MITRE

MP250397
MITRE WORK PRODUCT

SAFE-AI
A Framework for Securing AI-Enabled Systems

Authors:
J. Kressel
R. Perrella
E. Reed
N. Naik
J. Sidhu
Q. Hu
L. Booker
J. Cintron
L. Huffner

April 2025

Approved for Public Release; Distribution Unlimited. Public Release Case Number 25-1028.

McLean, VA

Executive Summary

The SAFE-AI framework emphasizes the importance of thoroughly evaluating the risks introduced by AI technologies when they are integrated into system architectures. It advocates for the careful selection of security controls that align with the level of risk posed by these advancements. SAFE-AI aims to strengthen the processes of security control selection and assessment by ensuring that AI-specific threats and concerns are systematically identified and addressed. SAFE-AI is based on National Institute of Standards and Technology (NIST) standards and the MITRE Adversarial Threat Landscape for Artificial Intelligence Systems (ATLAS)™ framework.

AI contributes to the attack surface through its inherent dependency on data and corresponding learning processes. Attacks include adversarial inputs, poisoning, exploiting automated decision-making, exploiting model biases, and exposure of sensitive information.

Another significant concern is the presence of supply chain vulnerabilities and the associated risks stemming from unclear provenance of AI models. AI systems often rely on third-party libraries, frameworks, and pre-trained models, which may contain hidden vulnerabilities or malicious code. The high cost of training new LLM models, for example, means that most organizations will acquire and execute models either from open source or proprietary sources with little or no method of determining the risks associated with executing such a model.

To effectively manage the threat landscape, SAFE-AI recommends controls from the NIST control catalog, Special Publication 800-53, Rev 5 and provides potential mitigations and a discussion of residual risk.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Purpose

The purpose of this paper is to provide guidance on securing AI-enabled systems. AI is rapidly changing the nature of Information Technology (IT) systems, incorporating advanced techniques for information processing, and is introducing new vectors for adversarial actions that greatly expands the attack cross-section of IT systems.

Federal departments and agencies generally follow the NIST Risk Management Framework (RMF) for securing their information systems. The RMF calls for implementing security controls from the NIST control catalog, Special Publication 800-53, Rev 5 [1]. NIST issues new guidance when nascent technologies become prevalent, however, as of this writing NIST has not issued a revised control catalog to address securing AI-enabled systems.

The guidance presented in this paper is focused on selection of security controls that address threats specific to AI and should be considered for implementation in AI-enabled systems.

1.2 Audience

The primary audience for this paper is cybersecurity and AI professionals with technical responsibilities for securing information systems, developing system security plans (SSPs), planning and performing security control assessments (SCAs), or developing system architectures that defend against adversarial AI.

The content and concepts herein should also benefit technical leadership tasked with managing cybersecurity for AI-enabled systems.

1.3 Scope

The scope of this paper covers adversarial threats, concerns, security controls, and guidance associated with the secure use of AI-enabled systems.

This paper considers the breadth of AI, including but not limited to Large Language Models (LLMs) and Generative AI (GenAI) and more conventional machine learning (ML) techniques since there are many specific applications of AI beyond GenAI.

This paper does not address AI objectives identified in various Executive Orders (EOs) concerning the principles of fairness, equity, explainability, and reliability. These topics are not directly related to the security of AI-enabled systems and, as such, are out of scope.

Some security controls are especially crucial when AI technologies are included in system architectures. They may also require better understanding of the adversarial threat landscape and require corresponding attention during control mitigation.

1.4 Background

As AI technologies have become more prominent in IT systems, MITRE undertook developing methods for addressing the AI threat landscape by producing the MITRE ATLAS™.

MITRE ATLAS™ uses a similar taxonomy as the MITRE ATT&CK®, which has been adopted by federal public and private organizations worldwide.

The MITRE ATLAS™ is a framework for managing the adversarial AI landscape covering tactics from reconnaissance to attacks on AI and their impact. ATLAS™ is based on real-world attack observations and techniques as well as demonstrations from AI red teams and security researchers. This paper comes from a complementary perspective, that of the defender, focusing on the protection of enterprise assets that use AI systems.

To meet the challenge, this paper presents a framework for addressing the security of AI-enabled systems.

1.4.1 Why AI-focused Security Guidance is Necessary

Machine learning is now commonly being used in document classification, text summarization, image identification, spam filtering, text, and image generation, etc.

AI systems have a set of risks that are not comprehensively addressed by current risk frameworks and approaches [2].

Because AI systems have unique and often non-deterministic behaviors which are unlike traditional IT systems, security assessors who have based their system security assessment criteria on traditional, non-AI-enabled IT systems, may fail to address important AI-specific threats and vulnerabilities.

The high cost of training new LLM models, for example, means that most organizations will acquire and execute models either from open source or proprietary sources with little or no method of determining the risks associated with executing such a model. Models are essentially opaque.

There is also the issue of unintended embedding of sensitive information (such as Personally Identifiable Information (PII), Protected Health Information (PHI), or Federal Tax Information (FTI) for example) into a model. With a locally executing model, if the model is not changed during exposure to sensitive information, we can be certain that no leakage can occur (as a direct result of this exposure). However, if through whatever process, the model changes after exposure to sensitive information, then we must assume that information is now part of the model. The model must now be treated as though it was sensitive information because it may be possible, through prompt engineering or other AI/ML query methods, to extract sensitive information. This influences the IT security of systems because, per the Bell-LaPadula model [3], sensitive information should not flow from high security to lower security environments. This is particularly important in software configuration management promotion models, where code and data are typically promoted from lower environments to higher security environments eventually reaching the "production" environment. This is the same promotion scheme used by Continuous Integration/Continuous Deployment (CI/CD) systems.

Another issue is that of supply chain vulnerabilities and the related risks associated with lack of clear provenance when it comes to AI models. Commercial LLM models are trained with huge, ill-defined data sets. The content of such data sets is not well characterized by vendors. As a result, it is unclear what has been trained into the data set. The AI risks identified earlier by NIST hold for these models. Adversaries, such as China, are now producing their own models, and the provenance of training data and optimization techniques are unknown.

To further exacerbate the problem, NIST has yet to publish guidance for securing AI systems.

To address this gap, MITRE set out to identify the threats and concerns with AI-enabled systems, using NIST Special Publication (SP)-800-53 Rev 5 and the associated NIST RMF frameworks to provide guidance to assessors that:

1. Identifies new threats for which assessors may not be prepared.
2. Identifies concerns specific to AI-enabled systems.
3. Provides supplemental assessment criteria that can be used to bolster existing security assessment criteria to meet the needs of AI-enabled systems.
4. Ultimately, to make it possible to develop and secure trustworthy AI-enabled systems.

Since existing frameworks are used to express the threats and controls, the impact on assessment plans should be limited to affected controls. In addition, the assessment process can remain largely the same, allowing assessment teams to be productive within familiar processes.

1.4.2 Terminology

This paper uses the following terms:

- **AI-enabled System:** An IT system using one or more AI technologies to infer, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" [4].
- **System Element**: One of four aspects of an IT system used for analyzing security risks. All four (4) elements form a comprehensive set. That is, a risk must be addressed in at least one of these system elements. The four (4) system elements are defined here:
  - **Environment** – the operational setting of an AI-enabled system, including infrastructure, network, compute, and storage.
  - **AI Platform** – the application software, AI software, and the operating system.
  - **AI Model** – a software program and its algorithms that receives input data, such as text, images, or numbers, and processes the input to produce output, such as predictions, recommendations, or generated content.
  - **AI Data** – data used for training and tuning AI models.
- **Residual Risk:** The portions of risk that remain after security measures have been applied. It is the portion of risk that is not eliminated or mitigated through risk management strategies and controls. Thus, residual risk is the risk that organizations must accept, transfer, ignore, or manage through other means, such as insurance or contingency planning. [5] It is important for organizations to identify and understand residual risk to make informed decisions about their risk tolerance and to ensure that they have appropriate measures in place to address potential impacts.
- **Organizational Common Controls:** The role of organizational common controls is to provide a common baseline of controls, shared across IT systems, providing the benefits of standardization and reusability.

2 SAFE-AI

SAFE-AI is the name given to the framework and associated processes to secure AI-enabled systems. The framework is based on existing MITRE ATLAS™ and MITRE ATT&CK frameworks but also integrates aspects of the NIST Risk Management Framework and the NIST AI Risk Management Framework. By harmonizing these frameworks, SAFE-AI provides practical guidance for securing AI-enabled systems.

2.1 Framework Basics

The SAFE-AI framework follows existing guidance but proposes a threat-informed model based on AI threats and concerns. The tabular framework tracks how each concern or threat is addressed by security controls, for each system element. The system element perspective ensures that the characteristics of the system element are addressed by assessment guidance. Often, the guidance differs by system element, even for the same control.

2.1.1 Relationship with MITRE ATLAS™

MITRE ATLAS™ is a "globally accessible, living knowledge base of adversary tactics and techniques against AI-enabled systems based on real-world attack observations and realistic demonstrations from AI red teams and security groups" [6]. The MITRE ATLAS framework follows from MITRE ATT&CK framework, and is a compendium of threats, tactics, and mitigations specific to AI-enabled systems. Not every adversarial action is knowable or controllable, but many can be anticipated, and that is where SAFE-AI elaborates concerns regarding threats to AI and identifies controls that are especially needed to secure AI-enabled systems. SAFE-AI addresses many of the ATLAS™ techniques, shown in Appendix A.

Note: There are also non-adversarial threats. For example, accidental or unintended threats, which may have similar consequences and should be addressed in any threat-informed defense of AI-enabled systems.

One important contribution from ATLAS™ is the taxonomy of tactics and techniques against AI-enabled systems. SAFE-AI leverages this taxonomy for identifying AI-specific concerns and determining controls that may be used to mitigate the concern.

2.1.2 Relationship with the NIST AI RMF

The NIST AI RMF "is intended for voluntary use and to improve the ability to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems. " [2].

Figure 1 NIST AI RMF Overview

The NIST AI Framework [2] identifies the following "characteristics of trustworthy AI systems":

- Safe
- Secure and Resilient
- Explainable and Interpretable
- Privacy-Enhanced
- Fair - with Harmful Bias Managed
- Accountable and Transparent
- Valid and Reliable

2.1.3 Framing the AI Risks

The primary goal of the NIST AI RMF is to frame the risks associated with AI, increase awareness, leading to better decision making with regards to AI system commissioning and deployment. This helps us understand the risk, impacts, and harms.

The NIST AI RMF includes a list of 12 AI Risks, reproduced here [2]:

1. Chemical, Biological, Radiological, and Nuclear (CBRN) Information or Capabilities: Eased access to or synthesis of materially nefarious information or design capabilities related to CBRN weapons or other dangerous materials or agents.
2. Confabulation: The production of confidently stated but erroneous or false content (known colloquially as "hallucinations" or "fabrications") by which users may be misled or deceived.
3. Dangerous, Violent, or Hateful Content: Eased production of and access to violent, inciting, radicalizing, or threatening content as well as recommendations to carry out self-harm or conduct illegal activities. Includes difficulty controlling public exposure to hateful and disparaging or stereotyping content.
4. Data Privacy: Impacts due to leakage and unauthorized use, disclosure, or de-anonymization of biometric, health, location, or other personally identifiable information or sensitive data.

5. Environmental Impacts: Impacts due to high compute resource utilization in training or operating Generative AI models, and related outcomes that may adversely impact ecosystems.
6. Harmful Biases or Homogenization: Amplification and exacerbation of historical, societal, and systemic biases; performance disparities between sub-groups or languages, possibly due to non-representative training data, that result in discrimination, amplification of biases, or incorrect presumptions about performance; undesired homogeneity that skews system or model outputs, which may be erroneous, lead to ill-founded decision-making, or amplify harmful biases.
7. Human-AI Configurations: Arrangements of or interactions between a human and an AI system which can result in the human inappropriately anthropomorphizing Generative AI systems or experiencing algorithmic aversion, automation bias, over-reliance, or emotional entanglement with Generative AI systems.
8. Information Integrity: Lowered barrier to entry to generate and support the exchange and consumption of content which may not distinguish fact from opinion or fiction or acknowledge uncertainties or could be leveraged for large-scale dis- and mis-information campaigns.
9. Information Security: Lowered barriers for offensive cyber capabilities, including via automated discovery and exploitation of vulnerabilities to ease hacking, malware, phishing, and offensive cyber.
10. Intellectual Property: Eased production or replication of alleged copyrighted, trademarked, or licensed content without authorization (possibly in situations which do not fall under fair use); eased exposure of trade secrets; or plagiarism or illegal replication.
11. Obscene, Degrading, and/or Abusive Content: Eased production of and access to obscene, degrading, and/or abusive imagery which can cause harm, including synthetic child sexual abuse material (CSAM), and nonconsensual intimate images (NCII) of adults.
12. Value Chain and Component Integrations: Non-transparent or untraceable integration of upstream third-party components, including data that has been improperly obtained or not processed and cleaned due to increased automation from Generative AI; improper supplier vetting across the AI lifecycle.

2.1.4 Relationship with the NIST Risk Management Framework

The NIST RMF is the principal organizing structure for identifying and managing risks in the Federal IT domain. It provides a structure for organization-wide Risk Management at three levels: organization, mission, and system [7].

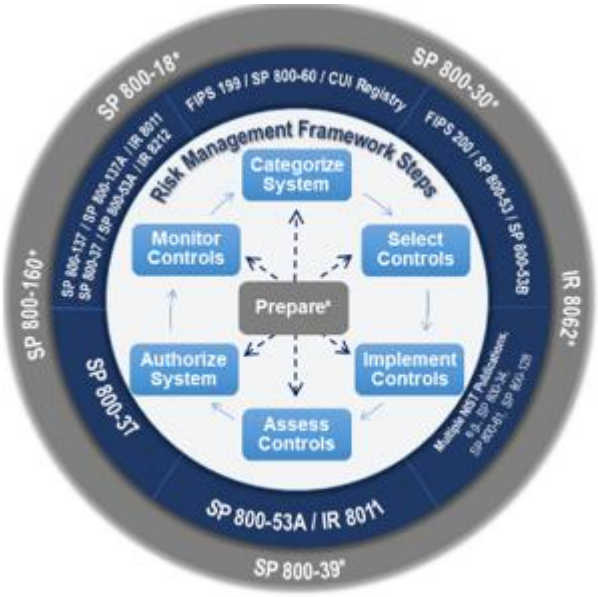

Figure 2 NIST Risk Management Framework

The NIST RMF [7] comprises 7 steps:

1. Prepare
2. Categorize
3. Select
4. Implement
5. Assess
6. Authorize
7. Monitor

SAFE-AI addresses the following three RMF steps: Prepare, Select, and Assess. The other steps, (i.e., Categorize, Implement, Authorize, and Monitor) are largely unaffected by this guidance and are well documented in NIST literature.

The following sections describe how the associated RMF Step is addressed by SAFE-AI.

2.2 RMF Step: Prepare

The NIST RMF outlines eighteen (18) tasks within the “Prepare" step. The SAFE-AI framework suggests enhancing four (4) of these tasks:

- **Risk Management Roles:** Organizations should identify AI subject matter experts (SMEs) to aid in managing security and privacy risks specific to AI-enabled systems.
- **Risk Management Strategy:** Organizations should reassess their risk tolerance levels specifically for AI-related risks, referencing the NIST AI RMF's guidelines on AI risks.
- **Organization-Wide Risk Assessment:** This task should incorporate the SAFE-AI framework to better prepare for assessing AI-enabled systems.
- **Common Control Identification:** Organizations should review and adjust their common controls to ensure they are applicable to AI-enabled systems.

The RMF Prepare step is unique in that it is executed prior to each of the other RMF steps.

2.2.1 Additional Considerations: Addressing AI Concerns

During the preparation stage, assessors must plan to address each AI concern.

The SAFE-AI framework maps MITRE ATLAS™ Threats to the four (4) system elements (Environment, AI Platform/Tools, AI Models, and AI Data).

At the intersection of each Threat and System Element, relevant NIST SP-800-53 Controls are enumerated, and an AI Concern is expressed in prose.

For example, for "Loss of Model", the AI Concern expresses the risk of model destruction or corruption by an attacker. For the "Environment" system element, it identifies Controls AC-03, AC-06, and CM-07 as being particularly relevant.

The SAFE-AI framework enumerates security controls for the four (4) system elements comprising AI-enabled systems: Environment, AI Platform, AI Model, and AI Data.

For each threat identified by SAFE-AI, a set of corresponding AI concerns is also identified. The concerns address specific vulnerabilities in the context of the four (4) system elements and the general class of threat. Some concerns also include associated mitigations, when appropriate.

For example, for the threat of "loss of models", the corresponding AI concerns include destruction or corruption of a model due to vulnerabilities in the system elements, caused by vulnerabilities such as outdated or unpatched software components, improperly enforced access control, or poor asset protection management practices. Specific controls are then allocated to each system element.

2.3 RMF Step: Select

The RMF Select Step includes Control Selection, Control Tailoring, Control Allocation, Document Planned Control Implementations, Continuous Monitoring Strategy – System, and Plan Review and Approval tasks.

In summary:

- Control selection is the task that selects the appropriate controls for the agency.
- Control Tailoring is the task that adapts the controls to the agency.
- Control Allocation identifies controls from the agency baseline and allocates them to the appropriate AI Threats and Concerns, categorized by System Elements. The same control may appear in several System Elements. This is normal and expected.
- Documenting planned control implementation is the task that documents, typically within the SSP.
- Continuous Monitoring Strategy – System is the task that defines how the system will be continuously monitored, including its integration into the agency's overall continuous monitoring strategy.
- Plan Review and Approval tasks by senior leadership occur after all prior tasks have completed.

In SAFE-AI, the tasks for Control Selection, Tailoring, Allocation, and Document Planned Control Implementations are augmented to address AI specific threats. The remaining tasks are unchanged.

Prepare for these tasks by reviewing the SAFE-AI matrix to help select controls for the system, across all four System Elements.

The result of the RMF Selection task is to augment the System Security Plan. This plan should address the AI threats and Concerns identified by SAFE-AI along with the usual risks.

2.3.1 Additional Considerations: AI Affected Controls

In the SAFE-AI framework, one hundred (100) NIST SP-800-53 controls are identified as potentially AI-affected. These are listed in Appendix D.

2.4 RMF Step: Assess

During the Assess Step, the appropriate assessor or assessment team is selected.

Next, the Security Assessment Plan (SAP) is produced by the assessment team to document how the assessment will be conducted, identify key inputs to the processor, establish rules for collaboration, and define the scope of the assessment. Control assessments are performed using the procedures defined in the SAP. Remediation actions are conducted based on the Security Assessment Report (SAR), leading to a reassessment of the remediated controls. Finally, a Plan of Action and Milestones (POA&Ms) is produced based on the findings and recommendations documented in the SAR.

The Q&A sets, found in Appendix E, are used by assessors to augment the SAP for conducting interviews with system owners and other appropriate stakeholders.

SAFE-AI includes guidance for planning SCAs but does not change any of the assessment and reporting requirements for the SAR or POA&Ms.

2.4.1 Additional Considerations: Hybrid Controls

AI-enabled systems may use agency or 3rd party (typically Cloud) services to implement AI services. Security controls that apply to AI services are typically Hybrid controls and not Common Controls. That is, the system that uses the AI service (whether on-premises or cloud-based) has some number of responsibilities to implement all or part of each of the associated controls. The remaining parts may be inherited from the AI service provider. During control allocation and documentation, the exact parts which may be inherited, or which must be directly addressed by the service consumer vs. the service producer must be determined.

