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vulnerabilities an attacker could exploit to gain access to a system and its models need to be anticipated, including outdated or unpatched software components, weak or improperly enforced access control, and poor asset protection management practices.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	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
	configuration management policies and practices for the AI-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	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
						system, firmware, or tool updates 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	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		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	AML.T0050 "Command and Scripting Interpreter"

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
code or other malicious activities	AI models use procedural representations of knowledge and models (e.g., as in a rule-based system). These 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.					possible that unknown flaws in that logic may be exploited by 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	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
	measuring known risks, errors, incidents, or negative impacts. Metrics should also account for known AI design and implementation failure modes associated with properties 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,	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

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
	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 databases, other organizations with similar systems, and system users and stakeholders.						Analysis", AML.T0006 "Active Scanning"
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		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
	statistical properties, stringent access controls for data and models, and vigilant monitoring of system performance.						
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"

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
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 prompts that do not appear to be malicious.	AML.T0051 "Prompt Injection"
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"

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
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 inputs and secure access control guidelines, to mitigate this potential vulnerability.	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"
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		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	AML.T0051 "Prompt Injection"

AI Threats	AI Concerns	Environment	AI Platform	AI Models	AI Data	Residual Risk	Related ATLAS ID
	content manipulation, attackers can undermine trust in the automated document processing system.					data" and "Supply chain and life cycle infiltrations and unvetted changes of training and operational data" AI Concerns for additional controls to consider.	
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 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.	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 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.	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
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 & 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"
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"