss of intellectual property through model stealing, degradation of service, or other AI specific vulnerabilities.

We adapted the existing framework of controls and policies from the popular ISO27001:2013 standard [1] and mapped it across the AI system building process – from data collection phase to responding to threats to AI systems. Organizations may have some or all existing controls implemented from ISO27001:2013 or already be in compliance with several risk frameworks (NIST 800-53 [2], PCI-DSS [3], FedRamp [4], etc) as part of existing information security efforts.

Failing to adequately secure AI systems increases risk to not only the AI systems addressed in this assessment, but more generally to the entire information technology and compliance environment.

The goal of this document is not to replace any of these existing efforts – but to describe securing AI systems from the vantage point of existing tools and frameworks, and extend it to all parts of the AI building process.

The guidance listed here is not prescriptive, as that would require more context such as the underlying platform, the underlying data type and the choice of algorithm. If you are an Azure ML customer, please refer to Enterprise security and governance - Azure Machine Learning | Microsoft Docs.

3.0 Suggested severity, likelihood, impact

Not all controls will be of critical importance to the security of an AI system. Therefore, to properly prioritize work, each control should be rated by the organization with a severity rating that is relevant to the business impact of not implementing a given control. An organization may choose to accept the risk of a critical control, and instead implement a compensating control to lower the risk. Ultimately, these ratings are to help guide risk-based decision making rather than to prescribe activities.

3.1 Severity

The severity of a compromise is going to depend on the AI model use case. Fortunately, if the data or systems being used were of critical concern prior to machine learning being integrated, it should remain the same. Similarly, if the model used is "off-the-shelf" with no additional input, depending on the context the model is used in, the severity of a compromise is likely lower. Techniques like differential privacy can reduce the potential impact of a compromise. However, this would not reduce the criticality of the system, the data, or the model. We recommend that models be protected using a defense-in-depth strategy rather than relying on any one defensive implementation.

Suggested severity level

Suggested as critical

- If the AI model is trained on, or ingests sensitive personal data, classified data, or data governed by compliance requirements such as PCI, HIPAA, GLBA, etc.
- If the AI model is used in a business-critical application or system such that compromise would have a large negative impact of business operations
- If the AI model is used in applications where physical or harm or death is a possible outcome
- If the AI model is used in a system that supports critical infrastructure (e.g. water, power, health)

Suggested as high

- If the AI model is trained on or ingests sensitive personal data, confidential information, or data that is otherwise considered critical by the organization
- If compromise of this AI model would have large but scoped impact on business operations
- If the AI model is used in business-critical applications or systems

Suggested as medium

- If the AI model is trained on a subset of training data that contains sensitive data types
- If compromise of this AI model would have implications for models deployed in production
- If the AI model is used in non-critical but business facing applications
- If the AI model is not used in production but has information regarding production models

Suggested as low

- If the AI model is trained on data that is not used in production
- If the AI model is not used in production, and does not have information regarding production models

Suggested as informational

- If the data is unclassified from a vetted source
- If the AI model isn't used in production

3.2 Likelihood

Likelihood has two major components, availability of the model, and the availability of techniques. To reduce the likelihood of an attack, an organization should implement controls that:

1. Remove attack surface or make attack surface harder to enumerate.
2. Ensure logging and alerting are working as designed to ensure quick resolution of issues.
3. Ensure that all supporting systems are up to date with security requirements.

Controls could include gating endpoints, network segmentation, or rate limiting. Special attention should be paid to traffic flows and network or pipeline diagrams, for example, an attacker compromising and external facing endpoint and working backwards through a pipeline.

3.3 Impact

Impact is related to affects to the organization. We suggest that you begin with familiarizing yourself with different ways ML systems can be attacked and consider ways in which production models can affect the organization (see Failure Modes in Machine Learning - Security documentation | Microsoft Docs). Once this is done, it can be mapped to a severity matrix.

3.4 Severity matrix

Below is a basic risk and vulnerability severity matrix to get organizations started. We suggest filling up a similar categorization by convening security architects, machine learning engineers and AI red team members.