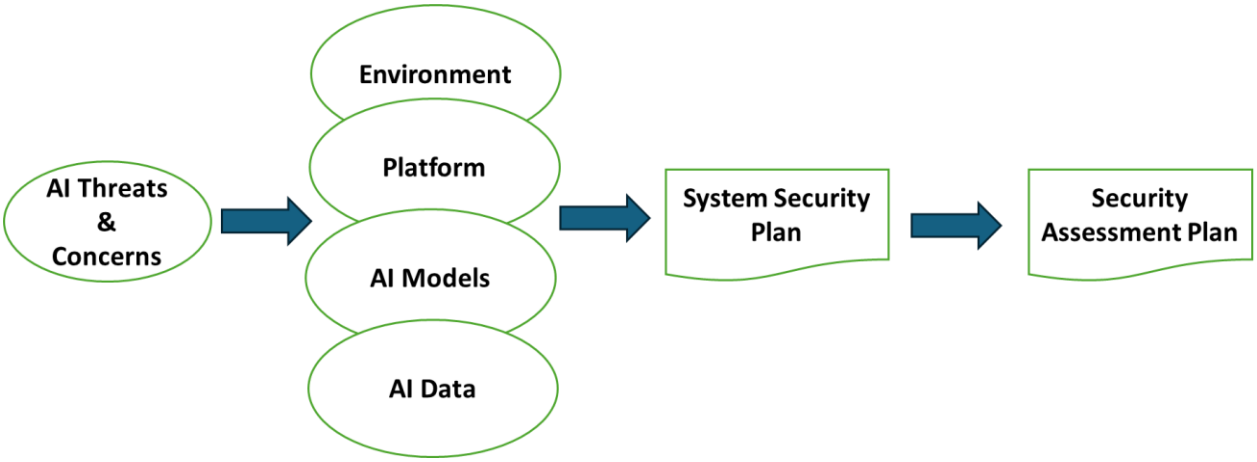

Figure 3 Threat-informed Assessment

When SCA teams prepare for conducting assessments of applications with AI technology, the assessors should think critically about each control in the context of AI concerns and in view of the system elements: Environment, Platform, AI Models, and AI Data. When applied to a specific SSP, it should inform the security assessment plan (see [5]).

The SAFE-AI framework maps each AI control to a set of questions, organized by system elements, that should be raised during security assessment planning to ensure that the risks posed by the associated AI concerns can be addressed and for determining necessary mitigation measures.

APPENDIX A. ATLAS™ FRAMEWORK WITH SAFE-AI MAPPING

This table depicts the entire ATLAS™ Framework with SAFE-AI mappings shown in blue highlighted cells. Cells containing a red ampersand (&) indicate techniques common with the MITRE ATT&CK framework.

Table A-1: ATLAS™ Framework with SAFE-AI Mapping

Reconnaissance	Resource Development	Initial Access	ML Model Access	Execution	Persistence	Privilege Escalation	Defense Evasion	Credential Access	Discovery	Collection	ML Attack Staging	Exfiltration	Impact
AML.TA0002	AML.TA0003	AML.TA0004	AML.TA0000	AML.TA0005	AML.TA0006	AML.TA0012	AML.TA0007	AML.TA0013	AML.TA0008	AML.TA0009	AML.TA0001	AML.TA0010	AML.TA0011
AML.T0006 & Active Scanning	AML.T0008 Acquire Infrastructure	AML.T0015 Evade ML Model	AML.T0040 AI Model Inference API Access	AML.T0050 & Command and Scripting Interpreter	AML.T0018 Backdoor ML Model	AML.T0054 LLM Jailbreak	AML.T0015 Evade ML Model	AML.T0055 & Unsecured Credentials	AML.T0063 Discover AI Model Outputs	AML.T0036 & Data from Information Repositories	AML.T0018 Backdoor ML Model	AML.T0025 Exfiltration via Cyber Means	AML.T0034 Cost Harvesting
AML.T0004 Search Application Repositories	AML.T0002 Acquire Public ML Artifacts	AML.T0049 & Exploit Public-Facing Application	AML.T0044 Full ML Model Access	AML.T0053 LLM Plugin Compromise	AML.T0051 LLM Prompt Injection	AML.T0053 LLM Plugin Compromise	AML.T0054 LLM Jailbreak		AML.T0062 Discover LLM Hallucinations	AML.T0037 & Data from Local System	AML.T0043 Craft Adversarial Data	AML.T0024 Exfiltration via ML Inference API	AML.T0029 Denial of ML Service
AML.T0001 Search for Publicly Available Adversarial Vulnerability Analysis	AML.T0017 & Develop Capabilities	AML.T0051 LLM Prompt Injection	AML.T0047 ML-Enabled Product or Service	AML.T0011 & User Execution	AML.T0061 LLM Prompt Self-Replication	AML.T0051 LLM Prompt Injection	AML.T0051 LLM Prompt Injection		AML.T0007 Discover ML Artifacts	AML.T0035 ML Artifact Collection	AML.T0005 Create Proxy ML Model	AML.T0057 LLM Data Leakage	AML.T0059 Erode Dataset Integrity
AML.T0000 Search for Victim's Publicly Available Research Materials	AML.T0021 & Establish Accounts	AML.T0010 ML Supply Chain Compromise	AML.T0041 Physical Environment Access		AML.T0020 Poison Training Data				AML.T0014 Discover ML Model Family		AML.T0042 Verify Attack	AML.T0056 LLM Meta Prompt Extraction	AML.T0031 Erode ML Model Integrity
AML.T0003 Search Victim-Owned Websites	AML.T0016 & Obtain Capabilities	AML.T0052 & Phishing							AML.T0013 Discover ML Model Ontology				AML.T0015 Evade ML Model
	AML.T0020 Poison Training Data	AML.T0012 & Valid Accounts							AML.T0056 LLM Meta Prompt Extraction				AML.T0048 External Harms
	AML.T0060 Publish Hallucinated Entities												AML.T0046 Spamming ML System with Chaff Data
	AML.T0019 Publish Poisoned Datasets												
	AML.T0058 Publish Poisoned Models												

APPENDIX B. CONVENTIONS

Control nomenclature

This paper uses a canonical **Control Identifier (ID)** that has been modified to conform with the NIST SP 800-53 Rev 5.1.1 patch and the NIST SP 800-53B, which removes parentheses from the identifier, so that sorting and other functions can be accomplished more easily. Zeroes are used to left pad the Control ID so that two numeric digits are used for the Control Identifier and two numeric digits used for the Control Enhancement number, are in the form of XX-nn-nn.

For example, AC-2 (13) is written as AC-02-13.

APPENDIX C. AI THREATS, CONCERNS, AND RESIDUAL RISK

This table shows a list of AI Threats and relevant concerns and a description of the potential residual risk for each threat. Related ATLAS™ identifiers for each threat are provided for reference.

Table 1: Threats and Concerns

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
Loss of models	The models used in an AI system are key components enabling system functionality. Malicious destruction or corruption of a model is therefore a critical AI concern. All potential vulnerabilities an attacker could exploit to gain access to a system and its models need to be anticipated, including outdated or unpatched software components, weak or improperly enforced access control, and poor asset protection management practices. The key consideration for avoiding model loss is access control in general and write access in particular.	AC-03-00, AC-06-00, CM-07-00, SC-37-00	AC-03-00, AC-05-00, AC-06-00, AU-02-00, CM-05-00	AC-03-00, AC-05-00, AC-06-00, AU-02-00, AU-03-00, CM-05-00, CM-07-00, SC-24-00, SI-20-00	AC-06-00	Risks from insider threats are not addressed by mitigations focused on access control. In addition, model corruption or tampering could occur undetected. See the "Insider Threat" AI Concerns elsewhere in this sheet for additional controls to consider.	AML.T0031 "Erode Model Integrity"
Model poisoning	AI-enabled systems may be vulnerable to attacks that perturb AI model inputs, or modify AI models to undermine their reliability, integrity, and availability. A wide variety of model poisoning attacks are possible - such as making changes to the code, objective functions, model parameters, or training data - so the attack surface is potentially very large. Mitigations include controlling access to models and data, continual/continuous testing, and establishing baselines for data distributions and model performance.			SR-03-00		Access controls can reduce but not eliminate the risk of insider threats. See the "Insider Threat" AI Concerns elsewhere in this sheet for additional controls to consider. Since the potential attack surface for poisoning attacks is so large and is not completely known, attacks may be undetected even with vigilant monitoring and testing.	AML.T0020 "Poison Training Data"

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
Insecure APIs	Insecure APIs can allow attackers unauthorized access, introduce malicious inputs, or disrupt AI systems. This includes risks like unauthorized data access and denial of service, as well as AI-specific threats such as manipulation of model inputs. AI systems often consist of both internal and externally facing APIs that need to be secured, so that data integrity is preserved as data is transmitted among the various components in the AI-enabled system. Standard mitigation strategies, such as data encryption, input validation, robust authentication and authorization mechanisms, are essential for ensuring security of both internal APIs and externally facing APIs. **Inference APIs are particularly vulnerable** as they are often exposed to external users. Adversaries may exploit legitimate access to inference APIs to gather detailed information about model ontology, structure, and behavior, enabling black-box and white-box attacks. Attackers can refine adversarial techniques to bypass model defenses and evade detection capabilities to introduce malicious data, potentially leading to incorrect predictions or compromised decision-making.	RA-05-00, SC-05-00, SC-23-00, SR-09-00	AC-24-00, SR-03-00, SR-11-00	SR-03-00		Authorized users may abuse their access privileges to compromise the security of an API. Open-source reconnaissance is difficult to prevent, so the security controls will not eliminate all risks. In particular, the risk posed by opensource information about APIs, particularly publicly available services, gives adversaries the opportunity to search for new zero-day exploits of a publicly available AI model. See the "Zero-day exploits", "Insider Threat", and "Vulnerability exploit" AI Concerns for additional controls to consider.	AML.T0040 "AI Model Inference API Access"
Configuration errors	As with all software systems, inadequate attention to configuration management issues can leave an AI-enabled system vulnerable to adversarial attacks. A particular AI concern is that AI-enabled systems are often designed as integrated systems comprised of several interacting components, each of which has its own configuration management requirements. Managing the interaction of all these configuration requirements can sometimes be a challenge. When configuration errors go unnoticed, a misconfigured component could provide the point of entry for attacks that poison data, perturb model inputs, or modify AI models to undermine their reliability, integrity and availability. Mitigations include vigilant attention to configuration management policies and practices for the AI-	CA-09-00, CM-03-00, CM-05-00, CM-07-00	CM-03-00, CM-05-00, CM-07-00, SA-10-00	CM-05-00, CM-07-00		Configuration errors can go unnoticed, especially when the AI-enabled system includes interacting subsystems that may rely on third-party or open-source components.	AML.T0006 “Active Scanning”

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
	enabled system and its components, as well as the infrastructure (e.g., host system or cloud service components like storage, computing resources, or databases) the system uses during operations.						
Data poisoning	Poisoned data can compromise the decision making of an AI-enabled system and bias its outputs. An adversary can poison data by compromising external data datasets, or by gaining access to the system and poisoning data stored for training, testing, or other operations. This is an important concern for AI because data poisoning attacks can embed vulnerabilities into an AI-enabled system that may be difficult to detect. For example, an adversary may embed a backdoor trigger that gets activated by designated input data during operations and generates the adversary's desired output rather than the correct response. Mitigations include preprocessing all data to sanitize and validate it before it is used, and continual/continuous testing	SC-07-00, SC-08-00		SC-08-00	AC-14-00, CM-07-00, SC-08-00, SI-04-00, SI-10-00	Data used as a "ground truth" baseline for validation could unknowingly be incomplete or unrepresentative of data the system encounters during operations, making preprocessing mitigations ineffective. See the "AI bias" AI Concerns elsewhere in this sheet for additional concerns to consider.	AML.T0020 "Poison Training Data"
Model exposure	Attackers may try to gain knowledge about the models in an AI-enabled system to steal intellectual property, enable unauthorized use of model capabilities, or achieve some competitive advantage. Attackers may exploit a variety of attack vectors to gain access to models, such as coding errors and software vulnerabilities in the system, weak access controls, and poor protection management practices for AI assets, to extract a trained AI model directly, or collect enough information about the model architecture to create a functionally equivalent copy of the model. Consequently, it is often prudent to treat models in AI-enabled systems as sensitive assets requiring protection and controlled access. Mitigations include stringent access controls (especially to prevent data exfiltration), and strong asset protection management practices.	AC-03-00, AC-06-00, AC-20-00, AC-24-00, CM-07-00, SC-04-00, SC-08-00, SC-28-00, SC-39-00	AC-03-00, AC-06-00, AC-20-00, AC-24-00, AU-02-00, CM-05-00, SC-04-00, SC-08-00, SC-39-00	AC-03-00, AC-05-00, AC-06-00, AC-20-00, AU-02-00, AU-03-00, CM-05-00, CM-07-00, SC-04-00, SC-12-00, SC-28-00, SI-20-00	AC-03-00, AC-06-00, AC-20-00, SC-04-00, SC-08-00, SC-28-00	Knowledge about model function could be disclosed indirectly through other means like public documents (such as academic papers). Sensitive information about the model may also be exposed inadvertently through authorized channels during routine use.	AML.T0024 "Exfiltration via ML Inference API"

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
Sensitive data exposure	Sensitive data must be safeguarded during the development, testing, and deployment of an AI-enabled system. When attackers obtain unauthorized access to sensitive data, the resulting privacy breach and data exposure can compromise the confidentiality of the data, in addition to facilitating data poisoning attacks that may be more effective when informed by privileged information. Countermeasures include implementing strong access controls, using secure and encrypted data storage, regularly updating and patching the AI models, and using data anonymization techniques.	PM-12-00, SC-04-00, SC-08-00, SC-28-00	PM-12-00, SA-17-00, SC-04-00, SC-08-00, SC-28-00	SC-04-00, SC-08-00, SC-28-00	SC-04-00, SC-08-00, SC-13-00, SC-28-00	Risks from insider threats are not addressed by mitigations focused on access control. The widespread use of third-party and open-source components to handle data in AI-enabled systems may expose sensitive data to external sources of risk that may be hard to identify and track. See the "Insider Threat" and "Lack of system, firmware, or tool updates and patches" AI Concerns for additional controls to consider.	AML.T0048 "External Harms"
Sensitive information disclosure	There are many ways AI applications can inadvertently disclose sensitive information, proprietary algorithms, or confidential data. For example, sensitive data may not be adequately filtered from AI responses, AI might memorize sensitive details during training, or there may be unintended data leaks due to misinterpretation of a query. In addition, adversaries may craft prompts that induce the AI to leak sensitive information from proprietary training data, data sources the AI component is connected to, or information from other users of the AI component. Disclosures like this can lead to unauthorized access, intellectual property theft, and privacy breaches. To mitigate these risks, AI applications should employ data sanitization, implement appropriate usage policies, and restrict the types of data returned by the AI component.	AC-04-00, AC-04-25, AC-06-00, AC-21-00, AC-24-00, PL-08-00, PM-07-00, PM-18-00	AC-04-00, AC-04-25, AC-06-00, AC-21-00, AC-23-00, AC-24-00, AU-06-00, SC-28-00, SI-07-00	AC-04-00, AC-04-25, AC-06-00, SC-04-00, SC-08-00, SC-28-00, SI-07-00, SI-20-00	AC-04-00, AC-04-25, AC-06-00, AC-21-00, SC-04-00, SC-08-00, SC-28-00, SI-07-00, SI-20-00	Given the complexity of typical AI-enabled systems, it can be difficult to identify and address all pathways that might be vulnerable to sensitive information disclosure. Security controls will mitigate the risks to some extent, but some degree of continuous monitoring will be needed to address the residual risks.	AML.T0048 "External Harms"

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
Supply chain and life cycle infiltrations and unvetted changes to the model (especially open source)	AI-enabled systems often incorporate pre-trained models obtained from external sources. Failure to assure that these models are securely sourced from external suppliers will enable attacks that insert malicious code into the system that can compromise the AI system's security and integrity. It is also important to scrutinize any updates or changes to these models, since the impact of a revised model on system behavior may not be immediately obvious. This makes change management and configuration management critically important for models. Continual/continuous testing may be needed to detect unexpected changes over time. It may also be important to establish baselines and understanding of operational data and its drift.		SR-01-00, SR-03-00, SR-04-00, SR-05-00, SR-06-00, SR-08-00, SR-11-00	SR-01-00, SR-02-00, SR-03-00, SR-04-00, SR-05-00, SR-06-00, SR-08-00, PM-30-00	SR-01-00, SR-02-00, SR-03-00, SR-06-00, SR-08-00	There is always the possibility that a trusted external source has been unknowingly compromised, which would make efforts to securely source models from suppliers ineffective at mitigating risk. Additional controls should be considered to address any concerns about residual risk associated with change management and configuration management (See AI Concerns associated with "Lack of system, firmware, or tool updates and patches" and "Errors in Configurations"). Note that some change and configuration management methods may be limited by the source and hosting of the model.	AML.T0010.003 "ML Supply Chain Compromise: Model"

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
Supply chain and life cycle infiltrations and unvetted changes of training and operational data	AI-enabled systems often rely on data obtained from external sources. Failure to assure that these data resources are securely sourced from external suppliers will enable attacks that insert malicious code into the system that can compromise the AI system's security and integrity. If the provenance of the data from external sources is not carefully documented (e.g., origins, transformations, dependencies, metadata, etc.), it may be challenging to recognize changes that render the data unsuitable for the current system. Mitigations include stringent vetting of AI model training, use of operational data, and thorough documentation of the data provenance.		SR-01-00, SR-04-00, SR-09-00	AT-03-00, SR-01-00, SR-04-00, SR-05-00	AT-03-00, SR-01-00, SR-04-00, SR-05-00	The size and scope of the data supply chain may be so large that the indicated controls will not sufficiently detect or mitigate all threats. There is always the possibility that a trusted external data source has been unknowingly compromised, which would make efforts to securely source data from suppliers ineffective at mitigating risk. Additional controls should be considered to address any concerns about residual risk associated with change management and configuration management. See the "Data Poisoning", "Backdoor and malware insertion" and "Errors in Configurations" AI Concerns for additional controls to consider.	AML.T0010.002 "ML Supply Chain Compromise: Data"
Supply chain infiltrations/ unvetted changes of AI tools/platforms (especially open source)	AI-enabled systems are often built using tools and platforms obtained from external sources. If one of these resources has vulnerabilities because it is outdated, unpatched or compromised, it could provide a point of entry for attacks that poison data, perturb model inputs, or modify AI models to undermine their reliability, integrity and availability. Similar concerns arise if the external resources are not securely sourced from external suppliers. The vulnerabilities of the tools and platforms from external sources must be carefully understood and managed. Otherwise, it may be challenging to identify the underlying cause of any adverse outcomes in the AI system. Mitigations include meticulous attention to the preparation and maintenance of relevant risk assessment documents such as software bills of materials (SBOMs), AI system bills of materials (AIBOMs), data cards, and model cards	SR-03-00	SR-03-00, SR-04-00, SR-05-00, SR-11-00			The size and scope of the supply chains may be so large that the indicated controls will not sufficiently detect or mitigate all threats. There is always the possibility that a trusted external source of AI tools/platforms has been unknowingly compromised, which would make efforts to securely source those components from suppliers ineffective at mitigating risk. Additional controls should be considered to address any concerns about residual risk associated with change management and configuration management. See the "Lack of system, firmware, or tool updates	AML.T0010.001 "ML Supply Chain Compromise: ML Software"

