GUIDELINES ON **SECURING AI SYSTEMS**

CYBER SECURITY AGENCY OF SINGAPORE

OCTOBER 2024

CSA SINGAPORE

TABLE OF CONTENTS

1. INTRODUCTION

Artificial Intelligence (AI) poses benefits for economy, society, and national security. It has the potential to drive efficiency and innovation in almost every sector – from commerce and healthcare to transportation and cybersecurity.

To reap the benefits of AI, users must have confidence that the AI will behave as designed, and outcomes are safe and secure. However, in addition to safety risks, AI systems can be vulnerable to adversarial attacks, where malicious actors intentionally manipulate or deceive the AI system. **The adoption of AI can introduce or exacerbate existing cybersecurity risks to enterprise systems. These can lead to risks such as data leakage or data breaches, or result in harmful or otherwise undesired model outcomes.**

As such, **as a key principle, AI should be secure by design and secure by default**, as with all software systems. This will enable system owners to manage security risks upstream. This will complement other controls and mitigation strategies that system owners may take to address the safety of AI, and other attendant considerations such as fairness or transparency, which are not addressed here.

The Cyber Security Agency of Singapore (CSA) has developed **Guidelines on Securing AI Systems** for system owners to secure the use of AI throughout its lifecycle. As AI is increasingly integrated into enterprise systems, security should be considered holistically at the system level. As such, these guidelines should be used alongside existing security best practices and requirements for IT environments. While these guidelines are not mandatory, we strongly encourage system owners to consider these key principles, so that they can make informed decisions on their adoption of AI vis-à-vis the potential risks.

AI security is a developing field of work, and mitigation controls continue to evolve. As such, CSA is also collaborating with AI and cybersecurity practitioners on the **Companion Guide on Securing AI Systems**. This is intended as a community-driven resource, and the Companion Guide complements the Guidelines as a useful reference containing practical measures and controls that system owners may consider as part of adopting the Guidelines, depending on their use case. The Companion Guide is not mandatory, prescriptive, or exhaustive. As the field of AI security continues to evolve rapidly, the Companion Guide will be updated to account for material developments in this space.

1.1. PURPOSE AND SCOPE OF THIS DOCUMENT

Purpose

These guidelines are designed to support systems owners that are adopting, or considering the adoption of AI systems. It identifies potential security risks associated with the use of AI and sets out guidelines for mitigating security risks at each stage of the AI lifecycle.

This document can be read together with the Companion Guide on Securing AI Systems, which provides an informative compilation of practical security control measures, that system owners may consider in implementing these guidelines.

Scope

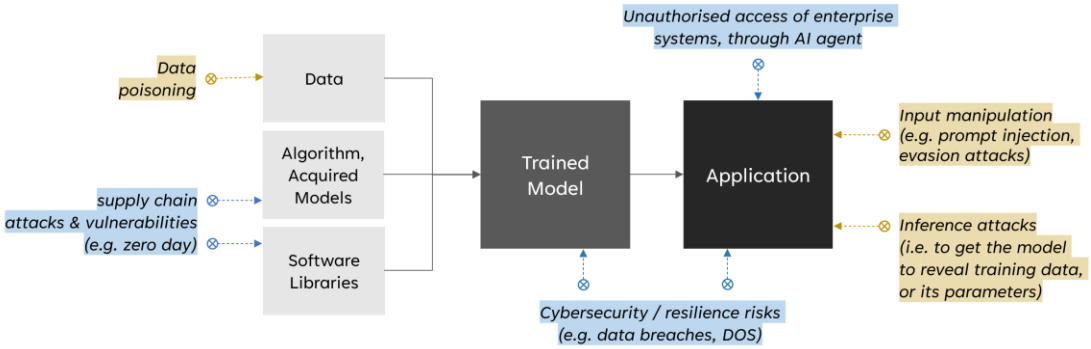
These guidelines address the cybersecurity risks to AI systems. It does not seek to address AI safety, or other common attendant considerations for AI such as fairness, transparency or inclusion, or cybersecurity risks introduced by AI systems, although some of the recommended actions may overlap. It also does not address the misuse of AI in cyberattacks (AI-enabled malware), mis/disinformation, and scams (deepfakes).

