COMPANION GUIDE ON **SECURING AI SYSTEMS**

OCTOBER 2024

This document is meant as a community-driven resource with contribution from the AI and cybersecurity practitioner communities. It puts together available and practical mitigation measures and practices. This document is intended for informational purposes only and is not mandatory, prescriptive nor exhaustive.

System owners should refer to this Companion Guide as a resource, alongside other available resources in observing the Cyber Security Agency of Singapore's (CSA) Guidelines on Securing AI systems. This Companion Guide is a living document that will be continually updated to address material developments in this space.

DEVELOPED IN CONSULTATION WITH

This document is published by the CSA, developed with partners across the AI and Cyber communities, including:

Accenture
Artificial Intelligence Technical Committee, Information Technology Standards Committee (AITC, ITSC)
Association of Information Security Professionals (AiSP)'s Artificial Intelligence Special Interest Group (AI SIG)
Alibaba Cloud (Singapore) Pte Ltd
Amazon Web Services Singapore
Amaris.AI
BSA | The Software Alliance
Ensign InfoSecurity Pte Ltd
F5
Google Asia Pacific Pte Ltd
Huawei International Pte Ltd
Information Technology Industry Council (ITI)
Kaspersky Lab Singapore Pte Ltd
KPMG in Singapore
Microsoft Singapore
Pricewaterhouse Coopers Risk Services Pte Ltd
Rajah & Tann Cybersecurity Pte. Ltd.
Rajah & Tann Technologies Pte. Ltd.
Resaro.AI
US-ASEAN Business Council
AI & Cyber practitioners across the Singapore Government

DISCLAIMER

These organisations provided views and suggestions on the security controls, descriptions of the security control(s), and technical implementations included in this Companion Guide. CSA and its partners shall not be liable for any inaccuracies, errors and/or omissions contained herein nor for any losses or damages of any kind (including any loss of profits, business, goodwill, or reputation, and/or any special, incidental, or consequential damages) in connection with any use of this Companion Guide. Organisations are advised to consider how to apply the controls within to their specific circumstances, in addition to other additional measures relevant to their needs.

VERSION HISTORY

VERSION	DATE RELEASED	REMARKS
0.1	29 July 2024	Draft release of Companion Guide
1.0	15 Oct 2024	First release

TABLE OF CONTENTS

QUICK REFERENCE TABLE

Stakeholders in specific roles may use the following table to quickly reference relevant controls in section "2.2 IDENTIFY THE RELEVANT MEASURES/CONTROLS"

The roles defined below are included to guide understanding of this document and are not intended to be authoritative.

Decision Makers:

Responsible for overseeing the strategic and operational aspects of AI implementation for the AI system. They are responsible for setting the vision and goals for AI initiatives, defining product requirements, allocating resources, ensuring compliance, and evaluating risks and benefits.

Roles Included: Product Manager, Project Manager

AI Practitioners:

Responsible for the practical application (i.e. designing, developing, and implementing AI models and solutions) across the life cycle. This includes collecting or procuring and analysing data that goes into systems, building the AI system architecture and infrastructure, building and optimising the AI system to deliver the required functions, as well as conducting rigorous testing and validation of AI models to ensure their accuracy, reliability, and performance. In cases where the AI system utilizes a third-party AI system, AI Practitioners include the third-party provider responsible for these activities, e.g. as contracted through a Service Level Agreement (SLA). AI practitioners would be in charge of implementing the required controls across the entire system.

Roles Included: AI/ML Developer, AI/ML Engineer, Data Scientist

Cybersecurity Practitioners:

Responsible for ensuring the security and integrity of AI systems. This includes implementing security measures to protect AI systems in collaboration with AI Practitioners, monitoring for potential threats, ensuring compliance with cybersecurity regulations.

Roles Included: IT Security Practitioner, Cybersecurity Expert

The following sections may be relevant to **Decision Makers**:	The following sections may be relevant to **AI Practitioners**:	The following sections may be relevant to **Cybersecurity Practitioners**:
1.1 Team competency on threats and risks 1.2 Conduct security risk assessment	1.1 Team competency on threats and risks 1.2 Conduct security risk assessment	1.1 Team competency on threats and risks 1.2 Conduct security risk assessment
2.1 Secure the supply chain	2.1 Secure the supply chain 2.2 Model development 2.3 Identify, track and protect assets 2.4 Secure the AI development environment	2.1 Secure the supply chain 2.3 Identify, track and protect assets 2.4 Secure the AI development environment
3.1 Secure the deployment infrastructure and environment 3.2 Have well developed incident management procedures	3.1 Secure the deployment infrastructure and environment 3.2 Have well developed incident management procedures 3.3 Release AI responsibly	3.1 Secure the deployment infrastructure and environment 3.2 Have well developed incident management procedures 3.3 Release AI responsibly
4.4 Vulnerability disclosure process	4.1 Monitor system outputs and behaviour 4.2 Monitor system inputs 4.3 Have a secure-by-design approach to updates and continuous learning 4.4 Vulnerability disclosure process	4.1 Monitor system outputs and behaviour 4.2 Monitor system inputs 4.4 Vulnerability disclosure process
5.1 Proper data and model disposal	5.1 Proper data and model disposal	5.1 Proper data and model disposal

Table 1: User Quick Reference Table

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, users must have confidence that the AI will behave as designed, and outcomes are safe, secure, and responsible manner. 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, unfair, or otherwise undesired model outcomes.** As such, the Cyber Security Agency of Singapore (CSA) has released the **Guidelines on Securing AI Systems** to advise **system owners on securing their adoption of AI.**

Nonetheless, **AI security is a developing field of study, and understanding of the security risks associated with AI continues to evolve internationally. As such, government agencies, our industry partners, AI and cybersecurity practitioners have put together this Companion Guide on Securing AI Systems**. The Companion Guide is a community-driven resource. It puts together available and practical mitigation measures and practices, drawing from industry and academia, as well as key resources such as the MITRE ATLAS database and OWASP Top 10 for Machine Learning and for Generative AI. System owners can refer to this Companion Guide as a resource, alongside other available resources in observing the Guidelines. This document is **intended for informational purposes only and is not mandatory, prescriptive nor exhaustive**. They should not be construed as comprehensive guidance or definitive recommendations.

> This Companion Guide is a living document that will be continually updated to address material developments in this space.

1.1. PURPOSE AND SCOPE

Purpose

This Companion Guide curates practical treatment measures and controls that system owners of AI systems may consider to secure their adoption of AI systems. **These measures/controls are voluntary, and not all the treatment measures/controls listed in this Companion Guide will be directly applicable** to all organisations or environments. Organisations may also be at different stages of development and release (e.g. POC, pilot, beta release). Organisations should consider relevance to their use cases/applications.

The Companion Guide is also meant as a resource to support system owners in addressing CSA's **Guidelines on Securing AI Systems**.

Scope

The controls within the Companion Guide primarily address the cybersecurity risks to AI systems. It does not 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 controls may overlap. It also does not cover the misuse of AI in cyberattacks (AI-enabled malware), mis/disinformation, and scams (deepfakes).

2. USING THE COMPANION GUIDE

The Companion Guide puts together potential treatment measures/ controls that can support secure adoption of AI. However, not all of these controls might apply to your organisation.

Our goal is to put together a comprehensive set of treatment measures that system owners can consider for their respective use cases across the AI system lifecycle. These span the categories of People, Process and Technology.

There are two categories of measures/controls: (1) based on classical cybersecurity practices, which continue to be relevant to AI systems; and (2) others unique to AI systems. Measures/controls marked with an asterisk (*) next to their number indicates that they are unique to AI systems.

Each measure/control is **designed to be used independently**, to offer flexibility in customising which measures to evaluate and what mitigations to adopt, based on the specific needs of your organisation.

2.1. START WITH A RISK ASSESSMENT

As in CSA's Guidelines for Securing AI Systems, system owners should consider starting with a risk assessment. This will enable organisations to identify potential risks, priorities, and subsequently, the appropriate risk management strategies (including what measures and controls are appropriate).

You can consider the following **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.

2.2. IDENTIFY THE RELEVANT MEASURES/CONTROLS

Based on the risk assessment, system owners can identify the relevant measures/controls from the following tables. Each treatment measure/ control plays a different role, and should be assessed for relevance and priority in addressing the security risks specific to your AI system and context (Refer to section "2.1 START WITH A RISK ASSESSMENT").

Checkboxes are included to help users of this document to keep track of which measures/controls are applicable, and have (or have not) been implemented.

Related risks and Associated MITRE ATLAS Techniques[1] indicated serve as examples and are not exhaustive. They might differ based on your organisation's use case.

Example implementations are included for each measure/control as a more tangible elaboration on how they can be applied. These are also not exhaustive.

Additional **references and resources** are provided for users of this document to obtain further details on applying the treatment measure/control if required.

Asterisks (*) indicate measures/controls that are unique to AI systems (those without an asterisk indicate more classical cyber practices).