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
						and patches" and "Supply chain and life cycle infiltrations and unvetted changes to the model" AI Concerns for additional controls to consider.	
Supply Chain infiltration/ unvetted changes of Environment components (especially open source)	Since AI is such a rapidly evolving technical area, AI-enabled systems tend to incorporate open-source components, or capabilities provided by external suppliers. Adversaries can sometimes gain initial access to a system by infiltrating and compromising targeted portions of the AI supply chain. This could include specialized hardware like GPUs, software stacks for AI code development, or pre-trained AI models. Failure to assure that AI-related components are securely sourced from external suppliers will enable attacks that insert malicious code into the system that can compromise the AI system's security and integrity. Moreover, it is especially important to scrutinize all updates and patches needed for any third party or open-source capabilities that may be integrated into an AI component. This implies that relevant risk assessment documents must be carefully prepared, such as software bills of materials (SBOMs), AI system bills of materials (AIBOMs), data cards, and model cards.	SR-03-00, SR-04-00, SR-09-00	SR-03-00, SR-09-00			The size and scope of the supply chain is so large that it is likely that controls will not sufficiently mitigate threats. In addition, SBOMs are not in widespread use, limiting their usefulness as a control mechanism. Finally, it is unclear if SBOMs will mitigate adversaries that target the means of production (as in the Solar Winds attack of 2023).	AML.T0010 "ML Supply Chain Compromise"
Loss of data	AI-enabled systems have a strong dependence and reliance on data during all phases of the lifecycle. Malicious destruction or corruption of data is therefore a critical AI concern. All potential vulnerabilities an attacker can exploit to gain access to a system and its data need to be anticipated, including outdated or unpatched software components, weak or improperly implemented access controls, and poor asset protection management practices. Key considerations for avoiding data loss include access controls in general (and write access in particular), along with careful assessment of backup and recovery capacity to avoid any backup capability limitations.		PL-02-00, SI-04-00	SC-28-00	SC-28-00	Risks from insider threats are not addressed by mitigations focused on access control. In addition, data corruption or tampering could occur undetected. See the "Insider Threat" AI Concerns elsewhere in this sheet for additional controls to consider.	AML.T0059 "Erode Dataset Integrity"

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
Unauthorized access to data	Unauthorized access to data is a concern that is important across all aspects of an AI-enabled system (including AI platforms, tools, and models) and all phases of the system lifecycle. The associated security risks include data breaches, manipulation of AI models, exposure of sensitive information, and potential misuse of AI systems for malicious purposes. Preventive measures include implementing strong access controls, using encrypted data storage, regularly updating and patching the AI models, and using AI-powered security tools for threat detection and response. Regular audits and security assessments can also help identify potential vulnerabilities. It may also be useful to recognize an expanded set of roles for the purposes of access control in AI-enabled systems (e.g. roles for prompt templating and model generation).	AC-01-00, SC-02-00, SC-03-00	AC-01-00, AC-03-00, AC-06-00, SC-02-00, SC-03-00, SC-10-00	AC-06-00	AC-06-00, SC-17-00	Risks from insider threats are not addressed by mitigations focused on access control. The widespread use of third-party and open-source components in AI-enabled systems can leave these systems vulnerable to many external sources of risk that may be hard to identify and track. See the "Insider Threat" and "Lack of system, firmware, or tool updates and patches" AI Concerns for additional controls to consider.	AML.T0012 "Valid Accounts" AML.T0055 "Unsecured Credentials"
Unauthorized access to environment, platform/tool	Malicious actors can exploit unauthorized access to perturb valid inputs to AI models, causing them to consistently generate incorrect decisions. If safeguards are not in place to validate inputs, AI-enabled systems may be vulnerable to attacks that inject instructions or commands to an AI model, causing it to execute unauthorized tasks or generate erroneous outputs. Also, be wary of "off-label use" where an AI component is developed outside of an organization's security safeguards, or a component has been lifted from one context or application and then "fine-tuned" to be used in a different setting.	AC-03-00, AC-06-00, AC-24-00, SC-37-00	AC-03-00, AC-06-00, AC-24-00, SC-23-00, SC-37-00			Risks from insider threats are not addressed by mitigations focused on access control. Data used as a "ground truth" baseline to validate inputs could unknowingly be incomplete or unrepresentative of data the system encounters during operations, making some safeguards for validating inputs ineffective. See the "Unauthorized access to data" and "Faulty authentication and authorization settings" AI Concerns for additional controls to consider.	AML.T0041 "Physical Environment Access" AML.T0047 "ML-Enabled Product or Service"
Insecure deserialization – embedding and executing remote unapproved code or other	Malicious users can sometimes exert control over an AI-enabled system by finding a command sequence that abuses the logic of the system and causes it to execute unauthorized tasks or generate incorrect behavior. The most prominent recent examples of this are the prompt attacks that some large language models are vulnerable to. Additionally, some AI models use procedural representations of knowledge and models (e.g., as in a rule-based system). These		SC-05-00, SI-10-00	SI-10-00	SI-10-00	The logic in some AI-enabled systems (such as those using neural networks) does not lend itself to the kind of transparency and scrutiny available for traditional software systems. Consequently, it is possible that unknown flaws in that logic may be exploited by	AML.T0050 "Command and Scripting Interpreter"

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
malicious activities	representations might be compromised by an adversary to enable the execution of malicious code. All procedural data and input data must be rigorously sanitized and validated.					procedural data or input data even if that data has been carefully sanitized and validated.	
Backdoor and malware insertion	Given the tendency of AI-enabled systems to depend on massive data stores and somewhat complex models and decision logic, it can be difficult identify all unsecured points of entry vulnerable to backdoor and malware insertion attacks. Attackers can manipulate data in all phases of the system lifecycle, exploit vulnerabilities in AI algorithms and models, or use a variety of other techniques to insert backdoors into AI systems that get triggered once the AI is deployed. It is important to anticipate potential threats and AI-related attack surfaces during the design phase, secure and verify data and software during development, and establish testing procedures to regularly monitor system components for data/model drift, changes in performance, or other AI system behavior issues once it is deployed. If a potential attack is detected, information about errors and attack patterns must be shared with incident databases.	CA-08-00, IA-03-00, RA-05-00, SC-04-00, SC-23-00, SC-28-00, SC-37-00, SI-03-00, SI-04-00, SI-05-00, SI-07-00, SI-16-00, SR-05-00, SR-09-00	AC-17-00, AU-02-00, CA-08-00, RA-05-00, SC-18-00, SC-23-00, SI-03-00, SI-04-00, SI-05-00, SI-16-00, SR-05-00, SR-09-00	SC-28-00, SI-03-00, SI-04-00, SI-07-00	SC-28-00, SI-03-00, SI-04-00, SI-07-00	Unsecured points of entry may be hidden in third-party and open-source software components. Mitigations that depend on securing data and software may not adequately address risks from insider threats. See the "Insider threat" and "Supply chain and life cycle infiltrations and unvetted changes of training and operational data" AI Concerns for additional controls to consider.	AML.T0018 "Backdoor ML Model"
Vulnerability exploit	Code development and testing practices for AI-enabled systems do not always conform to traditional software development practices. Consequently, assessing AI system vulnerabilities may raise unexpected challenges (e.g., in some cases it may be difficult to even identify what to test). These challenges are of course a prime opportunity for attackers to exploit gaps in the vulnerability assessment and initiate attacks. Avoiding these undesirable outcomes requires stringent approaches to vulnerability assessment and monitoring. All known potential threats, vulnerabilities, and attack vectors associated with an AI-enabled system must be identified early during the design phase (e.g., by using ATLAS) and the risks must be managed. It is critical to define metrics and procedures for detecting, tracking, and measuring known risks, errors, incidents, or negative impacts. Metrics should also account for known AI design and implementation failure modes associated with properties	AC-20-00, CA-02-00, CA-08-00, IA-06-00, RA-03-00, RA-05-00, SC-08-00, SI-02-00, SI-03-00	AC-20-00, CA-02-00, CA-08-00, RA-03-00, RA-05-00, SC-08-00, SI-02-00, SI-03-00	AC-20-00	AC-20-00	Some of the potential threats, vulnerabilities, and attack vectors associated with an AI-enabled system may be unknown during the design and implementation phases. Existing tests for known vulnerabilities may be inadequate. See the "Zero-day exploits" AI concerns for additional controls to consider.	AML.T0011 “User Execution”

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
	like brittleness. The deployed AI-enabled system must be continuously tested for errors or vulnerabilities						
Lack of system, firmware, or tool updates and patches	Failure to apply patches and updates to AI-enabled systems is just as problematic as it is for any software system. Attackers can exploit unpatched vulnerabilities to compromise system integrity and gain access to sensitive information. When it comes to AI-enabled systems, though, an outdated or unpatched component could provide a point of entry for attacks that poison data, perturb model inputs, or modify AI models to undermine their reliability, integrity and availability. Note that it is particularly important to be aware of updates and patches needed for any third party or open-source capabilities that may be integrated into an AI component. This implies that relevant risk assessment documents have been reviewed, such as software bills of materials (SBOMs), AI system bills of materials (AIBOMs), data cards, and model cards.	CM-07-00, CM-11-00, CM-14-00, MA-03-00, MA-06-00, RA-05-00, SA-22-00, SI-02-00	CM-07-00, CM-11-00, CM-14-00, MA-03-00, MA-06-00, RA-05-00, SA-22-00, SI-02-00			There is always the risk that updates and patches have been compromised. In addition, SBOMs are not in widespread use, limiting their usefulness as a control mechanism. See the "Sensitive data exposure" and "Supply chain and life cycle infiltrations and unvetted changes to the model" AI Concerns for additional controls to consider.	AML.T0001 "Search for Publicly Available Adversarial Vulnerability Analysis"
Faulty authentication and authorization settings	Weak or improperly implemented authentication and authorization mechanisms can allow attackers to poison data, manipulate model input, or otherwise compromise the behavior of an AI component. Note that in AI-enabled systems, faulty settings for these mechanisms may be the result of deliberate attacks like model stealing and prompt extraction, or inadequate attention to data privacy during model development, testing, and deployment.	AC-03-00, AC-14-00	AC-03-00, AC-14-00			The security controls listed here do not eliminate the risk of having an insider maliciously compromise authentication and authorization settings. See the "Insider Threat" AI Concerns elsewhere in this sheet for additional controls to consider.	AML.T0055 "Unsecured Credentials"
Zero-day exploits	AI-enabled systems typically have failure modes that can be difficult to characterize, and those modes and their causes can often be poorly understood (or even unknown). For this reason, it is critically important to continuously monitor performance once a system is deployed and proactively investigate reports of anomalous events (using, for example, red team exercises to discover and assess failure modes). Information sharing is a key component of this monitoring activity. Information about errors and other potential precursors to security abuses must be shared with incident	CA-08-00, SI-02-00, SI-03-00	CA-08-00, SI-02-00, SI-03-00	SI-20-00	SI-20-00	Given the prevalence of poorly understood (or unknown) failure modes in AI-enabled systems, none of the mitigations listed here can eliminate the possibility of zero-day exploits	AML.T0001 "Search for Publicly Available Adversarial Vulnerability Analysis", AML.T0006 "Active Scanning"

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
	databases, other organizations with similar systems, and system users and stakeholders.						
Insider threats	Insiders can exploit access privileges to engage in unauthorized activities, including data theft or sabotage of AI models and data. These insider attacks may be especially difficult to address for AI-enabled systems, since AI system development and documentation practices do not tend to employ the same process controls as traditional software development. Moreover, AI systems may require more frequent maintenance and triggers for conducting corrective maintenance due to factors like data, model, or concept drift. This points to the need for extra vigilance when it comes to things like data provenance and access control mechanisms and policies.	AC-05-00, AC-06-00, AC-24-00, CM-11-00, IA-02-00, IA-08-00, MA-05-00, PM-12-00, SC-28-00, SI-03-00, SI-04-00, SR-09-00	AC-05-00, AC-06-00, AC-24-00, CM-11-00, IA-02-00, IA-08-00, MA-05-00, PM-12-00, SC-28-00, SI-03-00, SI-04-00, SR-09-00	PM-12-00, SC-28-00, SI-04-00, SI-20-00	PM-12-00, SC-28-00, SI-04-00, SI-20-00	These security controls can make it more difficult for insiders to engage in unauthorized activities, and make it easier to identify unauthorized activities, but they cannot completely eliminate the risk.	AML.T0012 "Valid Accounts"
Backup capability limitations	AI-enabled systems have a strong dependence and reliance on data during all phases of the lifecycle. Backup capability limitations that may weaken safeguards against data loss or data corruption are therefore an important AI concern. Several issues make it challenging to provide suitable backup capabilities for an AI-enabled system: the data volumes associated with AI systems are massive and far beyond those for other software systems; complex AI-enabled computations may produce data usage patterns that change drastically during routine operations; and data storage requirements may differ in the different phases of the AI lifecycle. Mitigations include careful assessment of backup and recovery capacity requirements for each stage of the system lifecycle, along with specific backup policies to address the key data usage patterns.	CP-01-00, CP-09-00	CP-01-00, CP-09-00	CP-09-00	CP-09-00	There is always the possibility of a backup recovery failure. It may be prudent to consider implementing redundant systems and using cloud-based solutions to enhance backup capabilities for AI-enabled systems.	T1490 "Inhibit System Recovery"

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
Denial of Service attack	Many AI-enabled systems require significant amounts of specialized computing resources. From an adversary's perspective, these computing resources can often be viewed as expensive bottlenecks that can be easily overloaded. Adversaries can exploit this vulnerability by flooding the system with inputs, or by intentionally crafting inputs that require heavy amounts of useless compute from the AI system. The increased computing load can eventually degrade or shut down the services supplied by the AI-enabled system. Mitigations include limiting the number of queries the AI system will handle at any one time and putting in place explicit monitors to detect adversarial input.	SC-05-00, SC-37-00	SR-03-00, SR-11-00	SR-03-00		The risk of denial of service is somewhat mitigated by SC-05. However, there will remain some risk of cost burden or service quality degradation that cannot be compensated for.	AML.T0034 "Cost Harvesting"
Network components attacks	The AI components in an AI-enabled system are often general-purpose capabilities that are customized for the needs of the system and its use cases. This means, in particular, that AI-enabled systems often do not have an adequately designed resilient security architecture. There are likely to be shortcomings regarding access controls and proper network configurations. Gaps in these capabilities need to be proactively identified and mitigated during the design, development and deployment phases of the AI lifecycle.	AC-07-00, AC-17-00, AU-02-00, CA-08-00, SC-37-00	CA-08-00, SC-15-00			Since AI-enabled systems tend to be integrated systems that include many complex interactions among components, controls may not sufficiently mitigate all vulnerabilities in access controls, APIs and network configurations.	AML.T0049 "Exploit Public Facing Application"
Power supply attacks	Power supply attacks are problematic for AI-enabled systems that need massive amounts of time and computing resources to process large volumes of complex data during some phase of the AI lifecycle. For example, this is a routine concern for AI pipelines that train modern deep learning systems such as large language models. While the AI architectures supporting these big data requirements are explicitly designed to provide safeguards like checkpoint and restore operations, the I/O bandwidth and storage needed to address these concerns are formidable. The AI concern here is arranging for the massive amount of resources needed for the safeguards, above and beyond what is needed to build and use the AI-enabled system itself.	PE-11-00				With proper power planning and contingency planning, there should be little residual risk.	T1584 " Compromise Infrastructure ", AML.T0041 "Physical Environment Access"

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
Physical breaches	For AI-enabled systems, exploitation of the physical environment to attack the system can occur in a variety of ways. For example, an attacker may physically access the location where data is being collected and modify the collection process in ways that will comprise subsequent AI model training or performance. When the AI system receives input data from real world sensors, it may be possible to employ attacks that make malicious changes to the physical environment that will compromise system behavior (e.g., using physical domain patch-based attack to deceive a ML classification model). Mitigations include stringent approaches to detect physical tampering and unexpected vulnerabilities in AI components. Common anti-tamper technologies for software systems should be applied to AI-enabled systems if a physical breach is suspected. Since behavior patterns in AI components can be difficult to specify precisely, it may also be helpful to establish a variety of behavior baselines for AI components. Given a baseline of normal behavior, behavior analysis techniques could identify anomalous behavior patterns that might be useful indicators of tampering.	SR-09-00				Controlling access to the host environment is generally easier to do than controlling access to sensors or IoT (internet of things) where the devices are in the field and may be inadvertently or maliciously accessed. In this respect, the residual risk is the same for AI-enabled systems as for non-AI-enabled systems.	AML.T0041 " Physical Environment Access"
AI bias	Biases in the data and models associated with an AI-enabled system can lead to inaccurate outcomes or discriminatory treatment of certain individuals or demographics, with a corresponding negative impact on the trustworthiness of the system. A key underlying issue is that the scale and complexity of many AI-enabled systems can result in high levels of statistical uncertainty and many potential sources of bias, making bias management a challenge. Unintended sources of bias like spurious correlations and unrepresentative data sources may be difficult to avoid. Malicious sources of bias caused by adversarial attacks on data and models may be challenging to detect. Mitigations include careful attention to the quality of data and its statistical properties, stringent access controls for data and models, and vigilant monitoring of system performance.		CA-02-00, CM-02-00, PL-02-00, PL-04-00, SA-10-00	CA-02-00, CM-02-00, PL-02-00, PL-04-00, SA-10-00	CA-02-00, CM-02-00, PL-02-00, PL-04-00, SA-10-00, SR-04-00	Given the myriad of possible input sources to an AI-enabled system, the residual risk is that some of these have not been sufficiently controlled and analyzed to prevent the introduction of biases. There is also the residual risk of "data drift" where statistical properties of the operational inputs change over time. This can cause inaccurate and biased outcomes in the AI model if the original training data is no longer representative of the operational data.	AML.T0020 “Poison Training Data”