Attack type	Likelihood	Impact	Exploitability
Extraction	High	Low	High
Evasion	High	Medium	High
Inference	Medium	Medium	Medium
Inversion	Medium	High	Medium
Poisoning	Low	High	Low

> “Designing and developing secure AI is a cornerstone of AI product development at BCG. As the societal need to secure our AI systems becomes increasingly apparent, assets like Microsoft’s AI Security Risk Management Framework can be foundational contributions. We already implement best practices found in this framework in the AI systems we develop for our clients and are excited that Microsoft has developed and open sourced this framework for the benefit of the entire industry.”
>
> —Jack Molloy, Senior Security Engineer, Boston Consulting Group

4.0 Basic use

The rest of the document follows this structure:

- a **risk control** contains a description of which area the control covers
- the **objective** of the control and what it is supposed to accomplish
- a **threat statement** that gives a description of the risk being mitigated
- and finally, **guidance** for implementing a control. We understand that not all guidance can be implemented for legitimate business reasons. We suggest documenting those that have not been implemented.

Below is a control pulled from the AI systems risk assessment, notes have been added to describe each part of a risk categories structure.

Example control	How to read it
1. Data collection	**Primary category**
Controls and policies relating to the collection and storage of data from all sources that are used for machine learning and artificial intelligence.	Describes what controls in this category will cover at a high level.
2. Data sources	**Control category**
Objective: To ensure the integrity of data collected that is used for trained models.	Should describe the risk being mitigated with the controls.
Threat statement: Data is collected from untrusted sources that could contain Sensitive Personal Data, other undesirable data that could affect the security of a model or presents compliance risks to the organization.	A statement that describes the outcome of not implementing the control.
Control: Data should be collected from trusted sources. A list of trusted sources should be kept and updated. Approvals for collecting untrusted data should be considered on a case-by-case basis.	Specific verbiage that describes best practice for the control.
Guidance: 1. All reasonable effort should be made to ensure that data can be trusted before training a model. Untrusted or unknown data could introduce security vulnerabilities later in the pipeline. 2. Data that contains sensitive personal data whether used for data science purpose or otherwise should either be cleaned or stored and access appropriately. 3. Collecting data without consideration for its context could result in datasets that contain illegal data. Data collection efforts should be mindful about copyrighted material, data breaches, unsecured endpoints that accidently leak data.	Guidance is recommendations for satisfying the above criteria. We provide them in a product and vendor agnostic way to give room for organizations to solve the issue in a way that makes sense for them.

5.0 Machine learning security assessment

5.1 Before getting started

The purpose of this assessment is to help organizations articulate, track, and remediate risks to business operations introduced by AI systems. This assessment should be used to:

1. Gather information about the current state of AI security within the organization.
2. Perform a gap-analysis and build a roadmap for implementing recommendations.
3. Track security progress by performing this assessment annually or bi-annually.

If an organization has no security program, this assessment is not the place to start. An organization should have a functioning information security program prior to implementing recommendations in this assessment (see Azure security guidance - Cloud Adoption Framework | Microsoft Docs).

5.2 Data collection

Controls and policies relating to the collection and storage of data from all sources that are used for machine learning and artificial intelligence.

Objective: To ensure the integrity of data collected used in AI systems.

5.2.1 Data sources

Control: Data should be collected from trusted sources. A list of trusted sources should be kept and updated. Management approvals for collecting untrusted data should be considered on a case-by-case basis. If an untrusted source is approved, it should be documented.

Threat statement: Data is collected from untrusted sources that could contain sensitive personal data, other undesirable data that could affect the performance of a model or presents compliance risks to the organization.

Guidance:

1. Input data should be validated and trusted via management approval prior to use in an AI system.
2. Data collected for the purpose of an AI system should be reviewed prior to use or storage.
3. If appropriate, collected data should be cleaned of undesirable entries.
4. The source of data should be documented and kept with the data.
5. Inference data used to train a model should not be implicitly trusted and should be treated as new data.
6. Data collection efforts should be documented and audited. Collected data should have an owner who is responsible for its adherence to documented policies.