[1] MITRE ATLAS Framework offer a structured way to understand cyber threats in relation to AI systems (see Annex A)

2.2.1. PLANNING AND DESIGN

1.1	**Raise awareness and competency on security risks** Security is everyone's responsibility. Staff are provided with proper training and guidance.						
	Suggested Treatment Measures/Controls for consideration	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
1.1.1*	Ensure system owners and senior leaders understand threats to secure AI and their mitigations. **Responsible parties:** Decision Makers				• Incidents occurring due to poor cyber hygiene and/or knowledge	Attending seminars on AI threats, policies, and compliance and get exposed to case studies to appreciate the many AI potential and associated risks. Internal workshops and eLearning courses can inform employees on AI basics, responsible use, and relevant regulations. Integrate regular security training as part of the company's AI innovation training for a balanced approach. Online resources, e.g. electronic newsletters and YouTube videos could provide a means to track AI security developments that are emerging almost daily. Documentary evidence that team members have relevant security knowledge and training. These can include, where applicable: • Training records • Attendance records • Assessments • Certifications Establish the right cross-functional team to ensure that security, risk, and compliance considerations are included from the start.	• Principles for the Security of Machine Learning (UK NCSC) • Secure by Design - Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Secure by Design Software • Failure modes in Machine Learning (Microsoft) • OWASP AI Exchange • Advisory Guidelines on use of Personal Data in AI Recommendation and Decision Systems
1.1.2*	Provide guidance to staff on Security by Design and Security by Default principles as well as unique AI security risks and failure modes as part of InfoSec training. e.g. LLM security matters, common AI weaknesses and attacks. **Responsible parties:** Decision Makers, AI Practitioners, Cybersecurity Practitioners						
1.1.3	Train developers in secure coding practices and good practices for the AI lifecycle. **Responsible parties:** Decision Makers, AI Practitioners				• Code vulnerabilities that could be exploited		

1.2	**Conduct security risk assessments** Apply a holistic process to model threats to the system.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
1.2.1*	Understand AI governance and legal requirements, the impact to the system, users, organisation, if an AI component is compromised or has unexpected behaviour or there is an attack that affected AI privacy. Plan for an attack and its mitigation, using the principles of CIA. **Responsible parties:** Decision Makers, AI Practitioners, Cybersecurity Practitioners				• No triage, leading to confusion and locked or overloaded resources in the event of an AI security incident • Slow incident response, leading to large damage done • Slow remediation, leading to prolonged operational outage • Slow response means that attackers could do more damage, cover their tracks e.g. using anti-forensics	Perform a security risk assessment to determine the consequences and impact to the various stakeholders, and if the AI component does not behave as intended. Understand the AI inventory of systems used and their implications and interactions.	• Reference the case studies in this document. • Singapore Model Governance Framework for Generative AI • NIST AI Risk Management Framework • ISO 31000: Risk Management • MITRE ATLAS • NCSC Risk Management Guidance • OWASP Threat Modelling • OWASP Machine Learning Security Top Ten • Threats to AI using Microsoft STRIDE • Advisory Guidelines on use of Personal Data in AI Recommendation and Decision Systems • Model Artificial Intelligence Governance Framework
1.2.2*	Assess AI-related attacks and implement mitigating steps. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners					Having/Developing a play book and AI incident handling procedures that will shorten the time to remediate and reduce resources wasted on unnecessary steps. Document the decision-making process of assessing potential AI threats and possible attack surfaces, as well as steps to mitigate these threats. This can be done through a threat risk assessment. Project risks may extend beyond security, e.g. newer AI models could obsolete	

1.2	**Conduct security risk assessments** Apply a holistic process to model threats to the system.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
						the entire use case and business assumptions.	
1.2.3	Conduct a risk assessment in accordance with the relevant industry standards/best practices. **Responsible parties:** Decision Makers, AI Practitioners, Cybersecurity Practitioners				• Failure to comply with industry standards/best practices may lead to insufficient, inefficient or ineffective mitigations	Refer to the industry standards and best practices when performing risk assessment.	

2.2.2. DEVELOPMENT

2.1	**Secure the Supply Chain** Assess and monitor the security of the supply chain across the system's life cycle.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
2.1.1	Implement Secure Coding and Development Lifecycle. **Responsible parties:** Decision Makers, AI Practitioners				• Introduction of bugs, vulnerabilities or unwanted and malicious active content, such as AI poisoning and model backdoors Associated MITRE ATLAS Techniques: • AML.T0018.000 Backdoor ML Model • AML.T0020.000 Poison Training Data • AML.T0010 ML Supply Chain Compromise	Adopt Security by Design. Apply software development lifecycle (SDLC) process. Use software development tools to check for insecure coding practices. Consider implementing zero trust principles in system design.	• CSA Critical Information Infrastructure Supply Chain Programme • NCSC Supply Chain Guidance • Supply-chain Levels for Software Artifacts (SLSA) • MITRE Supply Chain Security Framework • OWASP Top 10 LLM Applications • MITRE Supply Chain Security Framework • NIST Secure Software Development Framework for Generative AI and for Dual Use Foundation Models Virtual Workshop
2.1.2	Supply Chain Security: Ensure data, models, compilers, software libraries, developer tools and applications from trusted sources. **Responsible parties:** Decision Makers, AI Practitioners, Cybersecurity Practitioners					If procuring any AI System or component from a vendor, check/ensure suppliers adhere to policy and the equivalent security standards as your organisation. This could be done by establishing a Service Level Agreement (SLA) with the vendor. If the above is not plausible, consider using software components only from trusted sources. Verify object integrity e.g. hashes before using, opening, or running any files. Associated MITRE Mitigations: • AML.M0016 Vulnerability Scanning • AML.M0013 Code Signing • AML.M0007 Sanitize Training Data	

2.1	**Secure the Supply Chain** Assess and monitor the security of the supply chain across the system's life cycle.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
						• AML.M0014 Verify ML Artifacts • AML.M0008 Validate ML Model	
2.1.3*	Protect the integrity of data that will be used for training the model. **Responsible parties:** AI Practitioners				• Data poisoning attacks • Exposure of sensitive and classified data in the AI training Data Associated MITRE ATLAS Techniques: • AML.T0020.000 Poison Training Data • AML.T0019 Publish Poison Dataset	Use automated data discovery tools to identify sensitive data across various environments, including databases, data lakes, and cloud storage. Implement secure workflow and data flow to ensure the integrity of the data used. When viable, have humans look at each data input and generate notifications where labels differ. Use statistical and automated methods to check for abnormalities. Associated MITRE Mitigations: • AML.M0007 Sanitize Training Data • AML.M0014 Verify ML Artifacts	• ETSI AI Data Supply Chain Security • DSTL Machine Learning with Limited Data
2.1.4*	Consider the trade-offs when deciding to use an untrusted 3rd party model (with or without fine tuning). **Responsible parties:** Decision Makers, AI Practitioners, Cybersecurity Practitioners				• Model backdoors • Remote code execution Associated MITRE ATLAS Techniques: • AML.T0018 Backdoor ML Model • AML.T0043 Craft Adversarial Data • AML.T0050 Command and Scripting Interpreter	Untrusted 3rd party models are models obtained from public/private repositories, whose publisher's origins cannot be verified. While there are benefits to relying on 3rd party models, possible risks	

2.1	**Secure the Supply Chain** Assess and monitor the security of the supply chain across the system's life cycle.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
						include less control and visibility of model development. This reduced visibility may introduce backdoors injected by malicious actors. Consider the trade-offs based on your application's requirements Associated MITRE Mitigations: • AML.M0018 User Training • AML.M0013 Code Signing	
2.1.5*	Consider sandboxing untrusted models or serialised weight files where relevant. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners					Running the model within a virtual machine or isolated environment away from production environment. Associated MITRE Mitigations: • AML.M0008 Validate ML Model • AML.M0018 User Training • AML.M0013 Code Signing	
2.1.6*	Scan models or serialised weight files. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners					Use scanning tools such as Picklescan, Modelscan, on model files from an external source on a separate platform/system where the production system is on. Associated MITRE Mitigations: • AML.M0016 Vulnerability Scanning • AML.M0008 Validate ML Model	• Pickle Scanning (Hugging Face) • Stable Diffusion Pickle Scanner GUI • Also see Annex A – Technical Testing and System Validation

2.1	**Secure the Supply Chain** Assess and monitor the security of the supply chain across the system's life cycle.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
2.1.7	Consider the trade-offs associated with using sensitive data for model training or inference. **Responsible parties:** Decision Makers, AI Practitioners				• Data leaks • Compromised privacy Associated MITRE ATLAS Techniques: • AML.T0024 Exfiltration via ML Inference API • AML.T0057 LLM Data Leakage • AML.T0056 LLM Meta Prompt Extraction • AML.T0040 ML Model Inference API Access • AML.T0047 ML Model Product or Service • AML.T0049 Exploit Public Facing Application	Check that data uploaded is non-sensitive or protected before submitting to the external model according to enterprise data protection policy/requirements. Organisations may explore various risk mitigation measures to secure their non-public sensitive data, such as anonymisation and privacy-enhancing technologies, before making decision on the use of sensitive data for model training. Pay specific attention to supplier policies on the confidentiality of user data, most notably ensure that suppliers commit that user inputs and model outputs are not subsequently used for model training. If necessary, consider techniques such as anonymisation, before deciding to use sensitive data for training. Associated MITRE Mitigations: • AML.M0012 Encrypt Sensitive Information • AML.M0016 Vulnerability Scanning	• Advisory Guidelines on use of Personal Data in AI Recommendation and Decision Systems

2.1	**Secure the Supply Chain** Assess and monitor the security of the supply chain across the system's life cycle.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
2.1.8	Apply appropriate controls for data being sent out of the organisation. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners				• Data leaks • Compromised privacy Associated MITRE ATLAS Techniques: • AML.T0024 Exfiltration via ML Inference API • AML.T0057 LLM Data Leakage • AML.T0056 LLM Meta Prompt Extraction • AML.T0040 ML Model Inference API Access • AML.T0047 ML Model Product or Service • AML.T0049 Exploit Public Facing Application • Insecure or vulnerable libraries, which can introduce unexpected attack surfaces • Model Subversion Associated MITRE ATLAS Techniques: • AML.T0016 Obtain Capabilities	Implement an automated Data Loss Prevention, exfiltration countermeasures, alert triggers and possibly human intervention e.g. added confirmation via login and input confirmation. Associated MITRE Mitigations: • AML.M0012 Encrypt Sensitive Information • AML.M0004 Restrict Number of ML Model Queries • AML.M0019 Control Access to ML Models and Data in Production.	
2.1.9	Consider evaluation of dependent software libraries, open-source models and when possible, run code checking. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners					For example, ensure the library does not have arbitrary code execution when being imported or used. This can be done by using AI code checking, a vulnerability scanning tool, or checking against a database with vulnerability information. Associated MITRE Mitigations: • AML.M0008 Validate ML Model • AML.M0011 Restrict Library Loading • AML.M0004 Restrict Number of ML Model Queries • AML.M0008 Validate ML Model • AML.M0014 Verify ML Artifacts	• CVE List • Open-source Insights • OSS Insight