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
Identity Spoofing (e.g. deep fakes, synthetic identities, CAPTCHA threat)	The ability to generate new or altered identities using AI has become common place and represents a threat to some forms of identification and authentication (e.g., voice spoofing. keystroke dynamics, and biometrics). This is particularly true in phishing attacks designed to gain sensitive information that may put access control safeguards at risk. AI based systems can also mimic certain human inputs to break known CAPTCHA-types of systems, thereby increasing the vulnerability to compromise and fraud. Mitigations include stringent access controls, robust authentication and authorization mechanisms, and user education regarding the dangers of social engineering attacks like phishing.	AC-07-00, AC-14-00, IA-02-00, IA-02-01, IA-02-02, IA-08-00, IA-12-00	AC-07-00, AC-14-00, IA-02-00, IA-02-01, IA-02-02, IA-08-00, IA-12-00			Generative AI can generate deep fakes and other forms of spoofing - some of which may be detectable via other forms of machine learning. The risk here is that the fake-detectors lag the fake-generators and so there is a period of time where the system is vulnerable.	AML.T0052 "Phishing"
Indirect Prompt Injection	Adversaries may devise malicious prompts that cause the AI component to act in unintended ways. The attacks may be designed to bypass defenses or allow the adversary to issue privileged commands. The attack is indirect when the AI component ingests the malicious prompt from a separate data source (e.g., text or multimedia from a website, chat plugins) as part of its normal operation. Plugins may be vulnerable to an indirect prompt injection attack that uses the AI component to exfiltrate the history of a user conversation with an external website. The user may never be aware of the prompt injection. This type of injection can be used by the adversary to target the PII of the user.		AC-06-00, AU-06-00, CM-05-00, SI-03-00, SI-04-00, SI-10-00			Prompts may be injected from any uncontrolled data source, so there is a limit to how effective the controls can be. Moreover, the logic underlying how the AI component responds to a prompt does not lend itself to the kind of transparency and scrutiny available in traditional software systems. Consequently, it is possible that unknown flaws in that logic may be exploited by prompts that do not appear to be malicious.	AML.T0051 "Prompt Injection"
Direct Prompt Injection	Adversaries may devise malicious prompts to an AI component causing it to act in unintended ways. Direct prompt injections are often an attempt to manipulate the AI component to generate harmful content or issue privileged commands to gain a foothold on the system, including placing AI component in a state in which it will freely respond to any user input, bypassing controls or guardrails placed on the AI component.	AC-03-00	AC-03-00, SI-03-00, SI-04-00, SI-10-00			Prompts may be injected from any uncontrolled data source, so there is a limit to how effective the controls can be. Moreover, the logic underlying how the AI component responds to a prompt does not lend itself to the kind of transparency and scrutiny available in traditional software systems. Consequently, it is possible that unknown flaws in that logic may be exploited by	AML.T0051 "Prompt Injection"

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
						prompts that do not appear to be malicious.	
Cost Harvesting	AI services tend to use large amounts of computing resources and consume a great deal of energy during response generation. Adversaries can maliciously increase the cost of running these services by flooding the system with useless queries, or by crafting computationally expensive inputs. For example, systems that rely on massive neural networks may be vulnerable to adversarial data (e.g., "sponge" examples) designed to activate large numbers of nodes in the hidden network layers.	AU-06-05, SC-05-00, SC-06-00	AU-06-05, SC-05-00, SC-06-00			The risk of denial of service is somewhat mitigated by SC-05. However, there will remain some risk of cost burden or service quality that cannot be compensated for.	AML.T0034 "Cost Harvesting"
Excessive Agency	Excessive Agency refers to situations where AI components have access to APIs, plugins, extensions, and tools with capabilities that go beyond what is necessary to support the AI component operations. Excessive permissions, unnecessary functionality, and unchecked authority to act autonomously are all examples of excessive agency that can result in unintended and unacceptable application behaviors with potentially damaging consequences. To mitigate these risks, developers need to limit extension capabilities (functionality, permissions, and autonomy) to only what is absolutely necessary, track user authorization, require human approval for all actions, and implement authorization in downstream systems.	AC-06-00, CM-07-00	AC-05-00, AC-06-00, CM-07-00	CM-07-00		Unfettered access to authorized capabilities may lead to unintended consequences. Effectively this allows for privilege escalation, a general class of risk. The residual risk here is that with AI-enabled systems, it may be difficult to anticipate all the ways in which giving an AI component agency might be problematic.	AML.T0050 "Command and Scripting Interpreter" AML.T0053 "LLM Plugin Compromise" AML.T0011 "User Execution"
Insecure Plugin Design	AI software often extends its functionality by using plugins, extensions, or APIs to connect to other services or resources. Plugins may provide a variety of useful capabilities, such as integrations with other applications, access to public or private data sources, and the ability to execute code. If these plugins are not securely designed (e.g., plugins have insufficient access controls or inadequate input validation), adversaries may exploit their access to the AI software to compromise the plugins with attacks that have harmful consequences like data exfiltration, remote code execution, and privilege escalation. Developers must implement robust security measures for plugins, like strict parameterized	AC-06-00, CM-07-00, SC-08-00	AC-06-00, AC-24-00, CM-05-00, CM-07-00, CM-13-00, SA-08-00, SC-39-00, SI-03-00, SI-10-00			Plugins introduce a variety of risks that are plugin-specific. Consequently, it may be difficult to identify all potential sources of plugin vulnerability. See the "Lack of system, firmware, or tool updates and patches" AI Concerns for additional controls to consider.	AML.T0053 "LLM Plugin Compromise"

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
	inputs and secure access control guidelines, to mitigate this potential vulnerability.						
Content Manipulation	Content manipulation poses a significant threat to Google DocumentAI, particularly in the context of OCR. Malicious actors can intentionally alter documents by changing numbers or formatting to deceive the AI, resulting in errors and misclassifications. Subtle content alterations can result in incorrect data extraction and faulty decision-making, compromising the integrity of the documents processed by the AI. By deceiving AI-enable systems via document content manipulation, attackers can undermine trust in the automated document processing system.		AC-02-12, AC-04-15, SC-24-00, SI-10-00			Carefully crafted malicious input is difficult to prevent. Employing some form of human-in-the loop (HITL) or a monitoring service using a probabilistic loop to spot check for HITL actions can help mitigate the residual risk. See the "Unauthorized access to data" and "Supply chain and life cycle infiltrations and unvetted changes of training and operational data" AI Concerns for additional controls to consider.	AML.T0051 "Prompt Injection"
Evade AI model	Adversaries can craft input data designed to prevent AI models from correctly identifying the contents of the data. For example, an adversary might introduce subtle perturbations that cause the model to misclassify or overlook meaningful information. This technique can be used to evade downstream tasks where machine learning is utilized by exploiting weaknesses in the AI model algorithms. Additionally, the adversary may evade machine learning-based virus/malware detection or network scanning tools towards the goal of a traditional cyber-attack.		AC-02-12, AC-04-15, SI-10-00			Carefully crafted malicious input is difficult to prevent. For example, if forms are sent directly through without an IRS gatekeeper, then adversarial data attacks are possible. See the "Unauthorized access to data" and "Supply chain and life cycle infiltrations and unvetted changes of training and operational data" AI Concerns for additional controls to consider.	AML.T0015 "Evade AI Model"
Publicly-available product or service	Adversaries may research existing open source or other publicly-available implementations of machine learning attacks. Adversaries may target AI-enabled products or services to gain access to the underlying AI model. Adversaries may use the product or service to indirectly access the AI model, potentially revealing details about the model, its algorithms, or its inferences through logs or metadata. This type of access can expose sensitive information about the model's structure, parameters, or	AC-02-12	AC-02-12, SA-09-00		SA-09-05, SA-09-06, SA-09-08	Open-source reconnaissance is difficult to prevent, so the security controls will not eliminate all risks. There is always the possibility that a trusted external product or service has been unknowingly compromised, which would make efforts to securely source the product or service from suppliers	AML.T0001 "Search for Publicly Available Adversarial Vulnerability Analysis"

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
	decision-making processes, and business intelligence. By exploiting these vulnerabilities, adversaries can gain insights that may help them craft more effective adversarial attacks or reverse-engineer the model. The research community often publishes their code for reproducibility and to further future research. Libraries intended for research purposes, such as CleverHans, the Adversarial Robustness Toolbox, and FoolBox, can be weaponized by an adversary.					ineffective at mitigating risk. See the "Supply chain and life cycle infiltrations and unvetted changes to the model (especially open source)" AI Concerns for additional controls to consider.	
Metadata exposure	The threat of metadata exposure involves the temporary logging of metadata about API requests, such as the time received, frequency and size of the request, and the IP addresses from which the requests originated. While this logging aims to improve the service and combat abuse, it could inadvertently reveal patterns or usage information. Adversaries could analyze the metadata to infer sensitive details about document processing activities, operational behaviors, and timelines. Additionally, IP addresses could be exploited to track user locations or launch targeted attacks against specific networks. Although the document content itself is not directly exposed, the metadata could still provide valuable insights to the adversaries attempting to cause harm.		AU-09-00, SA-03-02		AU-09-00, SA-03-02	Metadata exposure may occur in application logs. In an AI-enabled system consisting of several components, it may be difficult to assess the risks associated with combinations of information gleaned from all of the component application logs. See the "Sensitive data exposure" AI Concerns for additional controls to consider.	AML.T0057 “LLM Data Leakage” AML.T0002 “Acquire Public ML Artifacts
Robotic Process Automation Permissions	Robotic Process Automation (RPA) bots are responsible for handling and manipulating sensitive data. If access controls and policies are not properly implemented, the bots can cause damage to systems and data due to errors. RPA bots are vulnerable to adversarial attacks, such as Evasion Attack, where malicious actors manipulate input data to deceive the bots, causing them to perform unintended actions, misclassify data, or corrupt data. Monitoring unauthorized access and insider threats is crucial, as bots with excessive permissions can be misused for malicious purposes. Ensuring robust security measures and continuous monitoring can mitigate these risks and protect sensitive data.	AC-06-00, AC-24-00, AU-02-00, SC-02-00	AC-04-09, AC-24-00, SC-24-00, SI-10-00, SI-11-00			Mitigations focused on access control will not address risks associated with bot compromise, bot manipulation, or bot software built without adequate error handling capabilities. See the "Excessive Agency", "Vulnerability exploit" and "Faulty authentication and authorization settings" AI Concerns for additional controls to consider.	AML.T0050 "Command and Scripting Interpreter"

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
Spamming the System with Chaff Data	Adversaries may spam AI applications with chaff data to flood them with false positives, overwhelming the system and increasing detections. This tactic forces analysts to waste time reviewing and correcting incorrect inferences, reducing their efficiency. Techniques include automated scripts, botnets, or tools like Faker to generate large volumes of synthetic data. By inundating the system with irrelevant data, adversaries aim to degrade performance and exhaust the resources.	AC-02-12	AC-02-12, AC-12-00, SA-08-00, SI-10-00			The risk of this threat is somewhat mitigated by the security controls. However, as with all types of denial-of-service attacks, there will remain some risk of cost burden or service quality degradation that cannot be compensated for. A monitoring service using rate-based detection of chaff activity may be helpful to mitigate residual risk.	AML.T0046 "Spamming ML System with Chaff Data"

APPENDIX D. AI CONTROLS

This table contains a list of 100 NIST SP 800 Rev 5 controls that should be included in SSPs for AI-enabled systems. Modifying control descriptions in the SSP may be warranted for addressing specific weaknesses. The table includes four columns indicating the relation between AI controls and the four system elements of an AI-enabled system as defined here:

- Environment – infrastructure, network
- AI Platform – AI components, software
- AI Models – ML models, LLMs
- AI Data – training data, validation data

Table 2: AI Controls

Control ID	Control Name	Environment	AI Platform	AI Models	AI Data	NIST baseline
AC-01-00	Policy and Procedures	X	X			LMH
AC-02-12	Account Management | Account Monitoring for Atypical Usage	X	X			H
AC-03-00	Access Enforcement	X	X	X	X	LMH
AC-04-00	Information Flow Enforcement	X	X	X	X	MH
AC-04-09	Information Flow Enforcement | Human Reviews		X			
AC-04-15	Information Flow Enforcement | Detection of Unsanctioned Information		X			
AC-04-25	Information Flow Enforcement | Data Sanitization	X	X	X	X	
AC-05-00	Separation of Duties	X	X	X		MH
AC-06-00	Least Privilege	X	X	X	X	MH
AC-07-00	Unsuccessful Logon Attempts	X	X			LMH
AC-12-00	Session Termination		X			MH
AC-14-00	Permitted Actions Without Identification or Authentication	X	X		X	LMH
AC-17-00	Remote Access	X	X			LMH
AC-20-00	Use of External Systems	X	X	X	X	LMH
AC-21-00	Information Sharing	X	X		X	MH
AC-23-00	Data Mining Protection		X			
AC-24-00	Access Control Decisions	X	X			
AT-03-00	Role-based Training			X	X	LMH
AU-02-00	Event Logging	X	X	X		LMH
AU-03-00	Content of Audit Records			X		LMH
AU-06-00	Audit Record Review, Analysis, and Reporting	X	X			LMH
AU-06-05	Audit Record Review, Analysis, and Reporting | Integrated Analysis of Audit Records	X	X			H
AU-09-00	Protection of Audit Information		X		X	LMH
CA-02-00	Control Assessments	X	X	X	X	LMH
CA-08-00	Penetration Testing	X	X			H
CA-09-00	Internal System Connections	X				LMH
CM-02-00	Baseline Configuration		X	X	X	LMH
CM-03-00	Configuration Change Control	X	X			MH

Control ID	Control Name	Environment	AI Platform	AI Models	AI Data	NIST baseline
CM-05-00	Access Restrictions for Change	X	X	X		LMH
CM-07-00	Least Functionality	X	X	X	X	LMH
CM-11-00	User-installed Software	X	X			LMH
CM-13-00	Data Action Mapping		X			
CM-14-00	Signed Components	X	X			
CP-01-00	Policy and Procedures	X	X			LMH
CP-09-00	System Backup	X	X	X	X	LMH
IA-02-00	Identification and Authentication (organizational Users)	X	X			LMH
IA-02-01	Identification and Authentication (organizational Users) | Multi-factor Authentication to Privileged Accounts	X	X			LMH
IA-02-02	Identification and Authentication (organizational Users) | Multi-factor Authentication to Non-privileged Accounts	X	X			LMH
IA-03-00	Device Identification and Authentication	X				MH
IA-06-00	Authentication Feedback	X				LMH
IA-08-00	Identification and Authentication (non-organizational Users)	X	X			LMH
IA-12-00	Identity Proofing	X	X			MH
MA-03-00	Maintenance Tools	X	X			MH
MA-05-00	Maintenance Personnel	X	X			LMH
MA-06-00	Timely Maintenance	X	X			MH
PE-11-00	Emergency Power	X				MH
PL-02-00	System Security and Privacy Plans		X	X	X	LMH
PL-04-00	Rules of Behavior		X	X	X	LMH
PL-08-00	Security and Privacy Architectures	X				MH
PM-07-00	Enterprise Architecture	X				LMH
PM-12-00	Insider Threat Program	X	X	X	X	LMH
PM-18-00	Privacy Program Plan	X				LMH
PM-30-00	Supply Chain Risk Management Strategy			X		LMH
RA-03-00	Risk Assessment	X	X			LMH
RA-05-00	Vulnerability Monitoring and Scanning	X	X			LMH
SA-03-02	System Development Life Cycle | Use of Live or Operational Data		X		X	
SA-08-00	Security and Privacy Engineering Principles		X			LMH
SA-09-00	External System Services		X		X	LMH
SA-09-05	External System Services | Processing, Storage, and Service Location				X	
SA-09-06	External System Services | Organization-controlled Cryptographic Keys				X	
SA-09-08	External System Services | Processing and Storage Location — U.S. Jurisdiction				X	
SA-10-00	Developer Configuration Management		X	X	X	MH
SA-17-00	Developer Security and Privacy Architecture and Design		X			H

Control ID	Control Name	Environment	AI Platform	AI Models	AI Data	NIST baseline
SA-22-00	Unsupported System Components	X	X			LMH
SC-02-00	Separation of System and User Functionality	X	X			MH
SC-03-00	Security Function Isolation	X	X			H
SC-04-00	Information in Shared System Resources	X	X	X	X	MH
SC-05-00	Denial-of-service Protection	X	X			LMH
SC-06-00	Resource Availability	X	X			
SC-07-00	Boundary Protection	X				LMH
SC-08-00	Transmission Confidentiality and Integrity	X	X	X	X	MH
SC-10-00	Network Disconnect		X			MH
SC-12-00	Cryptographic Key Establishment and Management			X		LMH
SC-13-00	Cryptographic Protection				X	LMH
SC-15-00	Collaborative Computing Devices and Applications		X			LMH
SC-17-00	Public Key Infrastructure Certificates				X	MH
SC-18-00	Mobile Code		X			MH
SC-23-00	Session Authenticity	X	X			MH
SC-24-00	Fail in a Known State		X	X		H
SC-28-00	Protection of Information at Rest	X	X	X	X	MH
SC-37-00	Out-of-band Channels	X	X			
SC-39-00	Process Isolation	X	X			LMH
SI-02-00	Flaw Remediation	X	X			LMH
SI-03-00	Malicious Code Protection	X	X	X	X	LMH
SI-04-00	System Monitoring	X	X	X	X	LMH
SI-05-00	Security Alerts, Advisories, and Directives	X	X			LMH
SI-07-00	Software, Firmware, and Information Integrity	X	X	X	X	MH
SI-10-00	Information Input Validation		X	X	X	MH
SI-11-00	Error Handling		X			MH
SI-16-00	Memory Protection	X	X			MH
SI-20-00	Tainting			X	X	
SR-01-00	Policy and Procedures		X	X	X	LMH
SR-02-00	Supply Chain Risk Management Plan			X	X	LMH
SR-03-00	Supply Chain Controls and Processes	X	X	X	X	LMH
SR-04-00	Provenance		X	X		
SR-05-00	Acquisition Strategies, Tools, and Methods	X	X	X	X	LMH
SR-06-00	Supplier Assessments and Reviews		X	X	X	MH
SR-08-00	Notification Agreements		X	X	X	LMH
SR-09-00	Tamper Resistance and Detection	X	X			H
SR-11-00	Component Authenticity		X			LMH

APPENDIX E. ASSESSMENT INTERVIEW QUESTION AND ANSWER SETS

This table is a set of question-and-answer pairs that equips assessors with supplemental assessment criteria for planning and conducting SCAs. Specifically, it contains canned interview questions, along with expected answers, detailing specifically how the control should be implemented for AI systems. Where [XYZ] appears, organizationally defined parameters shall be substituted, such as for agency-specific assignments. In some instances, an information type is shown inside square brackets – e.g., [methods], [processes], to further specify the type of information needed.

Table 3: Assessment Interview Question and Answer Sets

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
AC-01-00	Q. What roles are used to restrict access to the AI environment and which users have them? A. The list of roles is [XYZ], and the users assigned to each role are [XYZ].	Q. What access control policies are specific to AI systems? A. AI-specific access control policies are [XYZ]. Q. What roles are used to restrict access to the AI Models and which users have them? A. The list of roles is [XYZ], and the users assigned to each role are [XYZ].	Q. What roles are used to restrict access to the AI Models and which users have them? A. The list of roles is [XYZ], and the users assigned to each role are [XYZ].	Q. What are the access control policies specific to AI training datasets in Production and lower environments? A. Access control policies specific to AI training datasets in Production and lower environments are [policies].
AC-02-12	Q. How are system accounts monitored for [atypical usage]? A. System accounts are monitored for [atypical usage]. Q. How is the atypical usage of system accounts reported to [personnel or roles]? A. Atypical usage of system accounts is reported to [personnel or roles].	Q. How are platform and application accounts monitored for [atypical usage]? A. System accounts are monitored for [atypical usage]. Q. How is the atypical usage of platform and application accounts reported to [personnel or roles]? A. Atypical usage of platform/application accounts is reported to [personnel or roles].		
AC-03-00	Q. How are authorizations for logical access to information and system resources authorized and enforced throughout the environment? A. Authorized access to the AI environment is enforced by [access control policies] using [access control methods]. Q. In the AI environment, how is access approved to the LLM chat to mitigate direct prompt injections? A. In the AI environment, access is approved via [method] for the LLM chat to mitigate direct prompt injections.	Q. How are authorizations for logical access to information and AI platform resources authorized and enforced? A. Authorized access to AI resources is enforced by [access control policies] using [access control methods]. Q. In the AI platform, how is access approved to the LLM chat to mitigate direct prompt injections? A. In the AI platform, access is approved via [method] for the LLM chat to mitigate direct prompt injections.	Q. How are authorizations for logical access to information within AI models authorized and enforced throughout the environment? A. Authorized access to the AI models is enforced by [access control policies] using [access control methods].	Q. How are authorizations for logical access to information and AI platform resources authorized and enforced? A. Authorized access to AI resources is enforced by [access control policies] using [access control methods]. Q. In the AI data, how is access approved to the LLM chat to mitigate direct prompt injections? A. In the AI data, access is approved via [method] for the LLM chat to mitigate direct prompt injections.