5.2.2 Sensitive data types

Control: To ensure stored data for AI systems is properly secured, tracked, and classified according to its sensitivity and use case. This includes appropriate, data classification labels, access policies, license information, descriptive statistics, originating source, and date of collection.

Threat statement: Data used in AI systems is used, stored, or accessed inappropriately due to a lack of required attributes, metadata, or documentation.

Guidance:

1. Develop a data policy that encompasses the privacy and protection of sensitive data types and communicate the policy to all personnel involved with the use or creation of AI systems.
2. Implement training and deployment pipelines that protect the confidentiality and integrity of the data used in AI Systems.

5.2.3 Data storage

Control: Data should be appropriately stored according to a documented classification process. Datasets should be indexed and considered an asset that is subject to asset management and access control policies.

Threat Statement: Data is stored insecurely and can be tampered with or altered by unauthorized parties or systems. Data is not correctly classified, leading to the disclosure of confidential information or sensitive personal data.

Guidance

1. Ensure development or AI research systems or accounts do not have access to production databases and vice versa.
2. Data used in AI systems should be classified and protected according to a documented classification policy.
3. Data used in AI systems is tracked under a documented asset management policy.
4. Data used for sensitive AI use cases are stored on approved and managed systems.
5. Access to data should be audited, and users requesting access should go through a formal access control process that includes management approval.
6. Data used in machine learning processes should not be exposed to the internet.
7. Data that is pulled from the internet (or other untrusted sources) should go through a filtering process that include management approval.
8. Datasets should be versioned with formal change control processes in place.

5.2.4 Data access

Control: Datasets should be appropriately tracked and verified via cryptographic hash before use.

Threat statement: Datasets are altered without authorization.

Guidance:

1. Role based access control for datasets should be enforced.
2. Perform regular access audits to ensure those with access to datasets should have access to datasets. Ensure that each account is operating within normal bounds.
3. If a central tracking platform is not used, access to data through raw access logs should be reviewed for purpose. Ensure that each account is operating within normal bounds.
4. Third party resource providers, contractors or other external parties should not have excess or inappropriate access to a company's train/test data assets without contracts in place.

5.2.5 Data integrity

Control: Datasets should be trusted and remain trusted throughout the AI system lifecycle.

Threat statement: Datasets are altered during the AI life cycle without the ability to audit or track changes.

Guidance:

1. Datasets should be uniquely identified such that unauthorized changes to an approved dataset would cause a review of the dataset.
2. Datasets and their cryptographic descriptions should be tracked in a central location. Access to the dataset should be audited.
3. Changes to the dataset should include an updated cryptographic descriptions and management approval before being submitted to the central tracking service.

5.3 Data processing

Controls and policies relating to the processing of data that is used for machine learning and artificial intelligence.

Objective: To ensure the secure processing of data from its raw form to an intermediary form ready for training.

5.3.1 Processing pipelines

Control: Processing pipelines should be adequately secured.

Threat statement: A threat actor can make unauthorized changes to the system by altering the data processing pipelines.

Guidance:

1. Not all data that moves through a production system is relevant to data science efforts. It is important to parse out only the required data, and ensure all data moved from a secure production setting into a development setting is appropriately tracked. Consider that certain types of data may not be able to be moved into a development environment and that data science may need to occur in a secure intermediary environment.

2. Proper auditing of data access throughout data processing lifecycle is important. Without separate accounts there is no can be no sufficient auditing of access. Further, the ability to respond to an incident cannot happen without potentially affecting business processes. Compromise of a single account would result in compromise of all data leaving the secure production environment.

3. Data science processes could require resources that are outside of a strict compliance boundary.

4. Data science processes should always be compliant with existing requirements. This could include moving data science resources and processes into a compliant environment.