2. UNDERSTANDING AI THREATS

AI is a type of software system, and is itself vulnerable to cyber threats, while also posing a new attack surface for the broader enterprise system that it is integrated to, or interfaces with. As such, securing AI is in addition to practising good 'classical' cybersecurity hygiene.

Securing an AI system introduces new challenges that may be unfamiliar in traditional IT systems. In addition to classical cybersecurity risks, the AI itself is vulnerable to novel attacks such as Adversarial Machine Learning (ML) that set out to distort the model's behaviour. For more details on the security threats to AI, refer to **Annex A**.

Figure 1. Classical and AI-specific risks of AI systems– diagram adapted from OWASP[1]



[1] Threats overview - https://owaspai.org/docs/ai_security_overview/

CLASSICAL CYBERSECURITY RISKS TO AI SYSTEMS

AI systems require vast amounts of data for training; some also require importing external models and libraries. If inadequately secured, AI systems can be undermined by **supply chain attacks**, or may be **susceptible to intrusion or unauthorised access**, through vulnerabilities in the AI model or the underlying IT infrastructure. In addition, **organisations and users risk losing the ability to access and use AI tools** if there are disruptions to cloud services, data centre operations, or other digital infrastructure (e.g. through Denial of Service attacks), this could in turn **disable systems that depend on AI tools to function**.

ADVERSARIAL MACHINE LEARNING

Malicious actors may use novel Adversarial ML techniques to attack AI models and data, influencing machine learning models to produce inaccurate, biased, or harmful output; and/or reveal confidential information. Adversarial ML[2] attacks include: *data poisoning* (injecting malicious or corrupted data into training data sets) or *evasion attacks* (on trained models) to **distort outcomes**, *inference attacks* or *extraction attacks* (probing the model) to **expose sensitive or restricted data,** or to **steal the model.**

[2] A Taxonomy and Terminology of Attacks and Mitigations https://csrc.nist.gov/pubs/ai/100/2/e2023/final. The MITRE ATLAS is a useful reference to understand and situate classical cybersecurity risks from Adversarial ML.

3. SECURING AI

The security of AI is a widely cited concern, but this field of work is still relatively nascent. While practitioners continue to grow the body of research and resources on the security threats to AI, these guidelines lay out key considerations that system owners should take to support secure adoption of AI. Given the rapid speed of AI development, system owners should continue to apprise themselves on the latest developments in AI security, and refresh their risk management strategies accordingly.

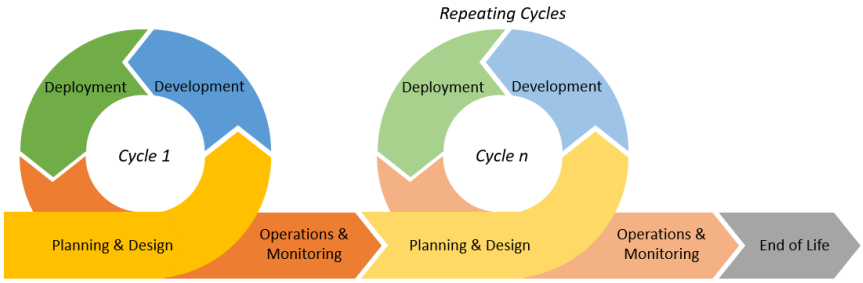
3.1. TAKE A LIFECYCLE APPROACH

There are **five key stages** – Planning and Design, Development, Deployment, Operations and Maintenance, and End of Life.

As with good cybersecurity practice, CSA recommends that system owners take a lifecycle approach to consider security risks. Hardening only the AI model is insufficient to ensure a holistic defence against AI related threats. All stakeholders involved across the lifecycle of an AI system should seek to better understand the security threats and their potential impact on the desired outcomes of the AI system, and what decisions or trade-offs will need to be made.