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
AC-04-00	Q. How is information flow controlled to prevent sensitive information from being disclosed to those that do not require access to it? A. Information flow is controlled by [means].	Q. How does the AI platform controls information flow to prevent sensitive information from being disclosed? A. The AI Platform controls information flow by [means].	Q. How does the AI Model controls information flow to prevent sensitive information from being disclosed? A. The AI Model controls information flow by [means].	Q. How is AI Data protected to prevent sensitive information from being disclosed? A. AI Data is protected by [methods] to prevent sensitive information from being disclosed.
AC-04-09		Q. How are human reviews of the AI platform performed with regards to information flow enforcement? A. Human review is enforced for [selected processes] under [conditions].		
AC-04-15		Q. How is the transfer of unsanctioned information between security domains detected and prevented? A. Unsanctioned information is detected by [methods] and its transfer between security domains is prevented in accordance with [security and/or privacy policy].		
AC-04-25	Q. How is information within the environment sanitized to minimize harmful or malicious code? A. Data within the AI environment is sanitized to minimize transfer of [malware/command and control code/ encoded data, sensitive data] according to [policy].	Q. How is data flowing within the AI platform sanitized by malicious code or destroyed when no longer in service? A. Data on AI platforms is [sanitized/destroyed] to minimize transfer of [malware/command and control code/ encoded data, sensitive data] according to [policy].	Q. How is AI model data sanitized to prevent the flow of sensitive data and/or destroyed when no longer in service? A. AI model data is sanitized of sensitive data according to [policy] or destroyed according to [guidelines].	Q. How is AI data sanitized to prevent the flow of sensitive data and/or destroyed when no longer in service? A. AI training data is sanitized of sensitive data according to [policy] or destroyed according to [guidelines].
AC-05-00	Q. What kind of roles/groups are used to limit access to AI environment and which users are members of each of those groups? A. The groups to limit access are [XYZ] and the users in each group are [XYZ].	Q. What kind of roles/groups are used to limit access to AI platform and which users are members of each of those groups? A. The groups to limit access are [XYZ] and the users in each group are [XYZ]. Q. How is the principle of separation of duties applied to the LLM? A. Separation of duties is achieved by means of [XYZ}.	Q. What kind of roles/groups are used to limit access to AI Models and which users are members of each of those groups? A. The groups to limit access are [XYZ] and the users in each group are [XYZ].	

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
AC-06-00	Q. What permissions are assigned to each of the groups/roles that are used to limit access to the AI environment? Do any roles exceed necessary access permissions? A. Access to AI environment is enforced by [using/not using] least privilege access. Users [do/do not] only have access to systems necessary to perform their duties. Permissions for each of the groups are [XYZ]. Q. What permissions are assigned to each user/group/role to prevent the disclosure of sensitive information? A. Permissions assigned to each user/group/role are [XYZ]. Q. What permissions are granted to plugins, and how are they managed to prevent excessive privilege? A. Permissions for each plugin are [XYZ]. When new plugins are needed the process to assign privilege is [XYZ].	Q. What permissions are assigned to each of the groups/roles that are used to limit access to the AI platform? Do any roles exceed the necessary access permissions? A. Access to AI platform is enforced by [using/not using] the principle of least privilege. Users [do/do not] only have access to systems necessary to perform their duties. Permissions for each of the groups are [XYZ]. Q. Within the AI platform, what limitations are implemented to prevent the chat plugin from accessing sensitive information? A. [XYZ] limitations are implemented to prevent the chat plugin from accessing sensitive information within the AI platform. Q. What permissions are assigned to the AI Platform to prevent the disclosure of sensitive information? A. Permissions assigned to the AI Platform are [XYZ]. Q. What kind of controls are in place to prevent LLMs from having excessive autonomy and to limit access to only needed information? A. The following [XYZ] processes and procedures are in place to prevent the LLM from having too much autonomy. Additionally, the following [XYZ] permissions are in place to limit which data they may access. Q. What permissions, and how are they managed and assigned to LLM plugins to prevent excessive permissions? A. Permissions for each of the plugins are [XYZ]. When new plugins are needed the process to assign privileges is [XYZ].	Q. What permissions are assigned to each of the groups/roles that are used to limit access to the AI Models? Do any roles exceed necessary access permissions? A. Access to AI models is enforced by [using/not using] least privilege access. Users [do/do not] only have access to systems necessary to perform their duties. Permissions for each of the groups are [XYZ]. Q. What permissions are assigned to the AI Model to prevent the disclosure of sensitive information? A. Permissions assigned to the AI Model are [XYZ].	Q. What permissions are assigned to each of the groups/roles that are used to limit access to the AI and LLM training data? Do any roles exceed the necessary access permissions? A. Access to AI training data is enforced by [using/not using] least privilege access. Users [do/do not] only have access to systems necessary to perform their duties. Permissions for each of the groups are [XYZ]. Q. What permissions are assigned to AI Data to prevent the disclosure of sensitive information? A. Permissions assigned to the AI Data are [XYZ].

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
AC-07-00	Q. How many concurrent sessions can a user of the AI have at the same time? A. Users can have [XYZ] number of concurrent sessions, while using the AI. Q. Are users of the AI automatically disconnected if they are inactive for more than [XYZ] number of minutes? A. Users are disconnected after [XYZ] number of minutes. Q. What steps are being taken to prevent malicious users from using techniques such as (CAPTHCA, Deep Fakes, Identity Spoofing, Synthetic Identities) from logging into the AI environment? A. [steps] are being taken to prevent malicious users from using techniques such as (CAPTHCA, Deep Fakes, Identity spoofing, Synthetic identities) from logging into the AI environment.	Q. What steps are being taken to prevent malicious users from using techniques such as (CAPTHCA, Deep Fakes, Identity Spoofing, Synthetic Identities) from logging into the AI platform? A. [steps] are being taken to prevent malicious users from using techniques such as (CAPTHCA, Deep Fakes, Identity spoofing, Synthetic identities) from logging into the AI platform.		
AC-12-00		Q. How are user sessions terminated such that processing by AI components does not continue past the end of the user session? A. AI sessions are terminated according to [conditions/triggers/policy].		
AC-14-00	Q. Which parts, if any, of the environment are accessible without user authentication? A. The areas that do not require authentication are [list]. Q. What permitted actions are allowed within a system without identification or authentication within the AI environment? A. The [actions] are permitted without identification or authentication within the AI environment.	Q. Which parts of the AI are accessible without user authentication? A. The functions that do not require authentication are [list]. Q. What permitted actions are allowed within a system without identification or authentication within the AI platform? A. The [actions] are permitted without identification or authentication within the AI platform.		Q. What kind of user authentication is required to access the AI training data, and who can access this data? A. Users are required to use the following authentication methods [XYZ], and the users that have access to the AI training data are [XYZ].

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
AC-17-00	Q. How is remote access from external networks to the environment restricted? What access controls are in effect to only allow authorized parties from accessing the environment? Which mechanisms are employed to facilitate secure and encrypted remote access while allowing for visibility to scan and detect malicious code/files. A. Environment remote access is restricted by [XYZ]. The access controls in effect at the environment are [controls (CA-3)]. The mechanisms in use to facilitate secure remote connections are [mechanisms].	Q. How is remote access to AI components (platform) restricted from external networks? What access controls are in effect to only allow authorized parties from accessing the AI platform? Which mechanisms are employed to facilitate secure and encrypted remote access while allowing for visibility to scan and detect malicious code/files. A. AI component remote access is restricted by [XYZ]. The access controls in effect at the AI platform are [controls (CA-3)]. The mechanisms in use to facilitate secure remote connections are [mechanisms].		
AC-20-00	Q. How is access to the AI environment and its components restricted from external access? A. The following [restrictions] are in place to ensure that access from external systems is not possible. Q. When connected to the environment, are there any external services (i.e. Internet, networks not part of the system) that can be accessed? If yes, which services can be accessed and why? A. Yes/No. If yes, you can access [XYZ] for the following [purposes].	Q. How is access to the AI platform restricted from external access? A. The following [restrictions] are in place to ensure that access from external systems is not possible.	Q. How is access to the AI models restricted from external access? A. The following [restrictions] are in place to ensure that access from external systems is not possible.	Q. How is access to AI, production, model, and validation data restricted from external access? A. The following [restrictions] are in place to ensure that access from external systems is not possible.
AC-21-00	Q. How is information sharing with other organizations (internal or external) restricted and reviewed to minimize the possibility of sensitive information being invertedly disclosed? A. We review information that is being shared by [means]. Sensitive information is restricted in the following [manner].	Q. What kind of protections does the AI platform provide to protect against sensitive information disclosure? A. The AI platform provides the following [safeguards] to protect against inadvertent sensitive information disclosure.		Q. How is data sanitized to prevent the disclosure of sensitive information? A. Data is sanitized by [methods].

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
AC-23-00		Q. How is the platform protected from data mining by agents attempting to find patterns in the AI platform? A. The platform is protected from adversarial data mining by [methods].		
AC-24-00	Q. Is access control enforced prior to each access request and limited to only authorized actions and users? A. Access control is enforced prior to access requests by [methods and/or mechanisms] for authorized actions and users.	Q. Is access control enforced prior to each access request and limited to only authorized actions and users? A. Access control is enforced prior to access requests by [methods and/or mechanisms] for authorized actions and users.		
AT-03-00			Q. What kind of security training is provided to personnel involved in AI development, deployment, and maintenance, in terms of how AI Models should be protected? A. We provide the following [training] to all personnel involved in the life cycle of the AI.	Q. What kind of security training is provided to personnel involved in AI development, deployment, and maintenance, in terms of how production AI training data should be protected? A. We provide the following [training] to all personnel involved in the life cycle of the AI.
AU-02-00	Q. What kinds of events on the environment is the system capable of logging to support auditing? What channels are used to coordinate logging functions with other organizational entities to provide audit-related info to inform selection of logging criteria? Are the selected event criteria for logging sufficient for supporting after-the-fact investigations, and how would the audit trail be used to reconstruct an incident. A. At the environment level we log [list of events]. The events are communicated to other organizational entities to inform them of their selection of logging criteria. The data selected for logging is sufficient for conducting investigations because [reasons], as the [specific logs] can construct a cohesive audit trail.	Q. What kinds of events on the platform is the system capable of logging to support auditing? What channels are used to coordinate logging functions with other organization entities to provide audit related info to inform selection of logging criteria? Are the selected event criteria for logging sufficient for supporting after-the-fact investigations, and how would the audit trail be used to reconstruct an incident. A. The platform is capable of logging [capabilities], and it is used to log [selected event criteria]. These events are communicated to other organizational entities to inform their selection of logging criteria. The data selected for logging is sufficient for conducting investigations because [reasons], as the [specific logs] can construct a cohesive audit trail.	Q. What types of events related to access and use of AI models are logged? A. The [types of events] are logged.	Q. What types of events related to access and use of AI data are logged? A. The [types of events] are logged.

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
AU-03-00			Q. What kind of information is logged with regards to access to AI Models? A. The [list of information types] are logged with, at a minimum, time stamps, who accessed, and what kind of action was taken.	Q. What kind of information is logged with regards to access to AI Data? A. The [list of information types] are logged with, at a minimum, time stamps, who accessed, and what kind of action was taken.
AU-06-00		Q. What kind of audit records are generated for the chat plugin in the AI platform? A. The [kinds of audit records] are generated for the chat plugin in the AI platform. Q. What kind of audit records are generated by the AI platform when sensitive information is accessed? A. The [kinds of audit records] audit records are generated by the AI platform when sensitive information is accessed.		
AU-06-05	Q. How are audit records analyzed to identify inappropriate or unusual activity in the environment? A. Audit records are analyzed to identify inappropriate or unusual activity in the environment by [methods].	Q. How are audit records analyzed to identify inappropriate or unusual activity in the platform? A. Audit records are analyzed to identify inappropriate or unusual activity in the platform by [methods].		
AU-09-00		Q. What are the means of protecting audit information and logging tools from unauthorized access, modification, and deletion on the AI platform? A. Audit information and logging tools are protected through [means]. Alerts are generated upon discovery of unauthorized access, modification, and deletion of audit information.		Q. What are the means of protecting audit information from unauthorized access, modification, and deletion in the AI data. A. Audit information of AI data is protected through [means]. Alerts are generated upon discovery of unauthorized access, modification, and deletion of audit information.

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
CA-02-00	Q. What kinds of assessment procedures are used to ensure that security concerns are properly tested for the AI environment? A. The [list of procedures] are followed.	Q. What kinds of assessment procedures are used to ensure that security concerns are properly tested for the AI platform? A. The [list of procedures] are followed. Q. What assessment procedure checks are there to ensure bias has not been introduced into the AI platform? A. The [assessment procedure checks] are implemented to ensure bias has not been introduced into the AI platform.	Q. What kinds of assessment procedures are used to ensure that security concerns are properly tested for the AI models? A. The [list of procedures] are followed. Q. What assessment procedure checks are there to ensure bias has not been introduced into the AI models? A. The [assessment procedure checks] are implemented to ensure bias has not been introduced into the AI models.	Q. What kinds of assessment procedures are used to ensure that security concerns are properly tested for the AI data? A. The [list of procedures] are followed. Q. What assessment procedure checks are there to ensure bias has not been introduced into the AI data? A. The [assessment procedure checks] are implemented to ensure bias has not been introduced into the AI data.
CA-08-00	Q. How often are penetration tests performed on the AI environment? What is the scope of the penetration testing? A. Penetration testing is performed at [frequency]. The scope of these tests include [scope].	Q. How often are penetration tests performed on the AI platform? What is the scope of the penetration testing? A. Penetration testing is performed at [frequency]. The scope of the tests include [scope].		
CA-09-00	Q. Provide documentation of all internal system interfaces, including all AI components, that show the interface characteristics, security and privacy requirements, and the nature of the information communicated. A. The SSP and [other documents] show all internal interfaces, ports, and protocols. Q. How often is interface documentation reviewed? A. We review interface documentation [frequency].			
CM-02-00		Q. What baseline configuration is used to determine if bias was introduced into the AI platform? A. [baseline configurations] were used to determine if bias was introduced into the AI platform.	Q. What baseline configuration is used to determine if bias was introduced into the AI models? A. [baseline configurations] were used to determine if bias was introduced into the AI models.	Q. What baseline configuration is used to determine if bias was introduced into the AI data? A. [baseline configurations] were used to determine if bias was introduced into the AI data.
CM-03-00	Q. What kind of documentation is needed to request changes to the environment? A. The [set of documents] is needed for all proposed changes, including type of change, justification, testing procedures, and possible effects on other components.	Q. What documentation is needed to request changes to the AI platform? A. The [set of documents] is needed for all proposed changes, including type of change, justification, testing procedures, and possible effects on other components.		

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
CM-05-00	Q. What is the change management process associated with the AI environment, concerning changes to the hardware or firmware defined, documented, and enforced? A. The process to make changes to the environment is [XYZ]. This process [is/is not] documented, and [does/does not] require approval and enforcement of physical and logical access restrictions associated.	Q. How is the change management process associated with the AI platform or software defined, documented, and enforced? A. The process to make changes to the platform or software is [XYZ]. This process [is/is not] documented, and [does/does not] require approval and enforcement of physical and logical access restrictions associated. Q. How are changes to the chat plugin authorized within the AI platform? A. Changes to the chat plugin are authorized via [methods] within the AI platform.	Q. What is the change management process associated with systems that house AI Models? A. The process to make changes to the models(s) is [XYZ]. This process [is/is not] documented, and [does/does not] require approval and enforcement of physical and logical access restrictions associated.	
CM-07-00	Q. How can you access the environment that hosts the AI network? Include network protocols, locations, what permissions are needed, what controls mechanisms are in place. A. Access to the AI environment can be obtained via the [protocols], from the [network segments] and the [permissions] associated with your account. Enforcement is done via [methods]. Q. How is access to LLM plugins restricted? A. The following [XYZ] restrictions are in place to limit access to the LLM plugins.	Q. How can you access the system(s) that host the AI platform? Include network protocols, locations, what permissions are needed/ not needed to limit model access to necessary features, what controls mechanisms are in place. A. Access to the systems that host AI Platform can be obtained via the [protocols], from the [network segments] and the [permissions] associated with your account. Enforcement is done via [methods]. Q. How does the AI Platform restrict the information that LLM plugins can access? A. The platform restricts access to the LLM plugins by [methods].	Q. How can you access the system(s) that host AI Models? Include network protocols, locations, what permissions are needed/ not needed to limit model access to necessary features, what controls mechanisms are in place. A. Access to the systems that host AI Models can be obtained via the [protocols], from the [network segments] and the [permissions] associated with your account. Enforcement is done via [methods].	Q. How is access to training data restricted? A. Access to training data is restricted by [means].

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
CM-11-00	Q. Describe how users are prevented from installing software on the systems or introducing new tools/executables onto the environment? A. The [mitigations] are in place to ensure that users cannot install or introduce new tools to the environment. Q. What is the process that is followed before new software can be installed or new tools/executables can be introduced? A. The [process] ensures that users cannot install or introduce new tools to the environment. Q. How often are software and tools installed or copied to the environment reviewed to ensure that they are still needed? A. We review tools/executables used on the system on a [frequency] basis to make sure that they are still needed and safe.	Q. Describe how users are prevented from installing software on the systems or introducing new tools/executables onto the AI platform? A. The [mitigations] are in place to ensure that users cannot install or introduce new tools to the AI Platform. Q. What is the process that is followed before new software can be installed or new tools/executables can be introduced? A. The [process] ensures that users cannot install or introduce new tools to the AI Platform. Q. How often are software and tools installed or copied to the environment reviewed to ensure that they are still needed? A. We review tools/executables used on the system on a [frequency] basis to make sure that they are still needed and safe.		
CM-13-00		Q. How are system data actions that process PII and FTI mapped and documented for the AI platform? A. System data actions that process PII and FTI are mapped and documented for the AI platform by [methods].		
CM-14-00	Q. What is the process to install patches to the infrastructure components? A. The [process] is used to install patches.	Q. What is the process to install patches to the AI platform and its components? A. The [process] is used to install patches.		
CP-01-00	Q. What are the contingency plans to follow in case of an emergency and how are they disseminated to the appropriate personnel? A. Contingency plans are included in [set of documents] and they are provided to all necessary users via [methods].	Q. What are the contingency plans to follow in case of an emergency and how are they disseminated to the appropriate personnel? A. Contingency plans are included in [set of documents] and are provided to all necessary users via [methods].		
CP-09-00	Q. How are backups of AI Environment made? A. The AI environment is backed up using [methods and schedule].	Q. How are backups of the AI platform made? A. The platform is backed up using [methods and schedule].	Q. How are backups of the AI models made? A. The AI models are backed up using [methods and schedule].	Q. How are backups of AI Data made? A. The AI data is backed up using [methods and schedule].