2.1	**Secure the Supply Chain** Assess and monitor the security of the supply chain across the system's life cycle.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
						• AML.M0011 Restrict Library Loading	
2.1.10	Use software and libraries that does not have known vulnerabilities. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners				• Insecure or vulnerable libraries, which can introduce unexpected attack surfaces • Model Subversion Associated MITRE ATLAS Techniques: • AML.T0016 Obtain Capabilities	Update to the latest secure patch in a timely manner. Associated MITRE Mitigations: • AML.M0008 Validate ML Model • AML.M0014 Verify ML Artifacts	• CVE List • Open-source Insights • OSS Insight

2.2	**Consider security benefits and trade-offs when selecting the appropriate model to use**						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
2.2.1*	Assess the need to use sensitive data for training the model, or directly referenced by the model. **Responsible parties:** AI Practitioners				• Privacy compromise • Attackers may be able to extract data used for training or from vector stores via malicious queries and prompt injections Associated MITRE ATLAS Techniques: • AML.T0057 LLM Data Leakage	Classify your organisation data based on sensitivity and/or enterprise data policy. Consider the need to use PII or sensitive data to generate vector databases that will be referenced by the model e.g. when using Retrieval Augmented Generation (RAG). Consider the trade-offs associated with using sensitive data for model training. Organisations may wish to explore various risk mitigation measures to secure their non-public sensitive data, such as anonymisation and privacy-enhancing technologies, before they decide whether to use such sensitive data for model training. Associated MITRE Mitigations: • AML.M0018 User Training	• Advisory Guidelines on use of Personal Data in AI Recommendation and Decision Systems (PDPC) • Generative AI Scoping Matrix • OWASP Machine Learning Security Top 10 (2023 edition) - Draft release v0.3 • OWASP Top 10 for Large Language Model Applications
2.2.2*	Consider Model hardening if appropriate. **Responsible parties:** AI Practitioners				• Input-based attacks • Prompt Injection • Adversarial Attacks • Model overfitting • Privacy compromise Associated MITRE ATLAS Techniques:	Apply data augmentation and adversarial training to reduce the effect of adversarial robustness attacks. Adversarial training: Inject adversarial text or image transformations (e.g. random flips, crops, rotation). This might	

2.2	**Consider security benefits and trade-offs when selecting the appropriate model to use**						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
					• AML.T0043 Craft Adversarial Data • AML.T0015 Evade ML Model • AML.T0024 Exfiltration via ML Inference API • AML.T0051 LLM Prompt Injection • AML.T0057 LLM Data Leakage • AML.T0054 LLM Jailbreak	impact the effectiveness of the model. For LLMs, prompt engineering best practices such as usage of guardrails and wrapping instructions in a single pair of salted sequence tags can be methods to further ground the model. Overfitting can increase the chance of adversarial attacks through model inversion. Associated MITRE Mitigations: • AML.M0003 Model Hardening • AML.M0006 Use Ensemble Methods • AML.M0010 Input Restoration • AML.M0015 Adversarial Input Detection • AML.M0004 Restrict Number of ML Model Queries	
2.2.3*	Consider implementing techniques to strengthen/harden the system apart from strengthening the model itself. **Responsible parties:** AI Practitioners				• Adversarial attacks on the model Associated MITRE ATLAS Techniques: • AML.T0015 Evade ML Model • Infrastructure Attacks • AML.T0029 Denial of ML Service • Attacker Recon activities • AML.TA002 ATLAS Tactic Recon	Supporting Countermeasures: • Cyber threat Intelligence to analyse and predict attacks. • Involve beta users (better red teaming) to test, exploit the wisdom of the crowds. • Anti-recon measures via hiding, disinformation, deception (honeypots).	

2.2	Consider security benefits and trade-offs when selecting the appropriate model to use						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
						• High quality datasets to improve model performance. • Data security controls for data collection, data storage, data processing, and data use as well as code and model security. • For LLMs, implement guardrails or input validation. • Implement endpoint security. • Consider implementing Zero Trust Principles for the system. Associated MITRE Mitigations: • AML.M0003 Model Hardening • AML.M0006 Use Ensemble Methods • AML.M0010 Input Restoration • AML.M0015 Adversarial Input Detection • AML.M0004 Restrict Number of ML Model Queries • AML.M0019 Control Access to ML Models and Data in Production	

2.3	**Identify, track and protect AI-related assets** Understand the value of AI-related assets, including models, data, prompts, logs, and assessments. Have processes to track, authenticate, version control, and secure assets.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
2.3.1	Establishing a data lineage and software license management process. This includes documenting the data, codes, test cases and model, including any changes made and by whom. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners				• Loss of data integrity • Unauthorised changes to data, model or system • Insider threats • Ransomware attacks • Loss of intellectual property Associated MITRE ATLAS Techniques: • AML.T0018.000 Backdoor ML Model • AML.T0020.000 Poison Training Data • AML.T0011 User Execution	Model cards, Data cards, and Software Bill of Materials (SBOMs) may be used. Associated MITRE Mitigations: • AML.M0016 Vulnerability Scanning • AML.M0013 Code Signing • AML.M0007 Sanitize Training Data • AML.M0014 Verify ML Artifacts • AML.M0008 Validate ML Model • AML.M0005 Control Access to ML Models and Data at Rest • AML.M0018 User Training	• Cybersecurity Code of Practice for Critical Information Infrastructure (CSA) • ISO 27001: Information security, cybersecurity and privacy protection
2.3.2	Secure data at rest, and data in transit. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners				• Data loss and leaks. • Loss of data integrity. • Ransomware encryption. Associated MITRE ATLAS Techniques: • AML.T0024 Exfiltration via ML Inference API • AML.T0025 Exfiltration via Cyber Means • AML.T0054 LLM Jailbreak	Sensitive data (model weight and python code) is stored encrypted and transferred with proper encryption protocols, and secure key management. Consider saving model weights in secure formats such as safetensor, etc. Associated MITRE Mitigations: • AML.M0012 Encrypt Sensitive Information • AML.M0005 Control Access to ML Models and Data at Rest • AML.M0019 Control Access to ML Models and Data in Production	

2.3	**Identify, track and protect AI-related assets** Understand the value of AI-related assets, including models, data, prompts, logs, and assessments. Have processes to track, authenticate, version control, and secure assets.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
2.3.3	Have regular backups in event of compromise. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners					Identify essential data to backup more frequently. Implement a regular backup schedule. Have redundancy to ensure availability. Associated MITRE Mitigations: • AML.M0014 Verify ML Artifacts • AML.M0005 Control Access to ML Models and Data at Rest • AML.M0019 Control Access to ML Models and Data in Production	
2.3.4*	Implement controls to limit what AI can access and generate, based on sensitivity of the data. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners				• Data leaks • Privacy attacks Associated MITRE ATLAS Techniques: • AML.T0036 Data from Information Repositories • AML.T0037 Data from Local System • AML.T0057 LLM Data Leakage	For sensitive data such as PII, explore various risk mitigation measures to secure non-public sensitive data, such as data anonymisation and privacy-enhancing techniques, before input into the AI. Have filters at the output to prevent sensitive information from being leaked. Associated MITRE Mitigations: • AML.M0012 Encrypt Sensitive Information • AML.M0019 Control Access to ML Models and Data in Production • AML.M0014 Verify ML Artifacts	• Advisory Guidelines on use of Personal Data in AI Recommendation and Decision Systems

2.3	**Identify, track and protect AI-related assets** Understand the value of AI-related assets, including models, data, prompts, logs, and assessments. Have processes to track, authenticate, version control, and secure assets.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
						• AML.M0005 Control Access to ML Models and Data at Rest	
2.3.5	For very private data, consider if privacy enhancing technologies may be used. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners				• Data leaks Associated MITRE ATLAS Techniques: • AML.T0024 Exfiltration via ML Inference API	Examples include having a Trusted Execution Environment, differential privacy or homomorphic encryption. Associated MITRE Mitigations: • AML.M0012 Encrypt Sensitive Information	

2.4	**Secure the AI development environment** Apply good infrastructure security principles.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
2.4.1	Implement appropriate access controls to APIs, models and data, logs, and the environments that they are in. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners				• Unauthorised access to system, data and models • Data breaches • Model/system compromise • Loss of intellectual property Associated MITRE ATLAS Techniques: • AML.T0024 Exfiltration via ML Inference API • AML.T0025 Exfiltration via Cyber Means • AML.T0036 Data from Information Repositories • AML.T0037 Data from Local System • AML.T0012 Valid Accounts • AML.T0057 LLM Data Leakage • AML.T0053 LLM Plugin Compromise • AML.T0054 LLM Jailbreak • AML.T0044 Full ML Model Access • AML.T0055 Unsecured Credentials • AML.T0013 Discover ML Ontology • AML.T0014 Discover ML Family • AML.T0007 Discover ML Artifacts • AML.T0035 ML Artifact Collection	Have secure authentication processes. Rule and role-based access controls to the development environment, based on the principles of least privilege. Have periodic reviews for role conflicts or violations of segregation of duties, and documentation should be retained including remediation actions. Access should be promptly revoked for terminated users or when the employee no longer requires access. Associated MITRE Mitigations: • AML.M0005 Control Access to ML Models and Data at Rest • AML.M0019 Control Access to ML Models and Data in Production • AML.M0012 Encrypt Sensitive Information • AML.M0014 Verify ML Artifacts	• Cybersecurity Code of Practice for Critical Information Infrastructure (CSA) • ISO 27001: Information security, cybersecurity and privacy protection • Advisory Guidelines on use of Personal Data in AI Recommendation and Decision Systems