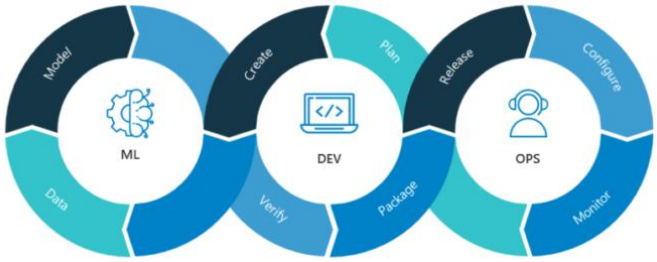
The AI lifecycle represents the iterative process of designing an AI solution to meet a business or operational need. As such, system owners will likely revisit the planning and design, development, and deployment steps in the lifecycle many times in the delivery of an AI solution.

Figure 2: AI System Development Lifecycle (AI SDLC)



Some organisations may have implemented the Machine Learning Operations (ML Ops) pipeline, which may not map exactly to the AI SDLC. Nonetheless, ML Ops teams that run a dev ops pipeline comprising ML Design, Development and Operation stages (similar to Figure 3), will find the guidelines across the AI SDLC's stages of Planning & Design, Development, Deployment and Operations relevant.

Figure 3: Example of ML-DevOps (source: Nvidia blog)



3.2. START WITH A RISK ASSESSMENT

Given the diversity of AI use cases, there is no one-size-fits-all solution to implementing security. As such, effective cybersecurity starts with conducting a risk assessment. This will enable organisations to identify potential risks, priorities, and subsequently, the appropriate risk management strategies.

A fundamental difference between AI and traditional software is that while traditional software relies on static rules and explicit programming, AI uses machine learning and neural networks to autonomously learn and make decisions without the need for detailed instructions for each task. As such, organisations should consider conducting risk assessments more frequently than for conventional systems, even if they generally base their risk assessment approach on existing governance and policies. These assessments may also be supplemented by continuous monitoring and a strong feedback loop.

We recommend these **four steps** to tailor a systematic defence plan that best addresses your organisation's highest priority risks – protecting the things you care about the most.

STEP 1

Conduct risk assessment, focusing on security risks to AI systems

Conduct a risk assessment, focusing on security risks related to AI systems, either based on best practices or your organisation's existing Enterprise Risk Assessment/Management Framework.

Risk assessment can be done with reference to CSA published guides, if applicable:

- Guide To Cyber Threat Modelling
- Guide To Conducting Cybersecurity Risk Assessment for Critical Information Infrastructure

STEP 2

Prioritise areas to address based on risk/impact/resources

Prioritise which risks to address, based on risk level, impact, and available resources.

STEP 3

Identify and implement the relevant actions to secure the AI system

Identify relevant actions and control measures to secure the AI system, such as by referencing those outlined in the **Companion Guide on Securing AI Systems** and implement these across the AI life cycle.

STEP 4

Evaluate residual risks for mitigation or acceptance

Evaluate the residual risk after implementing security measures for the AI system to inform decisions about accepting or addressing residual risks.

3.3. GUIDELINES FOR SECURING AI SYSTEMS

These guidelines apply across the various lifecycle stages of the AI system. System owners should read these as key issues to consider in securing their adoption of AI. In view of the diversity of use cases and developments in AI security, these guidelines do not provide prescriptive controls or requirements.

System owners should apply these to their specific context, and can reference the Companion Guide to Securing AI systems for potential controls.

1. PLANNING AND DESIGN

1.1. Raise awareness and competency on security risks

Organisations should understand the potential security risks posed by AI, in order to make informed decisions about adoption. Provide adequate training and guidance on the security risks of AI to all personnel, including developers, system owners and senior leaders.

1.2. Conduct security risk assessments

Risk management strategies should be informed by a security risk assessment, which will help to determine key risks and priorities. Apply a holistic process to model threats and risks to an AI system, in accordance with relevant industry standards/best practices.