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
IA-02-00	Q. How are users with access to the environment managed, authenticated, reviewed, and vetted to ensure that they require access? A. The [process] is used to add new users, and they are authenticated via [confirmation methods], and we review user accesses on a [frequency] to ensure that they still require access, if they do not require access we remove them. Q. In the AI environment, in what way is AI being used to authenticate as well as confirm the user is a human? A. Accounts are uniquely identified and authenticated as real people via [methods].	Q. How are users with access to the AI Platform managed, authenticated, reviewed, and vetted to ensure that they require access? A. The [process] is used to add new users, and they are authenticated via [methods], and we review user access on a [frequency] to ensure that they still require access, if they do not require access we remove them. Q. In the AI platform, in what way is AI being used to authenticate as well as confirm the user is a human? A. Accounts are uniquely identified and authenticated as real people via [methods].		
IA-02-01	Q. How is multi-factor authentication implemented for privileged accounts within the AI environment? A. Multi-factor authentication is implemented via [methods] for privileged accounts within the AI environment. Q. Is the system following NIST 800-63-3 in its choices for multi-factor authentication within the AI environment? A. The system follows [NIST-800-63-3] in its choices for multi-factor authentication within the AI environment.	Q. Is the system following NIST 800-63-3 in its choices for multi-factor authentication within the AI platform? A. The system follows [NIST-800-63-3] in its choices for multi-factor authentication within the AI platform.		
IA-02-02	Q. Is the system following NIST 800-63-3 in its choices for multi-factor authentication within the AI environment? A. The system follows [NIST-800-63-3] in its choices for multi-factor authentication within the AI environment.	Q. Is the system following NIST 800-63-3 in its choices for multi-factor authentication within the AI platform? A. The system follows [NIST-800-63-3] in its choices for multi-factor authentication within the AI platform.		

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
IA-03-00	Q. How are devices uniquely identified and authenticated before establishing a connection to an environment containing AI systems? A. Devices are uniquely identified by [methods/ identifiers] and authenticated by [methods].			
IA-06-00	Q. How is the feedback of authentication information, during the authentication process, protected from possible disclosure to and exploitation from unauthorized parties? A. Feedback from authentication information is protected by [methods].			
IA-08-00	Q. Do you have any external users that use the environment? If so, how do you identify them? A. We [do/do not] have external users. (If yes) Users are identified via [methods] and they are required to register by [methods]. Q. How are non-organizational users uniquely identified and authenticated within the AI environment? A. Non-organizational users are uniquely identified and authenticated via [methods].	Q. Do you have any external users that use the AI Platform? If you do, how do you identify them? A. We [do/do not] have external users. (If yes) Users are identified via [methods] and they are required to register by [methods]. Q. How are non-organizational users uniquely identified and authenticated within the AI platform? A. Non-organizational users are uniquely identified and authenticated via [methods].		
IA-12-00	Q. How do you validate the identity of a user who is attempting to access the AI environment with the following threats? [Deep fakes, Identity spoofing, Synthetic identity, CAPTCHA]? A. We validate the identity of the user using the following trusted sources of identity matching. We also ensure that the person is physically present at the point of verification when employing biometrics.	Q. How do you validate the identity of a user who is attempting to access the AI platform with the following threats? [Deep fakes, Identity spoofing, Synthetic identity, CAPTCHA]? A. We validate the identity of the user using the following trusted sources of identity matching. We also ensure that the person is physically present at the point of verification when employing biometrics.		
MA-03-00	Q. What is the process to install system patches to the production environment? A. The [process] is used to install patches.	Q. What is the process to install patches to the production AI platform? A. The [process] is used to install patches.		

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
MA-05-00	Q. What is the process to authorize maintenance personnel that will be accessing the environment? A. The [process] is used for authorizing maintenance personnel. Q. What is the process to ensure that maintenance personnel only have access to the components needing maintenance? A. Maintenance personnel are escorted while performing their duties by someone from the internal team to ensure that they do not attempt to access other parts of the environment.	Q. What is the process to authorize maintenance personnel that will be accessing the AI platform? A. The [process] is used for authorizing maintenance personnel. Q. What is the process to ensure that maintenance personnel only have access to the components needing maintenance? A. Maintenance personnel are escorted while performing their duties by someone from the internal team to ensure that they do not attempt to access other parts of the platform.		
MA-06-00	Q. How often are patches applied to devices (physical or virtual) in the environment? A. Patches are applied on a [frequency] basis.	Q. How often are patches applied to the AI platform? A. Patches are applied on a [frequency] basis.		
PE-11-00	Q. What backup power procedures are in place? A. The [procedures] are in place deal with power loss situations. [Consider AI architectures designed to adequately allow for checkpoint and restore operations in case of loss or corruption due to power loss]			
PL-02-00		Q. How is the context of AI usage and inventory biases reviewed? A. Context of AI usage to ascertain and inventory biases is reviewed using [documents]. Q. What security and privacy plans have been implemented to ensure bias is not introduced into the AI platform? A. The [security and privacy plan] has been implemented to ensure bias has not been introduced into the AI platform.	Q. What security and privacy plans have been implemented to ensure bias is not introduced into the AI models? A. The [security and privacy plan] has been implemented to ensure bias has not been introduced into the AI models.	Q. What security and privacy plans have been implemented to ensure bias is not introduced into the AI data? A. The [security and privacy plan] has been implemented to ensure bias has not been introduced into the AI data.
PL-04-00		Q. What methods are used to prevent bias from entering AI platforms and software? A. Bias is prevented from entering AI platform and software by [methods].	Q. What methods are used to prevent bias from entering AI models? A. Bias is prevented from entering AI models by [methods].	Q. What methods are used to prevent bias from entering the AI datasets? A. Bias is prevented from entering AI datasets by [methods].

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
PL-08-00	Q. What safeguards are in place to protect personally identifiable information? A. The [safeguards] are in place to protect PII.			
PM-07-00	Q. How is information privacy being protected? A. Information privacy is being protected by [methods].			
PM-12-00	Q. Describe the insider threat program that is in place and how it is used to safeguard the AI Environment. A. The [insider threat program] is in place and safeguards the AI environment using [methods]. (There is a good chance that you would be referred to a separate group/document that handles/describes this)	Q. Describe the insider threat program that is in place and how it is used to safeguard the AI platform. A. The [insider threat program] is in place and safeguards the AI Platform using [methods]. (There is a good chance that you would be referred to a separate group/document that handles/describes this)	Q. Describe the insider threat program that is in place and how it is used to safeguard the AI models. A. The [insider threat program] is in place and safeguards the AI models using [methods]. (There is a good chance that you would be referred to a separate group/document that handles/describes this)	Q. Describe the insider threat program that is in place and how it is used to safeguard the AI data. A. The [insider threat program] is in place and safeguards the AI datasets using [methods]. (There is a good chance that you would be referred to a separate group/document that handles/describes this)
PM-18-00	Q. What is the information privacy plan? A. The privacy plan is [XYZ].			
PM-30-00			Q. How is supply chain risk associated with the development, acquisition, maintenance, and disposal of AI Models managed? A. The supply change risk management strategy is [XYZ]. (It should include definitions for what acceptable risks are)	
RA-03-00	Q. What is the risk assessment process for the AI environment? A. The risk assessment process is [XYZ]. (ensure that threats and vulnerabilities are identified, also there should be an assessment of the likelihood of exploitation, possible effects in case of a successful attack)	Q. What is the risk assessment process for the AI platform? A. The risk assessment process is [XYZ]. (ensure that threats and vulnerabilities are identified, also there should be an assessment of the likelihood of exploitation, possible effects in case of a successful attack)		

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
RA-05-00	Q. By what frequency, or randomized protocol, are vulnerabilities monitored and scanned for within the system and hosted applications? A. The system and hosted applications are scanned at [frequency] or [randomized using XYZ]. Q. What tools and processes are employed to scan for vulnerabilities within the environment? How is interoperability among tools ensured, and parts of the vulnerability management process automated to update vulnerabilities to be scanned as well as [standards 1-3]? A. The [tools] are used within the environment to scan for vulnerabilities, and the [processes] are used. Interoperability is facilitated between tools by [configurations/process]. The parts of the management process that are automated are [automated processes]. The list of vulnerabilities to be scanned [is/is not] automatically updated. Q. What is the process that is undertaken when a vulnerability is detected? How is information shared across systems to help eliminate similar vulnerabilities? A. When a vulnerability is detected in the environment, it [is/is not] remediated using [XYZ] based on defined risk. Information [is/is not] shared throughout the organization using [XYZ].	Q. By what frequency, or randomized protocol, are vulnerabilities monitored and scanned for on the AI platform? A. The AI platform is scanned at [frequency] or [randomized using XYZ]. Q. What tools and processes are employed to scan for vulnerabilities on the AI platform? How is interoperability among tools ensured, and parts of the vulnerability management process automated to update vulnerabilities to be scanned as well as [standards 1-3]? A. The [tools] are used to scan for vulnerabilities on the AI platform, and the [processes] are used. Interoperability is facilitated between tools by [configurations/process]. The parts of the management process that are automated are [automated processes]. The list of vulnerabilities to be scanned [is/is not] automatically updated. Q. What is the process that is undertaken when a vulnerability is detected? How is information shared between teams managing various AI tools to help eliminate similar vulnerabilities? A. When a vulnerability is detected in the environment, it [is/is not] remediated using [XYZ] based on defined risk. Information [is/is not] shared throughout the organization using [XYZ].		
SA-03-02		Q. What is the approval process and documentation requirement for the use of live data in the preproduction AI platform? A. Approval is granted for the use of live data in preproduction AI platform according to [guidelines] and live data is protected with [controls].	Q. What is the approval process and documentation requirement for the use of live data in preproduction AI models? A. Approval is granted for the use of live data in preproduction AI models according to [guidelines] and live data is protected with [controls].	Q. What is the approval process and documentation requirement for the use of live data in preproduction AI datasets? A. Approval is granted for the use of live data in preproduction AI datasets according to [guidelines] and live data is protected with [controls].

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
SA-08-00		Q. How should AI systems be designed resiliently, with the ability to detect, respond to, and recover from disruptions and failures in a timely and effective manner? A. AI systems are designed resiliently, with the ability to detect using [methods], respond to using [methods], and recover from disruptions and failures in a timely and effective manner using [methods]. Q. What systems security and privacy engineering principles are applied when designing plugins? A. The [systems security and privacy engineering principles] are applied during design.	Q. How should AI systems be designed resiliently, with the ability to detect, respond to, and recover from disruptions and failures in a timely and effective manner? A. AI systems are designed resiliently, with the ability to detect using [methods], respond to using [methods], and recover from disruptions and failures in a timely and effective manner using [methods].	Q. How should AI systems be designed resiliently, with the ability to detect, respond to, and recover from disruptions and failures in a timely and effective manner? A. AI systems are designed resiliently, with the ability to detect using [methods], respond to using [methods], and recover from disruptions and failures in a timely and effective manner using [methods].
SA-09-00		Q. Do external providers of AI-enabled Privacy Enhancing Technologies (PETs) systems and services comply with agency [security and privacy requirements]. A. External providers of PETs must comply with [policies]. Q. How is access and documentation of trustworthiness for external model sources recorded? A. Access and documentation of trustworthiness for external model sources are recorded via [methods].		
SA-09-05		Q. How is the location of information processing, data storage, and system services being managed and controlled through an external provider? A. The location of information processing, data storage, and system services is managed by the external provider according to organizationally defined [criteria].		

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
SA-09-06				Q. How is control of cryptographic keys for encrypted AI data transmitted through external systems maintained? A. Exclusive control of cryptographic keys is maintained by [internal methods].
SA-09-08				Q. Where is AI data physically stored? A. Physical storage of AI data is contained exclusively within the legal jurisdictional boundary of the United States in [locations].
SA-10-00		Q. How are changes to the AI platform tracked? A. The [processes] are used to track changes to the AI platform. Q. How are security flaws tracked by the developer/vendor? A. The [processes] are used to track flaws. (ensure that this includes how flaws are fixed). Q. How is retrained data and development to model configurations updated within the AI platform to prevent AI bias? A. Retrained data and development to model configurations is updated via [methods] within the AI platform to prevent AI bias.	Q. How does the system audit model predictions for fairness and equity? A. The system regularly audits model predictions for fairness and equity via [methods]. Q. How is retrained data and development to model configurations updated within the AI models to prevent AI bias? A. Retrained data and development to model configurations is updated via [methods] within the AI models to prevent AI bias.	Q. How is retrained data and development to model configurations updated within the AI data to prevent AI bias? A. Retrained data and development to model configurations is updated via [methods] within the AI data to prevent AI bias.
SA-17-00	Q. How is access limited to the architecture of the system? A. Access to the architecture of the system is limited via [methods].	Q. Has the inclusion of AI-enabled PETs systems been vetted against the agency Enterprise Architecture? A. The agency Enterprise Architecture has vetted the use of PETs in the system. Q. How is access limited to model weights and hyperparameters of the system? A. Access to the model weights and hyperparameters are limited via [methods].		

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
SA-22-00	Q. What is the process to replace components that are no longer supported by the vendor? A. When an environment component is no longer supported the process of finding a replacement is [methods]. (this plan should include mitigations to be used while considering new components)	Q. What is the process to replace AI platform components that are no longer supported by the vendor? A. When an environment component is no longer supported the process of finding a replacement is [methods]. (this plan should include mitigations to be used while considering new components)		
SC-02-00	Q. How is system management functionality of the AI environment separated from user functionality and interface devices? A. The system management functionality of the AI environment is [physically/ logically] separated from the user functionality. This is done using [tools/ policies/ techniques].	Q. How is system management functionality of the AI platform separated from user functionality? A. The system management functionality of the AI environment is [physically / logically] separated from the user functionality. This is done using [tools/ policies/ techniques].		
SC-03-00	Q. How are security functions of the AI environment isolated from non-security functions (regular user functions and management functions)? A. Security functions of the AI environment are [physically/ logically] isolated from other functions within the AI environment using [techniques/ policies/ tools]	Q. How are security functions of the AI platform isolated from non-security functions (regular user functions and management functions)? A. Security functions of the AI platform are [physically/ logically] isolated from other functions within the AI environment using [techniques/ policies/ tools]		
SC-04-00	Q. Which measures are in place to prevent unintended information disclosures to unauthorized individuals or roles via shared system resources after they have been released back to the system? This includes encrypted data. A. The [measures] are in place to prevent unintended information transfer via shared system resources are [policies/ measures/ tools].	Q. What measures are in place to prevent the unauthorized disclosure of data from the shared resources of an AI platform? A. The [measures] taken to prevent the unauthorized disclosure of data through the AI platform are [policies], which are carried out using [tools].	Q. What measures are in place that limit the possibility of data exfiltration from systems that contain AI Models? A. The measures in place to mitigate the exfiltration of data relating to AI models are [measures/tools].	Q. What protection is implemented to prevent unauthorized and unintended information transfer via shared data systems concerning AI training data at rest? A. The protection mechanisms in use to protect disclosure of AI data at rest are [mechanisms/ policies] which are carried out using [tools].

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
SC-05-00	Q. What kind of mechanisms are used to prevent Denial of Service (DOS) attacks against the systems comprising of the AI environment? A. The protections implemented throughout the AI environment include [tools/devices] using [configuration]. Q. How have controls been employed in the AI environment to achieve the denial-of-service mitigations for each type of denial-of-service event? A. The controls that are used within the environment are [controls/policies]. Q. What controls have been employed within the AI environment to achieve denial-of-service mitigations for cost harvesting? A. [XYZ controls] have been employed within the AI environment to achieve denial-of-service mitigations for cost harvesting.	Q. What kind of mechanisms are used to prevent DOS attacks against the systems that host the AI platform? A. The protections implemented on the AI platform are [tools/devices] using [configuration]. Q. How have controls been employed on the AI platform to achieve denial-of-service (DOS) mitigations for each type of DOS event? A. The controls that are used within the environment are [controls/policies]. Q. What controls have been employed within the AI platform to achieve denial-of-service mitigations for cost harvesting? A. [XYZ controls] have been employed within the AI platform to achieve denial-of-service mitigations for cost harvesting.		
SC-06-00	Q. How is resource availability protected in AI environments to ensure proper prioritization? A. Resource availability is protected within the AI environment by allocating [resource amount] according to [priority levels/quotas/controls].	Q. How is resource availability protected for the AI platform? A. Resource availability is protected by the AI platform by allocating [resources] according to [priority/quotas/controls].		

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
SC-07-00	Q. How are communications monitored and controlled at the externally managed interfaces to the AI environment; and at key internal managed interfaces within the AI environment? A. Communications are monitored using [tools/methods] and controlled using [tools/policies], from external managed interfaces and key internal interfaces within the AI environment. Q. How are subnetworks configured for publicly accessible elements of the AI environment to be separated from internal organizational networks? A. The subnetworks for publicly accessible systems in the environment are [physically/ logically] separated from internal organization networks. Q. How are external networks permitted to be connected to from within the AI environment? How is boundary protection established at these connection points? A. External networks are only allowed to be connected to from managed interfaces with boundary protections. Boundary protection devices consist of [devices].	Q. How are communications monitored and controlled at the externally managed interfaces to the AI platform? A. Communications are monitored using [tools/methods] and controlled using [tools/policies], from externally managed interfaces connected to the AI platform. Q. How are subnetworks configured for publicly accessible elements of the AI platform to be separated from internal organizational networks? A. The subnetworks for publicly accessible elements of the AI platform are [physically/ logically] separated from internal organization networks. Q. How does the AI platform handle outgoing external network connections? How is boundary protection established at these connection points? A. External networks are only allowed to be connected to from managed interfaces with boundary protections. Boundary protection devices consist of [devices].		Q. How are communications with AI training data monitored and controlled at externally managed interfaces? A. Communications are monitored using [tools/methods] and controlled using [tools/policies], from externally managed interfaces.

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
SC-08-00	Q. How is the confidentiality and integrity of transmitted information protected to/from/within the AI environment, including internal and external networks, and all system components capable of transmitting information? A. The [physical and/or logical] means to protect the confidentiality and integrity of all transmitted information through AI environment are [tools, techniques, policies]. Q. How is the confidentiality and integrity of information, in transit, protected when using LLM plugins? A. CI in transit is protected via [methods].	Q. How is the AI platform designed to protect the confidentiality and integrity of transmitted information? A. The AI platform protects the confidentiality and integrity of transmitted information by using [tools, techniques, policies].	Q. How is the confidentiality and integrity of transmitted models assured? A. The confidentiality and integrity of transmitted models is assured by [tools, techniques, policies]. Q. How is cryptography used to prevent unauthorized disclosure of AI models and detect changes to models during data transmissions? A. Cryptography is used to protect AI models from disclosure and change by [methods.]	Q. How is the confidentiality and integrity of transmitted information protected within the AI training data or data storage systems during transmission? A. The confidentiality and integrity of AI data is protected during transmission because of [tools, techniques, policies] used within the AI training data and/or data storage systems. Q. How is cryptography used to prevent unauthorized disclosure of AI data and detect changes to data during transmission? A. Cryptography is used to protect AI data from disclosure and change by [methods.]
SC-10-00	Q. What is the policy for terminating network connections in the AI environment associated with communication sessions.? A. The network connection is to be terminated [at the end of the session or after the org-defined period of inactivity] to internal and external networks from the AI environment.	Q. What is the policy for terminating communication sessions on the AI platform? A. The network connection is to be terminated [at the end of the session or after the organizationally defined period of inactivity] to internal and external networks from the AI platform.		Q. How are communication sessions with AI data terminated? A. The network connection accessing AI data internally or externally is terminated [at the end of the session or after the organization-defined period of inactivity].
SC-12-00	Q. How are cryptographic keys established and managed within the AI environment in accordance with organization policy- containing standards for key generation, distribution, storage, access, and destruction? A. Cryptographic key management is an [automated/manual] procedure. Keys are generated, stored, accessed, and destroyed through [organization defined requirements], using managed trust stores from only approved trust anchors.		Q. How are cryptographic keys established to protect the confidentiality and integrity of AI models in storage in accordance with the organization-defined requirements for key generation, distribution, storage, access, and destruction? A. Cryptographic key management is an [automated/manual] procedure. Keys are generated, stored, accessed, and destroyed through [organization defined requirements], using managed trust stores from only approved trust anchors.	Q. How are cryptographic keys established and managed in the storage, encryption, and verification of integrity of AI data, in accordance with organization-defined requirements for key generation, distribution, storage, access, and destruction? A. Cryptographic key management is an [automated/manual] procedure. Keys are generated, stored, accessed, and destroyed through [organization defined requirements], using managed trust stores from only approved trust anchors.