2.4	**Secure the AI development environment** Apply good infrastructure security principles.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
2.4.2	Implement access logging and monitoring. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners				• Anomalies and suspicious activities not detectable • Failed compliance and audit. • Poor transparency and accountability • Insider threats Associated MITRE ATLAS Techniques: • AML.T0024 Exfiltration via ML Inference API • AML.T0025 Exfiltration via Cyber Means • AML.T0040 ML Model Inference API Access • AML.T0020.000 Poison Training Data	Log access with timestamps. Track changes to the data and model or configuration changes. Protect logs from being attacked (deleted, or tampered) Associated MITRE Mitigations: • AML.M0005 Control Access to ML Models and Data at Rest • AML.M0019 Control Access to ML Models and Data in Production	
2.4.3	Segregate production/ development environments. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners				• Data integrity and confidentiality being compromised • Limit the impact of potential attacks • Risk of disruptions or conflicts between different functions/ models • Insider attacks Associated MITRE ATLAS Techniques: • AML.T0024 Exfiltration via ML Inference API • AML.T0025 Exfiltration via Cyber Means	Consider keeping different project environments separate from each other. E.g. development separated from production. If you are using cloud services, consider compartmentalizing your projects using VPCs, VMs, VPNs, enclaves, and containers Associated MITRE Mitigations: • AML.M0005 Control Access to ML Models and Data at Rest • AML.M0019 Control Access to ML Models and Data in Production	

2.4	**Secure the AI development environment** Apply good infrastructure security principles.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
2.4.4	Ensure configurations are secure by default. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners				• Unauthorized access and data breaches • Insider threats Associated MITRE ATLAS Techniques: • AML.T0024 Exfiltration via ML Inference API • AML.T0025 Exfiltration via Cyber Means	Default option should be secure against common threats. E.g. Implicitly deny access to sensitive data. Associated MITRE Mitigations: • AML.M0005 Control Access to ML Models and Data at Rest • AML.M0019 Control Access to ML Models and Data in Production	

2.2.3. DEPLOYMENT

3.1	**Secure the deployment infrastructure and environment of AI systems** Apply good infrastructure security principles.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
3.1.1	Ensure contingency plans are in place to mitigate disruption or failure of AI services. **Responsible parties:** Decision Makers, AI Practitioners, Cybersecurity Practitioners				• Extended downtime to availability Associated MITRE ATLAS Techniques: • AML.T0029 Denial of ML Service	Having a manual or secondary system as a fail-over/fail-safe if the AI service becomes unavailable.	• Cybersecurity Code of Practice for Critical Information Infrastructure (CSA) • ISO 27001: Information security, cybersecurity and privacy protection • Advisory Guidelines on use of Personal Data in AI Recommendation and Decision Systems • NSA Guidance for Strengthening AI System Security
3.1.2	Implement appropriate access controls to APIs, models and data, logs, configuration files and the environments that they are in. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners				• Unauthorized access to sensitive AI models and data • Data breaches • Loss of model integrity • Loss of intellectual property Associated MITRE ATLAS Techniques: • AML.T0024 Exfiltration via ML Inference API • AML.T0025 Exfiltration via Cyber Means • AML.T0040 ML Model Inference API Access • AML.T0020.000 Poison Training Data	Have secure authentication processes. Rule and role-based access controls to the deployment environment, based on the principles of least privilege. Have periodic reviews for role conflicts or violations of segregation of duties, and documentation should be retained including remediation actions. Access should be removed timely for terminated users or when the employee no longer requires access. Associated MITRE Mitigations: • AML.M0005 Control Access to ML Models and Data at Rest • AML.M0019 Control Access to ML Models and Data in Production	

3.1	**Secure the deployment infrastructure and environment of AI systems** Apply good infrastructure security principles.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
3.1.3	Implement access logging, monitoring and policy management **Responsible parties:** AI Practitioners, Cybersecurity Practitioners				• Unauthorized access to deployment infrastructure and environment • Undetected Anomalies and suspicious activities • Nonadherence to compliance and audit requirements • Data integrity and accountability • Insider threats Associated MITRE ATLAS Techniques: • AML.T0024 Exfiltration via ML Inference API • AML.T0025 Exfiltration via Cyber Means • AML.T0040 ML Model Inference API Access	Keep a record of access to the model, inputs to the model, and output behaviour of the model. If necessary, track all AI applications, models and data. Have the ability to discover all AI apps, models, and data across the system, and who they are used by. Define and enforce data security policies across their environments. Associated MITRE Mitigations: • AML.M0005 Control Access to ML Models and Data at Rest • AML.M0019 Control Access to ML Models and Data in Production	
3.1.4	Implement segregation of environments. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners				• Data integrity and confidentiality being compromised • Limit the impact of potential attacks, Risk of disruptions or conflicts between different functions/models Associated MITRE ATLAS Techniques: • AML.T0029 Denial of ML Service • AML.T0025 Exfiltration via Cyber Means	Keep different project environments separate from each other. E.g. when working on the cloud, have a separate VPC. Keep the development and operational environment apart. Associated MITRE Mitigations: • AML.M0019 Control Access to ML Models and Data in Production	

3.1	**Secure the deployment infrastructure and environment of AI systems** Apply good infrastructure security principles.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
					• AML.T0031 Erode ML Model Integrity		
3.1.5	Ensure configurations are secure by default. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners				• Vulnerability exploitation, Unauthorized access, Data breaches • Insider threats Associated MITRE ATLAS Techniques: • AML.T0024 Exfiltration via ML Inference API • AML.T0025 Exfiltration via Cyber Means • AML.T0031 Erode ML Model Integrity	Default option should be secure against common threats. E.g. Implicitly deny access to sensitive data. Associated MITRE Mitigations: • AML.M0019 Control Access to ML Models and Data in Production	
3.1.6	Consider implementing firewalls. **Responsible parties:** Cybersecurity Practitioners				• Unauthorized access to AI systems, models, and data • Network-based attacks, such as denial-of-service (DoS) attacks. • Malware and intrusion attempts • Unauthorized access to specific components of the AI systems Associated MITRE ATLAS Techniques: • AML.T0029 Denial of ML Service • AML.T0046 Spamming ML System with Chaff Data	Consider implementing Firewalls if the model is accessible to users online. Associated MITRE Mitigations: • AML.M0005 Control Access to ML Models and Data at Rest • AML.M0019 Control Access to ML Models and Data in Production	

3.1	**Secure the deployment infrastructure and environment of AI systems** Apply good infrastructure security principles.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
3.1.7	Implement any other relevant security controls based on cybersecurity best practice, which has not been stated above. **Responsible parties:** Decision Makers, AI Practitioners, Cybersecurity Practitioners					Implement any other relevant security control based on best practice, such as ISO 27001.	

3.2	**Establish incident management procedures** Ensure proper incident response, escalation, and remediation plans.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
3.2.1	Have plans to address different attack and outage scenarios. Implement measures to assist investigation. **Responsible parties:** Decision Makers, AI Practitioners, Cybersecurity Practitioners				• Failed Incident Response • Disruption to business continuity Associated MITRE ATLAS Techniques: • AML.T0029 Denial of ML Service	Have different incident response plans that address different types of outages and the potential attack scenarios, which may be blended with DOS. Implement forensics support and protect against erasure of evidence. Use cyber threat intelligence to support investigation. Associated MITRE Mitigations: • AML.M0018 User Training	• CSA Incident Response Checklist
3.2.2	Regularly reassess incident response plans as the system changes. **Responsible parties:** Decision Makers, AI Practitioners, Cybersecurity Practitioners				• Failed Incident Response • Disruption to business continuity Associated MITRE ATLAS Techniques: • AML.T0029 Denial of ML Service	Assess how changes to the system and AI will affect the attack surfaces. Associated MITRE Mitigations: • AML.M0018 User Training	
3.2.3	Have regular backups in event of compromise. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners				• Data Loss • Ransomware attacks • Operational Disruptions • Data Integrity Associated MITRE ATLAS Techniques: • AML.T0029 Denial of ML Service • AML.T0031 Erode ML Model Integrity	Store critical data assets in offline backups. Associated MITRE Mitigations: • AML.M0014 Verify ML Artifacts	

3.2	**Establish incident management procedures** Ensure proper incident response, escalation, and remediation plans.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
3.2.4	When an alert has been raised or investigation has confirmed an incident, report to the relevant stakeholders **Responsible parties:** Decision Makers, AI Practitioners, Cybersecurity Practitioners				• Regulatory non-Compliance • Increased cost and damages to the enterprise	Use threat hunting to determine full extent of attack and investigate attribution.	