2. DEVELOPMENT

2.1. Secure the supply chain

The AI supply chain includes (but is not limited to) the training data, models, APIs, and software libraries. Each of these components may introduce new vulnerabilities (e.g, models may carry malware encoded as model parameters that could enable attackers to extract and inject malicious software onto user machines). Assess and monitor potential security risks of the AI system's supply chain across its life cycle. Ensure that suppliers adhere to security policies and internationally recognised standards, or that risks are otherwise appropriately managed. Consider evaluating supply chain components (e.g. through Software Bills of Material [SBOM], code checking, or against vulnerability databases).

2.2. Consider security benefits and trade-offs when selecting the appropriate model to use

Different AI models (e.g. machine learning, deep learning, generative) pose unique characteristics and risks (e.g. LLMs can be vulnerable to input manipulation attacks) and as such require different security measures. When developing or selecting an appropriate AI model for your system, consider factors which may affect its security (such as complexity, explainability, interpretability, and sensitivity of training data).

2.3. Identify, track and protect AI-related assets

As AI systems become increasingly integrated into business operations, they will become part of an organisation's strategic assets and should be secured accordingly. Otherwise, sensitive data, intellectual property and organisational assets are at risk of potential threats and breaches. Understand the value of AI-related assets, including models, data, prompts, logs and assessments. Have processes to track, authenticate, version control, and secure assets.

2.4. Secure the AI development environment

AI models require access to large amounts of training data, and an insecure development environment can introduce risks of data breaches (e.g. exposure of Personally Identifiable Information or confidential business information). Insecure development can also make AI models vulnerable to attacks (e.g. poisoning) that result in compromised model behaviour, or expose models and other intellectual property to theft, unauthorised replication or misuse. Apply standard infrastructure security principles, such as implementing appropriate access controls and logging/monitoring, segregation of environments, and secure-by-default configurations.

3. DEPLOYMENT

3.1. Secure the deployment infrastructure and environment of AI systems

Similar considerations as with 2.4 “Secure the AI development environment”. Apply standard infrastructure security principles, such as access controls and logging/monitoring, segregation of environments, secure-by-default configurations, and firewalls.

3.2. Establish incident management procedures

AI systems are complex and adaptive, and this can sometimes result in unpredictable behaviour. Given the diversity in AI use cases, incidents can range from minor issues such as malfunctioning chat bots to critical outcomes such as disruption in the operation of critical infrastructure. System owners should put in place appropriate incident response, escalation and remediation plans.

3.3. Release AI systems responsibly

AI systems can be vulnerable to the risks described above, including misuse, data breaches, and model manipulation. These have impact on the trust and confidence of users, and may have reputational implications for organisations. A good practice is to release models, applications or systems only after subjecting them to appropriate and effective security checks and evaluation.

4. OPERATIONS AND MAINTENANCE

4.1. Monitor AI system inputs

AI systems are dynamic and adaptive to input. There have already been real-life incidents, in which users/ attackers have deliberately crafted input to trick AI systems into making incorrect or unintended decisions. AI system owners may wish to monitor and log inputs to the AI system, such as queries, prompts and requests, as third-party providers may not do so due to privacy reasons. Proper logging allows for compliance, audit, investigation and remediation.

4.2. Monitor AI system outputs and behaviour

AI systems can break or degrade in production phase. Monitoring models after deployment will make sure that they are performing as intended, and alert system owners to potential issues (whether caused by adversarial attacks or otherwise). Operators should monitor for anomalous behaviour that might indicate intrusions, compromise, or data drift.

4.3. Adopt a secure-by-design approach to updates and continuous learning

Changes to the data and model can lead to changes in behaviour. System owners should ensure that risks associated to model updates have been considered and appropriately managed.

4.4. Establish a vulnerability disclosure process

Even with monitoring mechanisms in place, the adaptive nature of AI can make it challenging to detect attacks and unintended behaviour. There should be a feedback process for users to share any findings of concern, which might uncover potential vulnerabilities to the system.