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
SC-13-00			Q. How is cryptography being used to protect the confidentiality and integrity of AI models? A. Cryptography is being used by the AI model to protect its confidentiality through [methods] and integrity using [methods].	Q. How is cryptography being used to protect the confidentiality and integrity of AI data? A. Cryptography is being used in the AI data to protect its confidentiality through [methods] and integrity using [methods].
SC-15-00	Q. Is remote activation of collaborative computing devices and applications prohibited, and what are the exceptions? A. Yes, remote activation is prohibited, with the following [exceptions]. Q. What is the explicit indication of use provided to users physically present at the collaborative computing devices and applications within the AI environment when they are activated? A. The explicit indication of use is [indication].	Q. Is remote activation of collaborative AI powered applications prohibited, and what are the exceptions? A. Yes, remote activation is prohibited, with the following [exceptions]. Q. What is the explicit indication of use provided to users physically present at the AI powered applications when they are activated? A. The explicit indication of use is [indication].		
SC-17-00	Q. How are public key infrastructure (PKI) certificates issued for the AI environment? A. The [certificate policy or approved service provider] provides PKI for the AI environment. Only approved trust anchors are included in managed [trust stores or certificate stores].	Q. How are PKI certificates issued for the AI platform? A. The [certificate policy or approved service provider] provides PKI for the AI platform. Only approved trust anchors are included in managed [trust stores or certificate stores].		Q. How are PKI certificates issued for AI data? A. The [certificate policy or approved service provider] provides PKI for the AI data. Only approved trust anchors are included in managed [trust stores or certificate stores].
SC-18-00	Q. Within the AI environment, how are acceptable and unacceptable mobile code and mobile code technologies defined? A. The policies defining acceptable/ unacceptable uses of mobile code are [policies]. Q. How is the use of mobile code within the AI environment authorized, monitored, and controlled? A. The use of mobile code in the AI environment is authorized, monitored, and controlled using [tools] and [policies].	Q. On an AI platform, how are acceptable and unacceptable mobile code and mobile code technologies defined? A. The policies defining acceptable/ unacceptable uses of mobile code are [policies]. Q. How is the use of mobile code within the AI platform authorized, monitored, and controlled? A. The use of mobile code on the AI platform is authorized, monitored, and controlled using [tools] and [policies].		

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
SC-23-00	Q. How is confidence in data validity being established during communication sessions throughout the AI environment? How are users able to verify the authenticity of identities with parties at either end of a communication session? What are the protections against man-in-the-middle (MITM) attacks, session hijacking, and false information injections. A. Users can be confident that data is legitimate during a communication session on the environment because of [protections]. Integrity of identity if validated because of [protections] and ensure protection against MITMs, session hijacking, and false data injection.	Q. How does the AI platform ensure the validity of outgoing data and incoming requests during a communication session? How can a user ensure they are speaking directly with the AI platform, with no alteration or interception. A. The AI platform protects the authenticity of all communication sessions using [methods] to validate identity and data of both ends of a communication session.		
SC-24-00	Q. How is failure in a known state implemented to prevent loss or damage to the environment? A. Failing to a known state is accomplished by [means].	Q. How is system state information preserved to facilitate system restart and return to operational mode with minimal disruption? A. State information is preserved by [means].	Q. What are the organizational defined fail states defined for the AI model? A. The ODPs defined for failing to known states are [ODPs].	
SC-28-00	Q. What measures are employed to protect the confidentiality and integrity of information at rest in the AI environment? How are disk drives, network storage devices, and databases protected from breach, and in the event of a breach, how is confidentiality maintained? A. The measures employed to protect information at rest are [measures]. Environment storage devices are protected from access using [tools] and [protocols] are used to protect the data from disclosure.	Q. How is AI component software protected at rest? A. AI component software is protected at rest by [means].	Q. What measures are employed to protect the confidentiality and integrity of the AI models? How are devices and systems containing AI models protected from breach? How is the integrity of AI models protected from alteration and unauthorized disclosure in the event of a breach? A. The measures employed to protect models' information at rest are [measures]. Storage devices containing models are protected from unauthorized access using [tools] and [protocols] are used to protect the data from disclosure. The integrity of the models is verified using [protocols].	Q. What measures are employed to protect the confidentiality and integrity of the AI data? How are devices and systems containing AI data protected from breach? How is the integrity of AI training data protected from unauthorized alteration? A. The measures employed to protect the AI training data at rest are [measures]. Storage devices containing data are protected from unauthorized access using [tools] and [protocols] are used to protect the data from disclosure in the event of a breach. The integrity of the data is verified using [protocols] to ensure it has not been tampered with or poisoned via a breach.

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
SC-37-00	Q. How are out-of-band channels protected to prevent unauthorized introduction of AI components, models, and data into the environment? A. Out-of-band channels prevent the unauthorized introduction of AI components, models, and data into the environment by [methods].	Q. Are authorized out-of-band channels enabled to ensure business continuity and availability of AI-enabled systems? A. The [authorized out-of-band channels] are enabled to ensure business continuity and availability of AI-enabled systems.		
SC-39-00	Q. How is each executing system process assigned a separate address space to maintain separate execution domains within the AI environment? A. Separate address space for the AI environment is established by [means].	Q. How is each executing system process assigned a separate address space to maintain separate execution domains within the AI platform? A. Separate address space for the AI platform is established by [means]. Q. How does the AI platform enforce process isolation for LLM plugins? A. The AI platform isolates plugins by [means].		
SI-02-00	Q. What is the protocol for flaw remediation upon detection of a system flaw, and is remediation incorporated in organization configuration management process? How is software and firmware tested for effectiveness and potential side effects before installation? What is the lag time between the release of a security relevant software/firmware update and the installation of the update? A. Upon identifying a system flaw using [method/tool], which is [integrated/ not integrated] within organization config management process, flaws are reported and sent for correction. Software/firmware updates concerning flaw remediation [are/are not] tested for effectiveness and potential side effects before installation. Security relevant software/firmware updates are installed within [period] and [do/ do not] comply with organization installation standards.	Q. What is the protocol for flaw remediation upon detection of a flaw concerning AI platform, and is remediation incorporated in organization configuration management process? How is the AI platform, and accompanying software and firmware tested for effectiveness and potential side effects before installation? What is the lag time between the release of a security relevant software/firmware update and the installation of the update? A. Upon identifying a system flaw using [method/tool], which is [integrated/ not integrated] within organization config management process, flaws are reported and sent for correction. Software/firmware updates concerning flaw remediation [are/are not] tested for effectiveness and potential side effects before installation. Security relevant software/firmware updates are installed within [period] and [do/ do not] comply with organization installation standards.		

SI-03-00	Q. Which [signature/non signature based] mechanisms are deployed at system entry and exit points to detect and remove malicious code? How often are these mechanisms updated with the latest releases of software, tools, policies, and configurations? A. The environment employs [signature/non signature] based detection across [all/some] entry and exit points. These mechanisms are capable of scanning [all/certain] types of files, [including/not including] compressed files and hidden code (steganography). These mechanisms are updated with latest [code signatures/ reputation-based technology/heuristics]. Q. Are protection mechanisms configured to perform periodic scans of the entire environment, and real time scans of external files as they are [downloaded, opened, or executed]? How are protection mechanisms (anti exploitation software) configured to operate? How are falsely positive detections and removals of code addressed concerning the impact on availability of the environment? A. Protection mechanisms perform real time scans on all files as they are [downloaded/ opened/ executed] before they are allowed to enter the environment. The entire environment is also scanned [periodically/piece mail] to detect malicious code that made it past real time protection mechanisms. Anti exploitation protection mechanisms prevent compromise by blocking malicious code from running, while quarantining the file/program, and sending an alert to the security team automatically upon detection. The potential impact of false detection and removal of benign code on system availability is addressed by [protocol].	Q. Which [signature/non signature based] mechanisms are deployed at entry and exit points to the AI platform to detect and remove malicious code? How often are these mechanisms updated with the latest releases of software, tools, policies, and configurations? A. The AI platform employs [signature/non signature] based detection across [all/some] entry and exit points when processing data. These mechanisms are capable of scanning [all/certain] types of files, [including/not including] compressed files and hidden code (steganography). These mechanisms are updated with latest [code signatures/ reputation-based technology/heuristics]. Q. How are real-time scans of files performed as they are fed into the AI platform. How are protection mechanisms (anti exploitation software) configured to operate? A. Protection mechanisms perform real time scans on all files as they are imported into the AI platform, before they are allowed to be processed by the system. Periodic scans are performed on the AI platform to check for malicious code running on the platform. Anti exploitation protection mechanisms prevent compromise by blocking malicious code from being processed, while quarantining the file, and sending an alert to the security team automatically upon detection. The potential impact of false detection and removal of benign code on platform availability is addressed by [protocol]. Q. How is the AI platform robustness evaluated against adversarial attacks? A. Models are regularly evaluated for robustness against adversarial attacks by [methods].	Q. Which [signature/non signature based] mechanisms are used to detect model compromise? How often are these mechanisms updated with the latest releases of software, tools, policies, and configurations? A. The system employs [signature/non signature] based detection mechanisms, which are capable of scanning [all/certain] types of files, [including/not including] compressed files and hidden code (steganography). These mechanisms are updated with latest [code signatures/ reputation-based technology/heuristics]. Q. How are protection mechanisms configured to perform [periodic] scans of the entire model, and real time scans of all files and code as they are incorporated into the model? How are protection mechanisms (anti exploitation software) configured to preserve the integrity of the model? How are falsely positive detections and removals of code addressed concerning the impact on availability of the AI model? A. Protection mechanisms perform real-time scans on all files before they are allowed to enter the model. The entire model is also scanned [periodically] to detect malicious code that made it past real time protection mechanisms and to account for new intelligence. Anti exploitation protection mechanisms prevent compromise by blocking malicious code from running within the model, while quarantining the file/program, and sending an alert to the security team automatically upon detection. The potential impact of false detection and removal of benign code on model availability is addressed by [protocol]. Q. How is the AI model	Q. Which [signature/non signature based] mechanisms are deployed at storage devices to detect and remove malicious code? How often are these mechanisms updated with the latest releases of software, tools, policies, and configurations? A. The environment employs [signature/non signature] based detection across at entry points to storage devices. These mechanisms are capable of scanning [all/certain] types of files, [including/not including] compressed files and hidden code (steganography). These mechanisms are updated with latest [code signatures/ reputation-based technology/heuristics]. Q. How are protection mechanisms configured to perform periodic scans of the entire data storage architecture at rest, and real time scans of external files as they are downloaded? How are protection mechanisms (anti exploitation software) configured to operate within data storage? How are falsely positive detections and removals of code addressed concerning the impact on availability of the AI data? A. Protection mechanisms perform real-time scans on all files as they are downloaded and before they are allowed to enter the data storage system. The entire data storage is also scanned [periodically/piece mail] to detect malicious code that made it past real time protection mechanisms. Anti exploitation protection mechanisms prevent compromise by blocking malicious code from running, while quarantining the file/program, and sending an alert to the security team automatically upon detection. The potential impact of false detection and removal of benign code on data availability and integrity is addressed by

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
		Q. How is the system monitored for potential backdoors introduced into the AI platform during transfer learning? A. The system is monitored for potential backdoors introduced into the AI platform during transfer learning via [methods]. Q. In the AI platform, what malicious code mechanisms are implemented in the chat plugin to detect and eradicate malicious code? A. In the AI platform, [malicious code mechanisms] are implemented in the chat plugin to detect and eradicate malicious code. Q. What malicious code protection mechanisms are implemented in the AI platform to scan for indirect prompt injections? A. [malicious code protection mechanisms] are implemented in the AI platform to scan for indirect prompt injections. Q. What malicious code protection mechanisms are implemented in the AI platform to scan for direct prompt injections? A. [malicious code protection mechanisms] are implemented in the AI platform to scan for direct prompt injections. Q. What malicious code protection mechanisms are implemented in the AI platform to prevent malicious plugins from being used? A. The [procedures] are in place to ensure that plugins are safe.	robustness evaluated against adversarial attacks? A. Models are regularly evaluated for robustness against adversarial attacks by [methods]. Q. How is the system monitored for potential backdoors introduced into the AI models during transfer learning? A. The system is monitored for potential backdoors introduced into the AI models during transfer learning via [methods].	[protocol]. Q. How is the system monitored for potential backdoors introduced into the AI data during transfer learning? A. The system is monitored for potential backdoors introduced into the AI data during transfer learning via [methods]. Q. What malicious code protection mechanisms are implemented in the AI data to scan for indirect prompt injections? A. [malicious code protection mechanisms] are implemented in the AI data to scan for indirect prompt injections. Q. What malicious code protection mechanisms are implemented in the AI data to scan for direct prompt injections? A. [malicious code protection mechanisms] are implemented in the AI data to scan for direct prompt injections.

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
SI-04-00	Q. What policies/ tools/ devices are used to monitor the environment for unauthorized system use, and IOCs (including the detection of unauthorized local, network, and remote connections)? What is the process for analyzing detected events and anomalies using collected data? A. The environment is monitored using [methods]. Events [are/are not] analyzed for collection of threat intelligence. Q. What strategy is in play to deploy monitoring devices throughout the system to collect [essential info], and which locations are selected as ad hoc locations for tracking specific transactions of interest. How is monitoring level adjusted when considering changes in risk to organization? A. Locations for deploying monitoring devices throughout the environment and ad hoc locations are chosen through [means] and collect [certain info]. Monitoring [is/is not] expanded in times of increased risk to [organization operations and assets/ individuals/ other orgs/ the nation].	Q. How are attacks to AI detected on the platform? A. The system detects attacks on the AI platform using [methods/tools]. Q. What types of AI-specific information is monitored? A. The types of AI-specific information monitored is [methods]. Q. To what roles are monitoring results of AI-specific information directed? A. The monitoring results of AI-specific information is directed to [ISSO, System Owner, and the Insider threat group]. Q. How does system monitoring conduct thorough data quality assessments for the AI platform before training AI models? A. System monitoring conducts data quality assessments with [methods]. Q. How does system monitoring detect and remove malicious or biased data injected during training with the AI platform? A. System monitoring detects and removes malicious or biased data injected during training by [methods]. Q. How does the system monitor and detect indirect prompt injections within the AI platform? A. The system monitors and detects indirect prompt injections via [methods] within the AI platform. Q. How does the system monitor and detect direct prompt injections within the AI platform? A. The system monitors and detects direct prompt injections via [methods] within the AI platform.	Q. How are attacks to AI detected on the platform? A. The system detects attacks on the AI platform using [methods/tools]. Q: How is poisoned data detected for the AI Model? How are the tools that poisoned the data detected? A: Develop a set of the human ground-truth testing data set to assess the functionality enabled by AI then compare and evaluate the test results, analyze the nature of the deviation from the expected results to determine the level of severity of functional defects, source of the defects, traces and scope of data poisoning, and report and alert the Security office of the detection of data poisoning.	Q. How does the system monitor and detect data poisoning, injection of false or misleading information within the training dataset, and modifying or deleting a portion existing datasets? A. The system detects data poisoning, injecting false or misleading information within the training dataset and modifying or deleting portions of existing datasets by [means]. Q. How does the system ensure that data was not pre-poisoned prior to acquisition? A. The system detects pre-poisoning through [methods]. Q. How does system monitoring conduct thorough data quality assessments for the AI data before training AI models? A. System monitoring conducts data quality assessments with [methods]. Q. How does system monitoring detect and remove malicious or biased data injected during training AI data? A. System monitoring detects and removes malicious or biased data injected during training by [methods]. Q. How does the system monitor and detect indirect prompt injections within the AI data? A. The system monitors and detects indirect prompt injections via [methods] within the AI data. Q. How does the system monitor and detect direct prompt injections within the AI data? A. The system monitors and detects direct prompt injections via [methods] within the AI data.

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
SI-05-00	Q. What policies are in place for generating and distributing ongoing system security alerts, advisories, and directives? How does the organization distribute its own internal threat alerts? What is the speed of compliance with security directives? A. Security alerts, advisories, and directives are received from [source] on an ongoing basis. Internal alerts are also generated as needed as vulnerabilities are discovered within the environment. This information is disseminated to [personnel or roles/ elements/ external orgs]. The directives [are / are not] implemented within established time frames [or issuing organization is notified].	Q. What policies are in place for generating and distributing security alerts, advisories, and directives concerning threats to the AI platform? How does the organization distribute its own internal threat alerts? What is the speed of compliance with security directives? A. Security alerts, advisories, and directives are received from [source] on an ongoing basis. Internal alerts are also generated as needed using [tools] within the platform. This information is disseminated to [personnel or roles/ elements/ external orgs]. The directives [are / are not] implemented within established time frames [or issuing organization is notified].	Q. What policies are in place for generating and distributing security alerts, advisories, and directives concerning threats to the AI models? How does the organization distribute its own internal threat alerts? What is the speed of compliance with security directives? A. Security alerts, advisories, and directives are received from [source] on an ongoing basis. Internal alerts are also generated as needed using [tools] within the platform. This information is disseminated to [personnel or roles/ elements/ external orgs]. The directives [are / are not] implemented within established time frames [or issuing organization is notified].	Q. What policies are in place for generating and distributing security alerts, advisories, and directives concerning threats to the AI Data? How does the organization distribute its own internal threat alerts? What is the speed of compliance with security directives? A. Security alerts, advisories, and directives are received from [source] on an ongoing basis. Internal alerts are also generated as needed using [tools] within the platform. This information is disseminated to [personnel or roles/ elements/ external orgs]. The directives [are / are not] implemented within established time frames [or issuing organization is notified].
SI-07-00	Q. Which tools are in place to verify the integrity of software, firmware, and programs on the environment? What is the response plan when unauthorized changes are detected to the environment? A. The [tools] are performing integrity verification on the AI environment, which monitor [specific software/ firmware/programs]. Upon detecting unauthorized changes to software/ firmware/ programs/ component within the environment, [actions] are performed.	Q. Which tools are in place to verify the integrity of AI software and firmware on the AI platform? What is the response plan when unauthorized changes are detected to the platform? A. The [tools] are performing integrity verification on the AI platform, which monitor [software/ firmware/ tools]. Upon detecting unauthorized changes to the software/ firmware on the platform, [actions] are performed. Q. How are techniques such as adversarial training and robust optimization used to test the AI platform? A. Adversarial training and robust optimization are used via [methods] to test the AI platform.	Q. Which tools are in place to verify the integrity of AI models? What is the response plan when unauthorized changes are detected to the model? A. The [tools] performing integrity verification on the AI models, which monitor [specific elements/entire model] for unauthorized changes. Upon detecting an integrity violation to the model, [actions] are performed. Q. How are techniques such as adversarial training and robust optimization used to test the AI models? A. Adversarial training and robust optimization are used via [methods] to test the AI models.	Q. Which tools are in place to verify the integrity of AI data? What is the response plan when unauthorized changes are detected on the data? A. The [tools] performing integrity verification on the AI data, which monitor [training data/test data/data storage systems] for unauthorized changes. Upon detecting an integrity violation to the data, [actions] are performed. Q. How are techniques such as adversarial training and robust optimization used to test the AI data? A. Adversarial training and robust optimization are used via [methods] to test the AI data.