3.3	**Release AI systems responsibly** Release models, applications, or systems only after subjecting them to appropriate and effective security checks and evaluation						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
3.3.1*	Verify models with hashes/signatures of model files and datasets before deployment or periodically, according to enterprise policy. **Responsible parties:** AI Practitioners				• Model Tampering/Poisoning • Data Poisoning • Backdoor/ Trojan model Associated MITRE ATLAS Techniques: • AML.T0018.000 Backdoor ML Model • AML.T0020.000 Poison Training Data	Compute and share model and dataset hashes/signatures when creating new models or data and update the relevant documentation e.g. model cards. Associated MITRE Mitigations: • AML.M0014 Verify ML Artifacts	
3.3.2*	Benchmark and test the AI models before release. **Responsible parties:** AI Practitioners				• Failure to achieve trust and reliability • Adversarial Attacks • Lack of accountability • Model Robustness Associated MITRE ATLAS Techniques: • AML.T0048 External Harm • AML.T0043 Craft Adversarial Data • AML.T0031 Erode ML Model Integrity	Models have been validated and achieved performance targets before deployment. Consider using an adversarial test set to validate model robustness, where possible. Conduct AI Red-Teaming. Associated MITRE Mitigations: • AML.M0008 Validate ML Model • AML.M0014 Verify ML Artifacts	• Adversarial Robustness Toolbox (IBM) • CleverHans (University of Toronto) • TextAttack (University of Virginia) • Prompt Bench (Microsoft) • Counterfit (Microsoft) • AI Verify (Infocomm Media Development Authority, Singapore) • Moonshot (Infocomm Media Development Authority, Singapore)

3.3	**Release AI systems responsibly** Release models, applications, or systems only after subjecting them to appropriate and effective security checks and evaluation						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
3.3.3	Consider the need to conduct security testing on the AI systems. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners				• Security Vulnerabilities Associated MITRE ATLAS Techniques: • AML.T0048 External Harm • AML.T0031 Erode ML Model Integrity	Perform VAPT/security testing on AI systems. Prioritise and focus on the most realistic and practical attacks, based on the risk assessment during the planning phase. System owner and project teams to follow up on findings from security testing/red team, by assessing the criticality of vulnerabilities uncovered, apply additional measures and if necessary, seek approval from relevant entity e.g. CISO, for acceptance of residual risks, according to their enterprise risk management/cybersecurity policies. Create a feedback loop to maximise the impact of the findings from security tests. Associated MITRE Mitigations: • AML.M0003 Model Hardening • AML.M0006 Use Ensemble Methods • AML.M0016 Vulnerability Scanning	• OWASP Top 10 for Large Language Model Applications • Web LLM attacks (Portswigger)

2.2.4. OPERATIONS AND MAINTENANCE

4.1	**Monitor AI system inputs** Monitor and log inputs to the system, such as queries, prompts and requests. Proper logging allows for compliance, audit, investigation and remediation.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
4.1.1*	Validate/Monitor inputs to the model and system for possible attacks and suspicious activity. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners				• Adversarial Attacks • Data exfiltration Associated MITRE ATLAS Techniques: • AML.T0043 Craft Adversarial Data • AML.T0025 Exfiltration via Cyber Means	AI System owners may consider to monitor and validate input prompts, queries or API requests for attempts to access, modify or exfiltrate information deemed confidential by the organisation. Consider use of classifiers to detect malicious inputs and log them for future review to identify potential vulnerabilities. Note: Implementor should consider the current privacy regulations/guidelines when logging inputs. Associated MITRE Mitigations: AML.M0015 Adversarial Input Detection	• Introduction to Logging for Security Purpose (NCSC) • OpenAI usage policies • Advisory Guidelines On use of Personal Data In AI Recommendation and Decision Systems (PDPC)
4.1.2	Monitor/Limit the rate of queries. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners				• Denial of Service (DoS) Attacks Associated MITRE ATLAS Techniques: • AML.T0029 Denial of ML Service • AML.T0034 Cost Harvesting	If possible, prevent users from continuously querying the model with a high frequency e.g. API throttling. This mitigates the potential for membership-inference and extraction attacks. Associated MITRE Mitigations: • AML.M0004 Restrict Number of ML Model Queries	

4.2	**Monitor AI system outputs and behaviour** Monitor for anomalous behaviour that might indicate intrusions or compromise.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
4.2.1*	Monitor model outputs and model performance. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners				• Adversarial Attacks • Operational Impact Associated MITRE ATLAS Techniques: • AML.T0031 Erode ML Model Integrity • AML.T0020.000 Poison Training Data • AML.T0029 Denial of ML Service • AML.T0048 External Harms	Implement an alert system that monitors for anomalous or unwanted output. E.g. a customer facing chatbot that is safe for work begins to output profanity instead. Associated MITRE Mitigations: • AML.M0008 Validate ML Model	
4.2.2*	Ensure adequate human oversight to verify model output, when viable or appropriate. **Responsible parties:** AI Practitioners, Cybersecurity Practitioners				• False Positives from the model • Misinterpretation of Context • Adverse Impact on Operations Associated MITRE ATLAS Techniques: • AML.T0029 Denial of ML Service • AML.T0048 External	Manual investigation of unusual or anomalous alert notifications. For critical systems, ensure human oversight to verify decisions recommended by the model. Associated MITRE Mitigations: • AML.M0018 User Training • AML.M0015 Adversarial Input Detection	

4.3	**Adopt a secure-by-design approach to updates and continuous learning.** Ensure risks associated to model updates have been considered. Changes to the data and model can lead to changes in behaviour.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
4.3.1*	Treat major updates as new versions and integrate software updates with model updates and renewal. **Responsible parties:** AI Practitioners				• Model Tampering/Poisoning • Backdoor/ Trojan model Associated MITRE ATLAS Techniques: • AML.T0020.000 Poison Training Data • AML.T0018.000 Backdoor ML Model • AML.T0031 Erode ML Model Integrity • AML.T0010 ML Supply Chain Compromise	New models to be validated, benchmarked, and be tested before release. Associated MITRE Mitigations: • AML.M0008 Validate ML Model • AML.M0014 Verify ML Artifacts	• Principles for the Security of Machine Learning (UK NCSC)
4.3.2*	Treat new input data used for training as new data. **Responsible parties:** AI Practitioners				• Data Poisoning • Poison/Backdoor/Trojan model Associated MITRE ATLAS Techniques: • AML.T0020.000 Poison Training Data • AML.T0018.000 Backdoor ML Model • AML.T0010 ML Supply Chain Compromise	Subject new input to the same verification and validation as new data. Associated MITRE Mitigations: • AML.M0007 Sanitize Training Data	

4.4	**Establish a vulnerability disclosure process** Have a feedback process for users to share any findings of concern, which might uncover potential vulnerabilities to the system.						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Possible Risk Mitigated**	**Example Implementation**	**Reference or Resource**
4.4.1	Maintain open lines of communication. **Responsible parties:** Decision Makers, AI Practitioners, Cybersecurity Practitioners				• Regulatory non-Compliance	Set up channels to allow users to provide feedback on security and usage.	• SingCERT Vulnerability Disclosure Policy (CSA) • UK NCSC Vulnerability Disclosure Toolkit • CVE List • AI CWE List • ATLAS Case Studies
4.4.2	Share findings with appropriate stakeholders. **Responsible parties:** Decision Makers, AI Practitioners, Cybersecurity Practitioners				• Regulatory non-Compliance	Share discoveries of vulnerabilities to relevant stakeholders such as the company CISO.	

2.2.5. END OF LIFE

5.1	**Ensure proper data and model disposal**						
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Related Risks**	**Example Implementation**	**Reference or Resource**
5.1.1	Ensure proper and secure disposal/destruction of data and models in accordance with data privacy standards and/or relevant rules and regulations. **Responsible parties:** Decision Makers, AI Practitioners, Cybersecurity Practitioners				• Regulatory non-Compliance • Sensitive data loss	Examples include crypto shredding or degaussing	• Personal Data Protection Act (PDPA) • Advisory Guidelines on use of Personal Data in AI Recommendation and Decision Systems

3. USE CASE EXAMPLES

3.1. DETAILED WALKTHROUGH EXAMPLE

Case Study: Implementing Companion Guide on LLM-based Chatbot

- Company A is currently testing out an LLM to implement as their customer service chatbot, known as *SuperResponder*.
- The model is an LLM that is downloaded from an open-source model hosting website (Hugging Face) and further developed in-house on a cloud environment.
- The data is sourced from manually curated FAQs from customer service conversations, which will be converted to a vector database to implement Retrieval Augmented Generation (RAG) with the downloaded LLM model.

Supply Chain Attacks

In this example, Company A relies heavily on third party AI software components to develop *SuperResponder*.

The integrity and security of AI supply chains are essential for ensuring the reliability and trustworthiness of AI systems. AI vulnerabilities in the supply chain refer to weaknesses or exploitable points within the processes of acquiring, integrating, and deploying AI technologies. These vulnerabilities can stem from malicious or compromised components, including datasets, models, algorithms, and software libraries, which may introduce security risks and threats to AI systems[2].

[2] https://vulcan.io/blog/understanding-the-hugging-face-backdoor-threat/

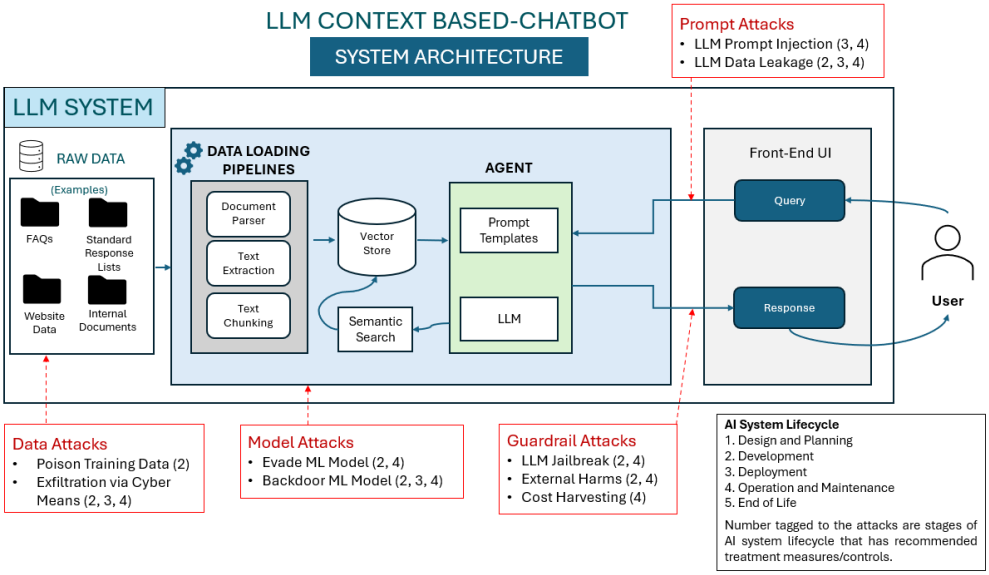

Figure 2. LLM Context Based-chatbot System Architecture

3.1.1. RISK ASSESSMENT EXAMPLE

Company A performed a risk assessment to identify and address potential risks to confidentiality, integrity, availability of their AI system. If the risks are not mitigated, there is a potential for an attacker to exploit the list of vulnerabilities, causing *SuperResponder* to be compromised. This could result in widespread customer dissatisfaction and damage to the company's reputation.