5. END OF LIFE

5.1. Ensure proper data and model disposal

As models are trained on large amounts of training data (incl. potentially confidential information), improper disposal can lead to incidents such as data breaches. There should be proper and secure disposal/destruction of data and models in accordance with relevant industry standards or regulations.

GLOSSARY

Term	Brief description
AI system	Artificial Intelligence. A machine-based system that for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.
Adversarial Machine Learning	The process of extracting information about the behaviour and characteristics of an ML system and/or learning how to manipulate the inputs into an ML system in order to obtain a preferred outcome.
Anomaly Detection	The identification of observations, events or data points that deviate from what is usual, standard, or expected, making them inconsistent with the rest of data.
API	Application Programming Interface. A set of protocols that determine how two software applications will interact with each other.
Backdoor attack	A backdoor attack is when an attacker subtly alters AI models during training, causing unintended behaviour under certain triggers.
Chatbot	A software application that is designed to imitate human conversation through text or voice commands
Computer Vision	An interdisciplinary field of science and technology that focuses on how computers can gain understanding from images and videos.
Data Breach	Data Breach occurs when a threat actor gains unauthorised access to sensitive/confidential data.
Data Integrity	The property that data has not been altered in an unauthorised manner. Data integrity covers data in storage, during processing, and while in transit.

Data Leakage	Unintentional exposure of sensitive, protected, or confidential information outside its intended environment.
Data Loss Prevention	A system's ability to identify, monitor, and protect data in use (e.g., endpoint actions), data in motion (e.g., network actions), and data at rest (e.g., data storage) through deep packet content inspection, and contextual security analysis of transaction (e.g., attributes of originator, data object, medium, timing, recipient/destination, etc.) within a centralised management framework.
Data Poisoning	Control a model with training data modifications.
Data Science	An interdisciplinary field of technology that uses algorithms and processes to gather and analyse large amounts of data to uncover patterns and insights that inform business decisions.
Deep Learning	A function of AI that imitates the human brain by learning from how it structures and processes information to make decisions. Instead of relying on an algorithm that can only perform one specific task, this subset of machine learning can learn from unstructured data without supervision.
Defence-in-Depth	Defence in depth is a strategy that leverages multiple security measures to protect an organization's assets. The thinking is that if one line of defence is compromised, additional layers exist as a backup to ensure that threats are stopped along the way.
Evasion attack	Crafting input to AI in order to mislead it into performing its task incorrectly.
Extraction attack	Copy or steal an AI model by appropriately sampling the input space and observing outputs to build a surrogate model that behaves similarly.
Generative AI	A type of machine learning that focuses on creating new data, including text, video, code and images. A generative AI system is trained using large amounts of data, so that it can find patterns for generating new content.
Guardrails	Restrictions and rules placed on AI systems to make sure that they handle data appropriately and don't generate unethical content.

Hallucination	An incorrect response from an AI system, or false information in an output that is presented as factual information.
Image Recognition	Image recognition is the process of identifying an object, person, place, or text in an image or video.
LLM	Large Language Model. A type of AI model that processes and generates human-like text. LLMs are specifically trained on large data sets of natural language to generate human-like output.
ML	Machine Learning. A subset of AI that incorporates aspects of computer science, mathematics, and coding. Machine learning focuses on developing algorithms and models that can learn from data, and make predictions and decisions about new data.
Membership Inference attack	Data privacy attacks to determine if a data sample was part of the training set of a machine learning model.
NLP	Natural Language Processing. A subset of AI that enables computers to understand spoken and written human language. NLP enables features like text and speech recognition on devices.
Neural Network	A deep learning technique designed to resemble the human brain's structure. Neural networks require large data sets to perform calculations and create outputs, which enables features like speech and vision recognition.
Overfitting	Occurs in machine learning training when the algorithm can only work on specific examples within the training data. A typical functioning AI model should be able to generalise patterns in the data to tackle new tasks.
Prompt	A prompt is a natural language input that a user feeds to an AI system in order to get a result or output.
Reinforcement Learning	A type of machine learning in which an algorithm learns by interacting with its environment and then is either rewarded or penalised based on its actions.