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
SI-10-00		Q. How are AI Platforms validated to ensure that they have not been tampered with? A. Answer should include things like hashes and other ways to detect changes to the data. Q. What kind of input or prompt validation is performed prior to executing queries or taking any actions? A. The [validation checks] are performed on all input provided to the platform. Q. How are prompt filters updated and reviewed to prevent bypassing in the AI Platform? A. Prompt filters are regularly updated and reviewed through [methods]. Q. What information input validation does the AI platform contain to prevent indirect prompt injection? A. The AI platform contains information [input validation checks] to prevent indirect prompt injections. Q. What information input validation does the AI platform contain to prevent direct prompt injection? A. The AI platform uses information [input validation checks] to prevent direct prompt injections. Q. How is the input provided to LLM plugins validated prior to use? A. Input is validated by using the [libraries] prior to processing the input.	Q. How are AI Models and wights matrices validated to ensure that they have not been tampered with? A. Answer should include things like hashes and other ways to detect changes to the data. Q. What kind of input or prompt validation is performed prior to executing queries or taking any actions? A. The [validation checks] are performed on all input provided to AI models. Q. How are prompt filters updated and reviewed to prevent bypass in the AI models? A. Prompt filters are regularly updated and reviewed through [methods].	Q. How is the AI training data validated to ensure that it has not been tampered with? A. Answer should include things like hashes and other ways to detect changes to the data. Q. How are prompt filters updated and reviewed to prevent bypass in the AI data or training process? A. Prompt filters are regularly updated and reviewed through [methods]. Q. What information input validation are used to prevent indirect prompt injection? A. The [validation checks] are used to prevent indirect prompt injections. Q. What information input validation are used to prevent direct prompt injection? A. The [validation checks] are used to prevent indirect prompt injections.
SI-11-00		Q. What is the behavior of error messages on the AI platform? A. Error messages are generated to provide necessary information for corrective actions without revealing exploitable information and messages are only revealed to [personnel].		

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
SI-16-00	Q. Which organization defined controls are implemented to protect system memory from unauthorized code execution? Are data-execution controls (if applicable) hardware or software enforced, and how are they combined with address space layout randomization within the environment to prevent writing backdoor executable code in prohibited memory locations? A. The system memory in the environment is protected from unauthorized code execution by using [organization defined controls]. These controls [include / don't include] hardware/software data execution prevention and address space layout randomization throughout the AI environment.	Q. Which organization defined controls are implemented to protect memory dedicated to AI platforms from unauthorized code execution? Are data-execution controls (if applicable) hardware or software enforced, and how are they combined with address space layout randomization on the platform's non-executable memory regions to prevent writing potentially backdoor executable code in prohibited memory locations? A. The AI platform memory is protected from unauthorized code execution by using [organization defined controls]. These controls [include/don't include] hardware/software data execution prevention and address space layout randomization on memory used for AI platforms.		
SI-20-00			Q. How are the AI models tainted to help determine whether the AI/ML models have been exfiltrated from the enterprise or improperly removed from the AI-enabled systems? A. The AI models are tainted by [methods].	Q. How are the AI models tainted to help determine whether the AI/ML models have been exfiltrated from the enterprise or improperly removed from the AI-enabled systems? A. The AI models are tainted by [methods].
SR-01-00		Q. How is the provenance of AI platform components assured for validity? A. The provenance of AI platform components is assured by [methods and assignment of organizationally defined parameters (ODPs)]. (Ref: SR-04)	Q. How is the provenance of AI models assured for validity? A. The provenance of AI models is assured by [methods and assignment of ODPs]. (Ref: SR-04)	Q. How is the provenance of AI datasets assured for validity? A. The provenance of AI datasets is assured by [methods and assignment of ODPs]. (Ref: SR-04)

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
SR-02-00		Q. How does the Supply Chain Risk Management Plan implement bias-awareness training and post processing techniques for the AI platform? A. The Supply Chain Risk Management Plan includes [bias-awareness training and post processing techniques].	Q. For models that are not maintained/ developed internally what is the process to ensure that changes to the model do not negatively affect the system? A. For these models, [processes] are used to review changes before they are moved into the production environment. Q. How does the Supply Chain Risk Management Plan implement bias-awareness training and post processing techniques for the AI models? A. The Supply Chain Risk Management Plan includes [bias-awareness training and post processing techniques].	Q. For AI data that is not maintained/developed internally what is the process to ensure that changes to the data do not, negatively, affect the system? A. For AI data, the [processes] are used to review changes before they are moved into the production environment. Q. How does the Supply Chain Risk Management Plan implement bias-awareness training and post processing techniques for the AI data and training? A. The Supply Chain Risk Management Plan includes [bias-awareness training and post processing techniques].
SR-03-00	Q. How is the environment protected against the introduction of unvetted public and open-source AI models, software, and tools? A. The environment is protected against the introduction of unvetted public and open-source AI models, software, and tools by [methods]	Q. How is the AI platform protected against the introduction of unvetted public and open-source AI models, software, and tools? A. The environment is protected against the introduction of unvetted public and open-source AI models, software, and tools by [methods]. Q. How are weakness in the AI platform identified and mitigated? A. Weakness are identified via [methods]. Once a weakness is identified we research available patches, test them in the development and test environments first to ensure that there are no problems with the current baseline, after they have been thoroughly tested, they are applied to the production environment.	Q. How is the chain of custody for the AI Model supply chain maintained and tracked? A. The chain of custody for AI Models is maintained by [methods].	Q. How is the chain of custody for the AI dataset supply chain maintained and tracked? A. The chain of custody for AI datasets is maintained and tracked by [methods].
SR-04-00		Q. How is the provenance of AI platform components assured for validity? A. The provenance of AI platform components is assured by [methods and assignment of ODPs].	Q. How is the provenance of AI models assured for validity? A. The provenance of AI models is assured by [methods and assignment of ODPs].	

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
SR-05-00	Q. What acquisition strategies, contract tools, and procurement methods are employed to protect against, identify, and mitigate supply chain risks in the AI environment? Which tools are techniques are employed in the environment to mitigate unauthorized production, theft, tampering, insertion of counterfeits, insertion of malicious software or backdoors, and poor development practices throughout the development life cycle. A. The strategies, tools, and procurement methods used to protect against supply chain risks are [XYZ]. Supply chain threats are mitigated in the environment by using [tools] to identify, monitor, and protect against risks.	Q. What acquisition strategies, contract tools, and procurement methods are employed to protect against, identify, and mitigate supply chain risks with AI platforms? Which tools and procedures are established to mitigate unauthorized production, theft, tampering, insertion of counterfeits, insertion of malicious software or backdoors, and poor development practices throughout the development life cycle of any AI platform. How is the organization ensuring AI developers follow the maximum level of supply chain integrity controls? A. The [strategies, tools, and procurement methods] are used to protect against supply chain risks within the AI platform. Organization uses strict assessment and verification of all AI software produced and employs [controls] throughout the entire supply chain. [Tools and techniques] are used to protect the integrity of the development process, providing protection for unauthorized production, theft, tampering, insertion of counterfeits, insertion of malicious software or backdoors, and poor development practices throughout the system development life cycle. Transparency into security and privacy practices [is/is not] promoted by the organization and implementing controls [is/is not] incentivized.		
SR-06-00		Q. What is the process to assess and review the supply chain-related risks associated with suppliers or contractors of the AI platform? A. The [processes] are used to review the suppliers or contractors of the AI platform.	Q. What is the process to assess and review the supply chain-related risks associated with suppliers or contractors of the AI models? A. The [processes] are used to review the suppliers or contractors of the AI models.	Q. What is the process to assess and review the supply chain-related risks associated with suppliers or contractors of the AI data? A. The [processes] are used to review the suppliers or contractors of the AI data.

Control ID	Environment (infrastructure, network)	AI Platform (AI components, AI software)	AI Models (ML models, LLMs)	AI Data (training data, validation data)
SR-08-00		Q. What agreements and procedures are in place with supply chain entities for the notification of compromises and potential compromises in the supply chain that can potentially adversely affect the AI platform? A. We have the [agreements] with the developers/vendors to ensure that we are promptly notified in these cases.	Q. What agreements and procedures are in place with supply chain entities for the notification of compromises and potential compromises in the supply chain that can potentially adversely affect the AI models? A. We have the [agreements] with the developers/vendors to ensure that we are promptly notified in these cases.	Q. What agreements and procedures are in place with supply chain entities for the notification of compromises and potential compromises in the supply chain that can potentially adversely affect the AI data? A. We have the [agreements] with the developers/vendors to ensure that we are promptly notified in these cases.
SR-09-00	Q. How is tamper protection provided for the components within the AI environment? What means of tamper protection is employed to protect these components during distribution and use? A. Tamper protection is provided for the AI environment and accompanying services by [methods]. Strong identification [is/ is not] used in conjunction with tamper [resistance/ detection].	Q. How is tamper protection provided for the AI system, components, and services? What means of tamper protection are employed to protect these systems during distribution and use? A. Tamper protection is provided for the AI system, components, and services by [methods]. Strong identification [is/ is not] used in combination with tamper [resistance/detection].		
SR-11-00		Q. How is assurance given that AI system components are not counterfeit? A. Assurance that AI system components are not counterfeit is given by [means].		

APPENDIX F. HIGH VALUE ASSET (HVA) OVERLAY

This table identifies controls from the **CISA HVA Overlay 2.0 (January 2021)** [8]
The CISA HVA Overlay was developed by the HVA Program Management Office (PMO) to provide technical guidance to federal civilian agencies for securing HVAs.

Source: https://www.cisa.gov/resources-tools/resources/high-value-asset-control-overlay

Table 4: System Level Controls

AC-02-00	AU-10-00	CP-07-00	PL-02-00	SC-03-00	SC-28-01
AC-02-02	AU-16-00	CP-07-03	PL-08-00	SC-03-02	SI-02-00
AC-03-00	AT-02-01	CP-09-01	PL-08-01	SC-05-00	SI-03-00
AC-03-09	CA-03-00	CP-10-04	PL-10-00	SC-05-01	SI-04-00
AC-04-00	CA-05-00	IA-02-00	PT-03-01	SC-05-02	SI-04-01
AC-05-00	CA-06-00	IA-02-01	PT-03-02	SC-05-03	SI-04-10
AC-06-00	CA-06-01	IA-02-02	RA-02-00	SC-07-00	SI-04-11
AC-06-05	CA-07-00	IA-02-12	RA-03-01	SC-07-03	SI-04-13
AC-06-07	CA-07-03	IA-03-00	RA-05-00	SC-07-05	SI-04-18
AC-17-00	CA-09-00	IA-05-00	RA-05-06	SC-07-10	SI-04-20
AC-17-02	CM-02-00	IA-05-01	RA-05-10	SC-07-11	SI-04-22
AC-20-00	CM-03-02	IR-04-02	SA-04-00	SC-07-12	SI-04-23
AU-02-00	CM-03-07	IR-04-08	SA-09-00	SC-07-14	SI-05-00
AU-06-00	CM-04-01	IR-04-10	SA-11-00	SC-07-17	SR-04-02
AU-09-00	CM-06-00	IR-05-00	SA-11-01	SC-07-21	SR-04-03
AU-09-02	CM-06-02	MP-06-00	SA-11-02	SC-07-22	SR-05-02
AU-09-03	CM-07-01	MP-06-08	SA-11-04	SC-08-00	SR-09-00
AU-09-05	CM-08-00	PE-03-00	SA-11-05	SC-18-04	SR-10-00
AU-09-06	CP-04-00	PE-03-01	SA-11-08	SC-28-00	

Table 5: Enterprise Controls

AU-06-03	CP-02-00	PM-07-00	PM-12-00	SR-06-00
AU-06-04	CP-08-05	PM-09-00	RA-03-00	
AU-06-05	IR-04-04	PM-10-00	SI-04-16	

APPENDIX G. GLOSSARY

The material for this glossary is adapted from Appendix V to House Bipartisan Task Force on Artificial Intelligence, "Report on Artificial Intelligence," 12 2024 [9]. The Office of Management and Budget (OMB) Circular No. A-130 - Security of Federal Automated Information Resources, "adequate security means security commensurate with the risk and magnitude of the harm resulting from the loss, misuse, or unauthorized access to or modification of information. This includes assuring that systems and applications used by the agency operate effectively and provide appropriate confidentiality, integrity, and availability, using cost-effective management, personnel, operational, and technical controls."

Table 6: AI Definitions

Term	Definition
Artificial Intelligence (AI)	Software systems capable of performing tasks typically expected to require human intelligence, e.g., voice recognition, image analysis, and language translation. The field of AI encompasses various subfields, including machine learning, natural language processing, and computer vision.
Machine Learning (ML)	The subfield of artificial intelligence that involves software learning and improving from data. ML algorithms can analyze large amounts of data, identify patterns in that data, and based on those patterns, make predictions or decisions without being explicitly programmed how to do so. Generally, using more data in training results in better performance on the task the software is trained for.
AI Model	A software program that receives input data, such as text, images, or numbers, and processes those inputs to produce specific types of outputs, such as predictions, recommendations, or generated content. Many AI models today use ML to "learn" how to produce outputs from inputs. The larger the model and the larger the training data set, the better the model performs.
Generative AI	AI systems that can generate new content, such as text, images, video, and music, with minimal or no human guidance on how exactly to create that content. Some

Term	Definition
	generative AI systems allow the user to specify the general nature or characteristics of the content to generate. A generative AI system is designed to produce content that is novel rather than copied from existing data. Generated content is also intended to be realistic in that it resembles human-created content. Typically, the content of the training data determines the types of content that can be generated.
Large Language Models (LLMs)	A powerful generative AI model trained in vast amounts of text data, giving it the capability to understand text and generate human-like text. LLMs are useful for a wide range of natural language processing tasks, such as chatbots, text summarization, and language translation.
Neural Network	A type of machine learning model consisting of interconnected nodes, or "artificial neurons," typically organized in one or more layers. Once the neural network is trained, the nodes cooperate to transform input data into output such as predictions, classification decisions, or generated content.
Deep Learning	An ML paradigm that uses neural networks with many layers. In general, the more layers in the neural network, the greater the performance of the neural network. Deep learning systems allow more sophisticated patterns to be recognized and more complex tasks to be performed. Deep learning has been responsible for many of the breakthroughs in AI over the past decade.
Natural Language Processing (NLP)	The subset of AI that involves processing human language, such as written text. NLP enables machines to understand, interpret, and generate human language, facilitating tasks like language translation, text summarization, and chatbots.

Term	Definition
Computer Vision	The subset of AI that involves understanding and interpreting visual information from images or videos. Computer vision allows machines to recognize objects, identify faces, and analyze various types of visual content.
Foundation Models	A type of AI model that, after training on vast amounts of data, is of general purpose to be used in a wide variety of different tasks.
Compute	The computational resources that are required to train and run AI models efficiently. It encompasses the computer hardware, memory, and other resources needed to create and use AI models. With the increasing complexity and size of AI models, compute has become a crucial resource.

APPENDIX H. ACRONYMS

AI	Artificial Intelligence
AIBOM	AI System Bills of Materials
AML	Adversarial Machine Learning
API	Application Programming Interface
AO	Authorizing Official
ATLAS	Adversarial Threat Landscape for Artificial-Intelligence Systems
CBRN	Chemical, Biological, Radiological, And Nuclear
CI/CD	Continuous Integration/Continuous Deployment
CISA	Cybersecurity and Infrastructure Security Agency
CSAM	Child Sexual Abuse Material
DOS	Denial-Of-Service
EO	Executive Order
FAR	Federal Acquisition Regulation
FISMA	Federal Information Security and Modernization Act
FTI	Federal Tax Information
GenAI	Generative Artificial Intelligence
HITL	Human-in-the-Loop
HVA	High Value Asset
ID	Identifier
I/O	Input/Output
IT	Information Technology
LLM	Large Language Model
MITM	Man-in-the-Middle
ML	Machine Learning
NCII	Nonconsensual Intimate Images
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
OCR	Optical Character Recognition
ODP	Organizationally Defined Parameter
OECD	Organization for Economic Cooperation and Development
OMB	Office of Management and Budget

PETs	Privacy Enhancing Technologies
PHI	Protected Health Information
PII	Personally Identifiable Information
PKI	Public Key Infrastructure
PMO	Program Management Office
POA&M	Plan of Action and Milestones
Q&A	Questions and Answers
RMF	Risk Management Framework
RPA	Robotic Process Automation
SAP	Security Assessment Plan
SAR	Security Assessment Report
SCA	Security Controls Assessment
SBOM	Software Bills of Materials
SME	Subject Matter Expert
SP	Special Publication
SSP	System Security Plan
USC	United States Code

APPENDIX I. REFERENCES

[1] NIST, "Special Publication 800-53 Revision 5," NIST, Gaithersburg, MD, 2020.

[2] NIST, "AI RIsk Management Framework 1.0," 1 2023. [Online]. Available: https://doi.org/10.6028/NIST.AI.100-1. [Accessed 27 1 2025].

[3] Wikipedia, "Bell-LaPadula Model," [Online]. Available: https://en.wikipedia.org/wiki/Bell%E2%80%93LaPadula_model. [Accessed 30 1 2025].

[4] OECD, "Explanatory memorandum on the updated OECD definition of an AI system," 2024. [Online]. Available: https://doi.org/10.1787/623da898-en. [Accessed 2 Jan 2025].

[5] NIST, "NIST Special Publication 800-30 R1 "Guide for Conducting Risk Assessments"," NIST, Gaithersburg, MD, 2012.

[6] The MITRE Corporation, "MITRE ATLAS," 20 12 2024. [Online]. Available: https://atlas.mitre.org.

[7] NIST, "NIST Risk Management Framework," [Online]. Available: https://nist.gov/rmf. [Accessed 29 1 2025].

[8] CISA, 1 2021. [Online]. Available: https://www.cisa.gov/resources-tools/resources/high-value-asset-control-overlay. [Accessed 25 12 2024].

[9] House Bipartisan Task Force on Artificial Intelligence, "Report on Artificial Intelligence," 12 2024. [Online]. Available: https://science.house.gov/2024/12/house-bipartisan-task-force-on-artificial-intelligence-delivers-report. [Accessed 2/20/2025].

Data Rights Legend