The hypothetical risk assessment* is as follows:

Risk Scenarios	Impact	Likelihood	Proposed Mitigations	Risk Level
Prompt injection attack Crafted input can be executed to instruct LLM to retrieve private customer information.	**Confidentiality: High** Confidential information such as PII data of customers may be leaked.	**Likelihood: Medium** Chatbot interface is public facing. Attack can be performed easily without privileged access and be repeated continuously.	Use automated tools to remove PIIs from datasets used. In addition, use data protection measures and output sanitisation mechanisms.	Initial Risk Level: **Medium** Residual Risk Level: **Low**
Supply Chain Vulnerabilities. Use of compromised pre-trained LLM can introduce other vulnerabilities such as model backdoor.	**Integrity: High** The chatbot may be prompted to regularly output the wrong answer or advice to customers.	**Likelihood: Medium** It is possible to upload compromised models onto public model hosting platforms. These models are downloaded and used to develop the chatbot.	Scanning the model. Sandboxing the model. Download models from trusted model developers or sources.	Initial Risk Level: **Medium** Residual Risk Level: **Low**
Model Denial of Service. Chatbot at risk of volumetric and continuous querying, consuming a large amount of resource.	**Availability: Medium** The chatbot service can be overwhelmed by a large volume of requests and become unavailable to other users.	**Likelihood: Medium** Volumetric and continuous querying of the chatbot can be performed with some scripting knowledge or automated tools.	API throttling.	Initial Risk Level: **Medium** Residual Risk Level: **Low**

* The above table is not exhaustive and is meant as an example of a risk assessment done.

3.1.2. WALKTHROUGH OF TABULATED MEASURES/CONTROLS

Following the risk assessment, Company A promptly referenced the CSA Guidelines for Securing AI Systems and the Companion Guide to mitigate the risks. The list of implemented actions are as follows:

3.1.2.1. PLANNING AND DESIGN STAGE

1.1	**Raise awareness and competency on security risks**				
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Implementation done / reason not done**
1.1.1	Ensure system owners and senior leaders understand threats to secure AI and their mitigations.	✓			System owners have attended seminars on AI security and understood potential risks associated with AI systems.
1.1.2	Provide guidance to staff on Security by Design and Security by Default principles as well as unique AI security risks and failure modes as part of InfoSec training. e.g. LLM security matters, common AI weaknesses and attacks.	✓			Trained staff on AI security and risks, e.g. attack vectors, and countermeasures (practical defence strategies). Developers were sent to attend a 3-day course on AI & Cybersecurity covering adversarial machine learning at a local tertiary institution. They also referred to online courses from Udemy on AI security essentials and AI risk management.
1.1.3	Train developers are trained in secure coding practices and good practices for the AI lifecycle.	✓			Developers have attended certified workshops on how to maintain secure coding practices when developing the model.

1.2	Conduct security risk assessments				
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Implementation done / reason not done**
1.2.1	Understand AI governance and legal requirements, the impact to the system, users, organisation, if an AI component is compromised or has unexpected behaviour or there is an attack that affected AI privacy. Plan for an attack and its mitigation, using the principles of CIA.	✓			Understood AI Verify framework, PDPA Guidance for AI and User Data, Model Governance Framework for Generative AI (IMDA)
1.2.2	Assess AI-related attacks and implement mitigating steps.	✓			Threat Modelling: Identify and assess potential attack vectors from adversarial attacks such as prompt injection, membership inference, data poisoning and backdoor attacks using mitre atlas framework
1.2.3	Conduct risk assessment is done in accordance with the relevant industry standards/best practices.	✓			Risk assessment was conducted in accordance with company risk management policy

3.1.2.2. DEVELOPMENT

2.1	Secure the Supply Chain				
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Example Implementation of Action**
2.1.1	Implement Secure Coding and Development Lifecycle.	✓			Attended secure coding courses and adopt secure coding practices for development of the LLM.
2.1.2	Supply Chain Security: Ensure data, models, compilers, software libraries, developer tools and applications from trusted sources.	✓			For the pre-trained 3rd party LLM model – Applied Source verification to ensure data and models obtained are from trusted and reputable sources. Verified the authenticity and integrity of the sources before incorporating them into the system (digital signatures).
2.1.3	Protect the integrity of data that will be used for training the model.	✓			Data to support Retrieval Augmented Generation (RAG) is sourced from company's own customer service conversations and internal FAQ documents.
2.1.4	Consider the trade-offs when deciding to use an untrusted 3rd party model (with or without fine tuning).	✓			Examples of risks considered: Data Breaches, Data Privacy Leakage, Service Disruptions, Model backdoor. Compensatory measures such as prompt filters and prompt engineering to mitigate adversarial attacks.
2.1.5	Consider sandboxing untrusted models or serialised weight files where relevant.	✓			Implemented virtual machines (VMs), to isolate and restrict execution environment of these components.
2.1.6	Scan models or serialised weight files.	✓			Scanned model files with Picklescan
2.1.7	Consider the trade-offs associated with using sensitive data for model training or inference			✓	Not using external APIs during development
2.1.8	Apply appropriate controls for data being sent out of the organisation.			✓	Not required as model is hosted locally and not a SaaS.

2.1.9	Consider evaluation of dependent software libraries, open-source models and when possible, run code checking.	✓			Used a vulnerability scanner to ensure safety of third-party libraries from known CVEs
2.1.10	Use software and libraries that does not have known vulnerabilities.	✓			Use of updated software and libraries with no known vulnerability in accordance with company IT policy

2.2	**Consider security benefits and trade-offs when selecting the appropriate model to use**				
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Implementation done / reason not done**
2.2.1	Assess the need to use sensitive data for training the model, or directly referenced by the model.	✓			Sensitive data not used for vector database. Training data has been carefully sanitised by sensitive data redaction methods to counter inference attacks.
2.2.2	Consider Model hardening if appropriate.	✓			Prompt engineering to prevent the model from producing output beyond what is intended. Implemented guardrails to ensure sensitive data is not disclosed.
2.2.3	Consider implementing techniques to strengthen/harden the system apart from strengthening the model itself.	✓			Added input prompt filters and output filters for unwanted topics, to mitigate against prompt injections.

2.3	Identify, track and protect AI-related assets				
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Implementation done / reason not done**
2.3.1	Establishing a data lineage and software license management process. This includes documenting the data, codes, test cases and model, including any changes made and by whom.	✓			Maintained documentation of the changes made to newer model versions on Model cards and verified it.
2.3.2	Secure data at rest, and data in transit.	✓			Encryption algorithms approved by enterprise security policy is used for data at rest and transit.
2.3.3	Have regular backups in event of compromise.	✓			Used git to maintain version control of the codebase and model artifacts.
2.3.4	Implement controls to limit what AI can access and generate, based on sensitivity of the data.	✓			Prompt engineering to ensure that the model is less likely to generate any unwanted topics.
2.3.5	For very private data, privacy enhancing technologies may be used.			✓	No private data used

2.4	Secure the AI development environment				
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Implementation done / reason not done**
2.4.1	Appropriate access controls to APIs, models and data, logs, and the environments that they are in.	✓			Rule and role-based access controls implemented for developers
2.4.2	Implement access logging and monitoring.	✓			Turned on cloud native logging.
2.4.3	Segregation production/ development environments.	✓			Developer environment is in a different VPC from the deployment environment.
2.4.4	Ensure configurations are secure by default.	✓			Implicit deny access to unauthorised users via cloud native identity and access management.

3.1.2.3. DEPLOYMENT

3.1	Secure the deployment infrastructure and environment of AI systems				
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Implementation done**
3.1.1	Ensure contingency plans are in place to mitigate disruption or failure of AI services.	✓			Deployed a backup availability zone to ensure availability of service.
3.1.2	Implement appropriate access controls to APIs, models and data, logs, configuration files and the environments that they are in.	✓			General users only have access to the LLM interface via the frontend chatbot, no access to the backend environment.
3.1.3	Implement access logging, monitoring and policy management	✓			Turned on cloud native logging.
3.1.4	Implementation segregation of environments.	✓			Deployment environment is in a different VPC from the development environment.
3.1.5	Ensure configurations are secure by default.	✓			Implicit deny access to unauthorised users via cloud native identity and access management.
3.1.6	Consider implementing firewalls.	✓			Configured firewalls in between access to environment and model
3.1.7	Implement any other relevant security controls based on cybersecurity best practice, which has not been stated above.			✓	Current controls are in line with company cybersecurity policy.

3.2	Establish incident management procedures				
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Implementation done / reason not done**
3.2.1	Have plans to depict different attack and outage scenarios. Implement measures to assist investigation.	✓			Conducted exercise to simulate outage of AI chatbot and fail over to another availability zone.
3.2.2	Regularly reassess incident response plans as the system changes.	✓			Will reassess system every 12 months or whenever there is an update to the system, according to company cybersecurity policy.
3.2.3	Have regular backups in event of compromise.	✓			Weekly backups in place, according to company IT policy.
3.2.4	When an alert has been raised or investigation has confirmed incident, to report to the relevant stakeholders.	✓			Procedure in place to report to CISO, in accordance with incident response standard operating procedure.