5. Data should be tracked through its entire lifecycle; this includes subsets of larger datasets. It should be required that a model can be traced back to the data it was trained on. Further, a copy of that data should exist in its entirety.

5.3.2 Dataset aperture

Control: To ensure subsets (e.g., temporal, categorical slices) of data included for model building and how might that impart security hazards (privacy leakage, poisoning/integrity via overemphasis on feedback, etc.).

Threat statement: Threat actor can recover parts of the data by reconstructing/recovering subsets of data.

Guidance:

1. Subsets of data are datasets themselves. These subsets are required to have the same metadata attached to them as the parent dataset and should be similarly reviewed for sensitive data types.

2. Depending on policies regarding machine learning practices (SLA's, bias metrics, etc), any given dataset (including subsets) should meet a minimum documented standard surrounding these metrics if they are to be used in model building. The metadata should be always attached to the dataset.

3. All datasets that violate existing policies should have a documented exception that has been approved by management. Included in the exception should be a documented reason for the exception in addition to the required metadata.

4. All data used for model building should be tracked in a central location. Data should be auditable at any time. Additionally, models found to be trained on untracked data should be pulled from production until they are matched with a known dataset with the required metadata.

5. Datasets should be appropriately versioned such that all metadata is updated, and users of the data understand the contents and the statistical properties. If necessary management approval for sensitive use cases should be required.

5.4 Model training

Controls and policies relating to the training of models and algorithms.

5.4.1 Model design

Control: Model training code is reviewed by a responsible party.

Threat statement: Improper code or vulnerabilities in model code produce availability, integrity, or confidentiality risks.

Guidance:

1. Model design and research should happen in the appropriate environment. Model design and architecture can have a large effect on the efficacy of a model. Production environments are not the place for research or to test non-provable claims about efficacy of a design.
2. Model selection for a production system should be reviewed and approved by management. This should happen early in the development phase and should be tracked through any available mechanism (Excel, DevOps, Git, etc). Exceptions should be documented.
3. Models are often domain specific and there should be adequate documentation accompanying the model throughout its use in an organization.
4. Ensure model metadata is accessible to users and unapproved uses of models are documented and enforced. It is acceptable for a user to fine-tune an existing model so long as new meta-data is attached and tracked appropriately.

5.4.2 Model training

Control: The model selection criterion (metric and holdout sets) mimics natural drift and any adversarial conditions that may be expected at deployment time.

Threat statement: A model that is trained under ideal conditions is likely to be brittle when deployed in adversarial settings.

Guidance

1. Training and validation sets should respect natural temporal dependencies. For example, for malware classifiers, a validation set should include only software versions later than those contained in the training set.
2. Explicitly add model robustness by augmenting datasets with common corruptions that could be reasonably be discovered in the wild.
3. Explicitly train against worst-case conditions using adversarial retraining.
4. Track experiments and associated meta.

5.4.3 Model selection

Model selection consists of choosing one model from a set of candidates, where each candidate has a unique set of model parameters, training algorithm and training hyper-parameters. The selection criterion for the winning model is often based on a single quantifiable metric (e.g., minimum loss, maximum detection rate) as measured on a common holdout dataset, or as averaged across a K-fold validation set.

Control: Model design and training algorithm includes explicit or implicit model regularization.

Threat statement: Models are overfit to a training and/or single validation dataset and are more vulnerable to failure modes.

Guidance:

1. Where computationally feasible, K-fold cross-validation should be used to prevent overfitting to a single holdout set.
2. Verify that selected models perform well on disparate holdout sets to validate that they have not overfit.
3. Ensure that processes exist.

5.4.4 Model versioning

Control: Models are continuously retrained as new training data flows into training pipelines.

Threat statement: An incident occurs but the model involved cannot be located for investigation.

Guidance:

1. Version models such that every time a model is trained it is assigned a new version. Qualifiers such as my_model_dev_1.1 or my_model_prod_1.1 should be used to delineate. Production from pre-production models. This will also help isolate issues to either a production or pre-production issue. Reference existing secure SDL processes or policies.