SDLC	Software Development Life Cycle The process of integrating security considerations and practices into the various stages of software development. This integration is essential to ensure that software is secure from the design phase through deployment and maintenance.
Training data	Training data is the information or examples given to an AI system to enable it to learn, find patterns, and create new content.

ANNEX A

UNDERSTANDING AI THREATS

Adversarial threats are caused by threat actors with deliberate intention to cause harm. Typically, these threat actors are referred to as attackers or adversaries.

To understand these threats, system owners can refer to resources such as the OWASP Top 10 for Large Language Model Applications, or OWASP Machine Learning Security Top 10, or the MITRE ATLAS™ (Adversarial Threat Landscape for Artificial-Intelligence Systems). The MITRE ATLAS in particular provides **a structured knowledge base** for AI and cybersecurity professionals to understand and defend against AI cyber threats. It compiles adversary tactics, techniques, and case studies for AI systems based on real-world observations, demonstrations from ML red teams and security groups, as well as state-of-the-possible from academic research.

> Any attempt to secure an AI system should be on top of the 'traditional' good cybersecurity hygiene, such as implementing the principle of least privileges, multi-factor authentication, continuous security monitoring and auditing.

The ATLAS[3] Matrix (see **Table A1)** covers 2 types of adversarial 'techniques'.

- Techniques specific to AI/ML systems (indicated in orange boxes), and
- Techniques that are conventional cybersecurity offensive techniques, but applicable to both AI and non-AI systems and come directly from the MITRE Enterprise ATT&CK Matrix (indicated in white boxes).

System owners should continue to build their awareness of security threats using these resources, to better understand emerging risks that may have implications on their adoption of AI. As this space continues to evolve, such resources will aid both AI and cyber teams in their security risk assessment and management activities.

[3] MITRE ALTAS Framework: https://atlas.mitre.org/. It leverages the same core principles and structure of the well-known MITRE ATT&CK (Adversarial Tactics, Techniques, and Common Knowledge) framework, which is widely used by cyber defenders to map the terminologies of cybersecurity attacks. The ATLAS adapts these to the unique context of AI systems and potential adversarial attacks.

Table A1: MITRE ATLAS Matrix

Reconnaissance	Search for Victim's Publicly Available Research Materials	Search for Publicly Available Adversarial Vulnerability Analysis	Search Victim-Owned Websites	Search Application Repositories	Active Scanning		
Resource Development	Acquire Public ML Artifacts	Obtain Capabilities	Develop Capabilities	Acquire Infrastructure	Publish Poison Datasets	Poison Training Data	Establish Accounts
Initial Access	ML Supply Chain Compromise	Valid Accounts	Evade ML Model	Exploit Public-Facing Application	LLM Prompt Injection	Phishing	
ML Model Access	ML Model Inference API Access	ML-Enabled Product or Service	Physical Environment Access	Full ML Model Access			
Execution	User Execution	Command and Scripting Interpreter	LLM Plugin Compromise				
Persistence	Poison Training Data	Backdoor ML Model	LLM Prompt Injection				
Privilege Escalation	LLM Prompt Injection	LLM Plugin Compromise	LLM Jailbreak				
Defence Evasion	Evade ML Model	LLM Prompt Injection	LLM Jailbreak				
Credential Access	Unsecured Credentials						
Discovery	Discover ML Model Ontology	Discover ML Model Family	Discover ML Artifacts	LLM Meta Prompt Extraction			
Collection	ML Artifact Collection	Data from Information Repositories	Data from Local System				
ML Attack Staging	Create Proxy ML Model	Backdoor ML Model	Verify Attack	Craft Adversarial Data			
Exfiltration	Exfiltration via ML Inference API	Exfiltration via Cyber Means	LLM Meta Prompt Extraction	LLM Data Leakage			
Impact	Evade ML Model	Denial of ML Service	Spamming ML System with Chaff Data	Erode ML Model Integrity	Cost Harvesting	External Harms	