3.3	Release AI systems responsibly				
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Implementation done / reason not done**
3.3.1	Verify models with hashes/signatures of model files and datasets before deployment or periodically, according to enterprise policy.	✓			Models are validated with hashes before deployment.
3.3.2	Benchmark and test the AI models before release.	✓			Prepared a golden dataset to validate and benchmark model. Conducted red teaming on the LLM model before release; incorporating test cases on prompt injection and supply chain attacks, which were identified during the security risk assessment.
3.3.3	Consider need to conduct security testing on the AI systems.	✓			Performed VAPT/security testing on LLM systems. The system owner followed up on findings from the red team, assessed the criticality of vulnerabilities uncovered, applied additional measures, and sought approval from CISO for acceptance of vulnerabilities that cannot be rectified.

3.1.2.4. OPERATIONS AND MAINTENANCE

4.1	Monitor AI system inputs				
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Implementation done / reason not done**
4.1.1	Validate/monitor inputs to the model and system for possible attacks and suspicious activity.	✓			All inputs to the LLM that has guardrails triggered are logged for future review and to identify potential vulnerabilities in prompt design.
4.1.2	Monitor/Limit the rate of queries.	✓			API throttling is in place to limit rate on queries to model.

4.2	Monitor AI system outputs and behaviour				
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Implementation done / reason not done**
4.2.1	Monitoring of model outputs and model performance.	✓			Implement a monitoring system to detect anomalous behaviour or outputs from the LLM system that could indicate an attack or vulnerability.
4.2.2	Ensure adequate human oversight to verify model output, when viable or appropriate.	✓			Manually investigate unusual, automated processes that are flagged as anomalous.

4.3	Adopt a secure-by-design approach to updates and continuous learning.				
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Implementation done / reason not done**
4.3.1	Treat major updates as new versions and integrate software updates with model updates and renewal.	✓			To validate and benchmark new models and updates against a 'Golden dataset'
4.3.2	Treat new input data used for training as new data.	✓			New data used for finetuning will be validated as they were new data.

4.4	Establish a vulnerability disclosure process				
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Implementation done / reason not done**
4.4.1	Maintain open lines of communication.	✓			Establish a vulnerability disclosure program (bounty program, etc.) to encourage responsible reporting and handling of security vulnerabilities by the users.
4.4.2	Share findings with appropriate stakeholders.	✓			New findings will be shared with company CISO

3.1.2.5. END OF LIFE

5.1	Ensure proper data and model disposal				
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Implementation done / reason not done**
5.1.1	Ensure proper and secure disposal/destruction of data and models in accordance with data privacy standards and/or relevant rules and regulations.	✓			All data related to the chatbot, and vector database will be deleted through the CSP data disposal process, in line with company data policy.

3.2. STREAMLINED IMPLEMENTATION EXAMPLE

Case Study: Patch attacks on image recognition surveillance system

- Company B has recently implemented an advanced AI-driven facial recognition gantry system at all access points at their office.
- The system is part of enhanced security measures to identify individuals and to streamline employee flow by reducing dependence on manual checks.
- Facial recognition systems utilise deep learning algorithms to identify individuals, by analysing visual data captured through cameras.

Patch Attacks	A patch attack is a type of attack that disrupts object classification in a camera's visual field by introducing a specific pattern or object. This disruption can lead to misinterpretation or evasion attacks.
In this example, the system owner has identified patch attack as a possible attack vector for this system	

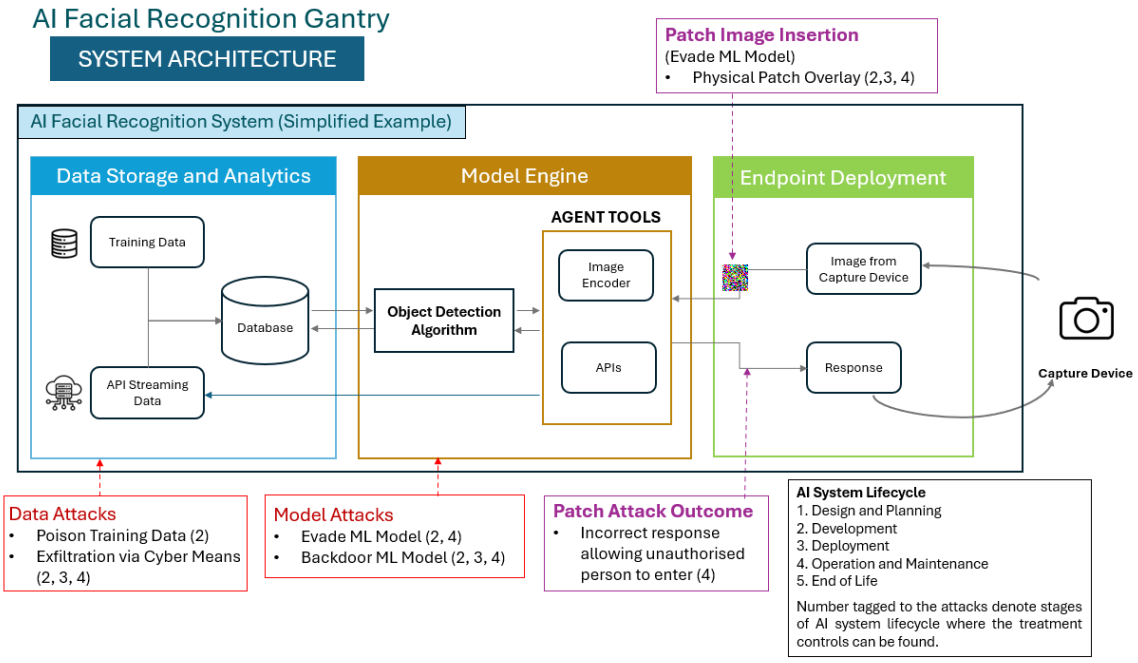

Figure 3. AI Facial Recognition Gantry System Architecture

3.2.1. RISK ASSESSMENT EXAMPLE – EXTRACT ON PATCH ATTACK

The following is an extract from a security risk assessment, specific to an image patch attack.

Risk Scenarios	Impact	Likelihood	Proposed Mitigations	Risk Level
Image Patch Evasion attack. Attacker can use adversarial patches to compromise physical security measures, leading to unauthorised access and potential security breaches.	**Integrity: High** Integrity of the AI facial recognition system will be impacted allowing unauthorised personnel to access the gantry	**Likelihood: Low** Threat actors need to know how the facial recognition AI model works in order to generate a malicious patch that is effective	• Adversarial training • Ensemble model • Multiple sensors • Input Filtering	**Initial Risk Level: High** **Residual Risk Level: Low**

3.2.2. RELEVANT TREATMENT CONTROLS FROM COMPANION GUIDE

To avoid repetition from section 5.1, we outline only the essential controls related to the Patch Attack scenario.

2.2	**Consider security benefits and trade-offs when selecting the appropriate model to use**				
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Implementation done / reason not done**
2.2.2	Consider Model hardening if appropriate.	✓			**Adversarial training** is implemented. **Ensemble Model**: Utilised ensemble approaches that combine multiple facial recognition algorithms. These measures can enhance robustness and resilience against image patch attacks, mitigating the impact of individual vulnerabilities
2.2.3	Consider implementing techniques to strengthen/harden the system apart from strengthening the model itself.	✓			**Multi-Sensor Fusion:** Multiple cameras and lasers used to detect the face.

4.1	**Monitor AI system inputs**				
	Treatment Measures/Controls	**Yes**	**No**	**NA**	**Implementation done / reason not done**
4.1.1	Validate/monitor inputs to the model and system for possible attacks and suspicious activity.	✓			**Additional input filtering layer to detect if abnormal patches are present.** **Having a staff to verify when one is detected.**

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

Technical Testing and System Validation

Efficient testing is an essential component for Security by Design and Privacy by Design, ensuring that AI systems meet the needs and expectations of end-users, deliver value, solve real-world problems, and are safe, reliable, accurate, and beneficial for intended users and purposes.

AI systems can be vulnerable to adversarial attacks where malicious actors manipulate inputs to cause the system to malfunction. Testing helps expose these vulnerabilities and implement safeguards to mitigate them. Repeated iterations can improve the design lifecycle and lead to a deeper understanding of how individual AI components are interacting with each other in an eco-system, which should be secured in its totality.

TYPES OF TESTING

There are three main categories of AI testing, each with varying levels of access to the internal workings of the AI system:

White-Box Testing: In white-box testing, you have complete access to the source code, model weights, and internal logic of the AI system. This allows for very detailed testing, focusing on specific algorithms and code sections. However, it requires significant expertise in the underlying technology and can be time-consuming

Grey-Box Testing: Grey-box testing provides partial access to the AI system. You might have knowledge of the algorithms used but not the specific implementation details. This allows for testing specific functionalities without getting bogged down in the intricate code.

Black-Box Testing: Black-box testing treats the AI system as a complete unit, with no knowledge of its internal workings. This is similar to how a user would interact with the system. Testers focus on inputs, outputs, and expected behaviours.

PROS AND CONS OF BLACK BOX TESTING FOR AI

Black-box testing offers several advantages, particularly for securing sensitive information:

Protects Intellectual Property: By not requiring access to source code or model weights, black box testing safeguards proprietary information and trade secrets.

Focus on User Experience: It prioritises real-world functionality from a user's perspective, ensuring the AI delivers the intended results.

Reduced Expertise Needed: Testers do not need in-depth knowledge of the underlying algorithms, making it more accessible for broader testing teams.

However, it is important to note that black box testing alone might not be sufficient for the most comprehensive form of AI testing, because:

Limited Visibility into Issues: Without understanding the internal workings, it can be difficult to pinpoint the root cause of errors or unexpected behaviours.

Challenges in Debugging: Debugging issues becomes more complex as you cannot isolate problems within the specific algorithms or code sections.

CHALLENGES OF AI TESTING

Despite considerable research to uncover the best methods for enhancing robustness, many countermeasures would fail when subjected to stronger adversarial attacks. The recommended approach would be to subject the AI system iteratively to robustness testing with respect to different defences, using a comprehensive testing tool or system, like running a penetration test.