5.5 Model deployment

Controls and policies relating to the deployment of models, algorithms, and supporting infrastructure.

5.5.1 Security testing

Control: Models put into production are adequately secured.

Threat statement: AI systems are not adequately tested for vulnerabilities prior to deployment.

Guidance:

1. Formal acceptance testing criteria have not been defined and documented for new AI systems, upgrades, and new versions.

2. New AI systems, upgrades or new versions should be implemented with formal testing.

3. Automated tools should be leveraged for testing information systems, upgrades, or new versions.

4. Test environment should closely resemble the final production environment.

5. The frequency, scope, and method(s) for independent security reviews should be documented.

5.5.2 Security and compliance review

Control: Robust management of the underlying network is key to securing the ML system and the infrastructure.

Threat statement: Compromise of the ML system by accessing the unsecured network.

Guidance:

1. Gateway devices to ML systems should be configured to filter traffic between domains and block unauthorized access.
2. Relevant statutory, regulatory, and contractual requirements should be explicitly defined and documented, and addressed, alongside specific controls and individual responsibilities.
3. Secure configuration guidelines should also be documented, implemented, or reviewed.
4. The criterion for the segregation of ML networks into domains should be consistent with the organization's access control policy or access requirements of the organization.
5. Mechanisms such as secure gateway, VPN, routing for ML systems should be implemented sufficiently to enable a graduated set of controls.
6. Users and ML engineers should employ or follow requirements for the implementation of controls to properly segregate and restrict use of publicly accessible systems, internal networks, and critical assets.

5.6 System monitoring

Controls and policies relating to the ongoing monitoring of machine learning systems and supporting infrastructure.

5.6.1 Logs and log review

Control: Logging and monitoring is vital for ML systems for security reasons.

Threat statement: During an investigation, logs for ML systems are not found.

Guidance:

1. Logging and monitoring should occur consistently across all AI systems and their components, including storage, pipelines, production servers, etc.
2. Event and security logs should be reviewed regularly for abnormal behavior.
3. Consolidated reports and alerts on system activity should be generated and reviewed by management or a security representative.

5.7 Incident management

5.7.1 Roles and responsibilities

Control: Security logs should be collected in a central location.

Threat statement: During an investigation, security analysts do not have a formalized playbook.

Guidance:

1. Organizations for must follow a formal process to report AI systems incidents in the context of loss of service, loss of equipment, loss of facilities, system malfunctions, system overloads, human errors, and non-compliances with policies or guidelines, breaches of physical security, uncontrolled system changes, software malfunctions, hardware malfunctions, and access violations.
2. Formal incident response and escalation procedures should be developed to document actions taken on receipt of a report of an information security event.
3. Incident response procedures should be tested on a periodic basis, tracking response metrics.

5.8 Business continuity planning

5.8.1 Planning, review and outcomes

Control: Ensure that ML systems can be remediated and recovered after an incident.

Threat statement: Incidents cause persistent confidentiality, integrity, or availability issues to critical ML systems.

Guidance:

1. Critical AI assets should be identified and inventoried.
2. The organization should develop a business continuity plan (BCP) or disaster recovery (DR) process in the face of attacks on AI systems.
3. The organization must identify prioritized the risks associated with the impact of losing critical AI systems to attacks.
4. Organizations must have a business continuity testing operated on a repeated schedule for critical AI systems.

References

[1] ISO 27001 Annex A Controls - Overview (isms.online)

[2] Official PCI Security Standards Council Site - Verify PCI Compliance, Download Data Security and Credit Card Security Standards

[3] Failure Modes in Machine Learning - Security documentation | Microsoft Docs

[4] Threat Modeling AI/ML Systems and Dependencies - Security documentation | Microsoft Docs

[6] AI/ML Pivots to the Security Development Lifecycle Bug Bar - Security documentation | Microsoft Docs

[7] Enterprise security and governance - Azure Machine Learning | Microsoft Docs

If you have questions, comments or feedback, please contact atml@microsoft.com