Such a platform would then subject the test system via not just multiple attacks that will scale upwards progressively but would manage the testing cycles with knowledge to optimise the attack evaluation process, e.g., Black box attacks that do not need the help of insiders. In addition, the project teams can also test the robustness of their AI systems against the full set of known and importantly, unknown adversarial attacks.

Other challenges are:

Non-determinism: resulting from self-learning, i.e. AI-based systems may evolve over time and therefore security properties may degrade.

Test oracle problem: where assigning a test verdict is different and more difficult for AI-based systems, since not all expected results are known a priori.

Data-driven paradigm: AI algorithms, where in contrast to traditional systems, (training) data will predominately determine the output behaviour of the AI.

Developing diverse test datasets: Creating datasets that represent various languages, modalities (text, image, audio), and potential attack vectors.

Evaluating performance across modalities: Measuring the effectiveness of attacks and model robustness across different data types.

Limited testing tools: The need for specialised tools to handle the complexities of blended AI models.

LIST OF AI TESTING TOOLS

AI testing is extremely complex, and the tools listed here will not be always able to reduce its complexity and difficulty.

The list of tools for AI model testing will be split into three categories: Offensive AI Testing Tools, Defensive AI Testing Tools, and Governance AI Testing Tools, based on the primary purpose and functionality of the tools.

Offensive AI Testing Tools

Offensive AI Testing Tools are designed to identify vulnerabilities and weaknesses in AI systems by simulating adversarial attacks or malicious inputs. These tools help evaluate the robustness and security of AI models against various types of attacks, such as adversarial examples, data poisoning, and model extraction.

Defensive AI Testing Tools

Defensive AI Testing Tools, on the other hand, focus on enhancing the robustness and resilience of AI systems against potential threats and vulnerabilities. These tools aim to detect and mitigate the impact of adversarial attacks, natural noises, or other forms of corrupted inputs, ensuring that AI models maintain their intended behaviour and performance. Tools that have both offensive and defensive elements are listed under Offensive Testing.

Governance AI Testing Tools

Governance AI Testing Tools are broader in scope and are primarily concerned with assessing the trustworthiness, fairness, and transparency of AI systems. These tools provide frameworks, guidelines, and resources to evaluate and ensure that AI systems align with principles of responsible AI development, deployment, and governance.

Note: The tools mentioned in these tables are often open-source projects or research prototypes that are still under active development. As such, their functionality, performance, and capabilities may change over time, and they might not always work as intended or as described. It is essential to regularly check for updates, documentation, and community support for these tools, as their features and effectiveness may evolve rapidly. Additionally, some tools might have limited support or documentation, requiring users to have a certain level of expertise and familiarity with the underlying concepts and technologies. Therefore, it is crucial to thoroughly evaluate and validate these tools in a controlled environment before deploying them in production or critical systems. Using highly automated settings may result in violations of cybersecurity misuse legislation that forbids any form of scanning or vulnerability scanning unless permission has been granted. For open-source tools, their long-term maintenance, ease of use, other tools integration, reporting and community adoption may be a concern, especially compared to commercial or enterprise-backed AI security solutions.

OFFENSIVE AI TESTING TOOLS

Tool Name Description	License Type	Model Type	Pros	Cons
Gymnasium[3] Malware Environment for single-agent reinforcement learning environments, with popular reference environments and related utilities (formerly Gym)	Open-source	Various	Provides a toolkit for developing and comparing reinforcement learning algorithms. This makes it possible to write agents that learn to manipulate PE files (e.g., malware) to achieve some objective (e.g., bypass AV) based on a reward provided by taking specific manipulation actions.	Limited to the malware domain.
Deep-Pwning[4] Metasploit for Machine Learning.	Open-source	Various	Comprehensive framework for evaluating robustness of ML models against adversarial attacks. Offers flexibility and customisation options, allowing testers to fine-tune attack parameters and strategies to suit their specific testing requirements.	Requires expertise in adversarial machine learning.
Garak[5] LLM Vulnerability Scanner.	Open-source	LLM, Hugging Face models and public ones.	Specifically designed for testing LLMs for vulnerabilities, i.e. probes for hallucination, data leakage, prompt injection, misinformation, toxicity generation, jailbreaks, and many other weaknesses.	Limited to LLMs, relatively new tool.
Adversarial Robustness Toolbox (ART)[6] Library that helps developers and researchers improve the security of machine learning models.	Open-source	Various but not LLMs	Originated from IBM. Was part of a DARPA project called Guaranteeing AI Robustness Against Deception (GARD). Good for research, with modules for attacks, defences, metrics, estimators, and other	Donated by IBM to the Linux Foundation AI & Data Foundation in 2020 and has lost steam, as no version updates since 2020 and has little new activities.

[3] https://github.com/Farama-Foundation/Gymnasium
[4] https://github.com/cchio/deep-pwning
[5] https://github.com/leondz/garak/
[6] https://github.com/Trusted-AI/adversarial-robustness-toolbox

			functionalities to help secure machine learning pipelines against adversarial threats.	Does not directly address LLM security issues like prompt injection.
CleverHans[7] A Python library for creating and evaluating adversarial examples, benchmarking machine learning models against adversarial examples.	Open-source	Various and developing LLM attacks as well.	Good educational and research library, offering a wide range of attack and defence methods via a modular design. Offers a comprehensive set of tools for generating and analysing adversarial examples. These are carefully crafted inputs designed to deceive machine learning models, helping researchers and developers identify weaknesses in their systems. Various evaluation metrics that go beyond standard accuracy measurements. It includes metrics like robustness, resilience, and adversarial success rates, providing a more comprehensive understanding of a model's performance.	Requires a steep learning curve for beginners to understand all the concepts and effectively utilise. Being a static framework, may not inherently keep pace with the rapidly evolving landscape of adversarial attacks and defence strategies. Documentation and tutorials are focused on computer vision models. While CleverHans offers implementations for popular machine learning frameworks like TensorFlow and PyTorch, it may not support all existing frameworks or the latest updates.
Foolbox[8] A Python toolbox for creating adversarial examples that fools machine learning models.	Open-source	Various but not LLMs	Open-source Python library that offers a wide variety of adversarial attack methods, including gradient-based, score-based, and decision-based attacks, hence more feature-rich, compared to the ART toolkit.	Specialised focus on image classification models and does not cover other areas well.

[7] https://github.com/cleverhans-lab/cleverhans
[8] https://github.com/bethgelab/foolbox

			Also provides defences against these attacks. Provides in-depth tools and techniques for analysing adversarial attacks and security in the context of computer vision tasks.	
Advertorch[9] A PyTorch library for generating adversarial examples and enhancing the robustness of deep neural networks.	Open-source	Various but not LLMs	Offers broader set of attack and defence techniques compared to the ART toolkit, such as universal adversarial perturbations and ensemble-based defences. Allows users to seamlessly apply adversarial attacks and defences to PyTorch models.	Steep Learning Curve. Specifically designed for PyTorch models, which may limit its applicability to frameworks or models from different libraries.
Adversarial Attacks and Defences in Machine Learning (AAD) Framework[10] Python framework for defending machine learning models from adversarial examples.	Open-source	Various but not LLMs	Provides a comprehensive set of tools for evaluating and defending against adversarial attacks on machine learning models, which includes a wider range of attack and defence techniques compared to the ART toolkit, covering areas like evasion, poisoning, and model extraction attacks. Defence techniques include adversarial training, defensive distillation, input transformations, and model ensembles.	High complexity.

[9] https://github.com/BorealisAI/advertorch
[10] https://github.com/changx03/adversarial_attack_defence

DEFENSIVE AI TESTING TOOLS

Tool Name Description	License Type	Model Type	Pros	Cons
CNN Explainer[11] Visualisation tool for explaining CNN decisions.	Open-source	CNN	Helps understand and validate CNN model decisions. A good visualisation system to educate new users via visualisation.	Limited to CNNs only and does not cover any other AI vision model.
Nvidia NeMo[12] A framework for generative AI.	Open-source	LLM	Includes guardrails specifically designed for LLM security, e.g. monitoring, and controlling LLM behaviour during inference, ensuring that generated responses adhere to predefined constraints. It provides mechanisms for detecting and mitigating harmful or inappropriate content, enforcing ethical guidelines, and maintaining user privacy. Guardrails are customizable and adaptable to different use cases and regulatory requirements.	Complex and GPU intensive, thus expensive and affects latency.
AllenAI's AllenNLP[13] An Apache 2.0 NLP research library, built on PyTorch, for developing deep learning models on a wide variety of linguistic tasks.	Open-source	LLM	NLP library that includes guardrails for LLM security: tools for bias detection, fairness assessment, and data governance, helping users build and deploy LLMs responsibly. Designed to be flexible and adaptable to different use cases.	Steep learning curve, complex setup, heavily focused on research and experimentation - some of its features might be more geared towards academic research rather than production-level applications. No new features to be added, tool is only maintained.

[11] https://poloclub.github.io/cnn-explainer/
[12] https://github.com/NVIDIA/NeMo
[13] https://github.com/allenai/allennlp

AI GOVERNANCE TESTING TOOLS

Tool Name Description	License Type	Model Type	Pros	Cons
Assessment List for Trustworthy AI[14] Self-assessment tool for trustworthiness of AI systems.	Open-source	Various	Fairly comprehensive framework for evaluating trustworthiness.	Not an automated tool, requires manual assessment.
OECD AI System Classification[15] Classification and tools for developing trustworthy AI systems.	Open-source	Various	Provides guidelines and resources for trustworthy AI development.	Not a specific testing tool, more of a framework.
Charcuterie[16] Collection of tools for data science and machine learning.	Open-source	Various	Provides a variety of tools for data analysis and model development.	Not specifically focused on testing, more of an assistance tool.
LangKit[17] Open-source text metrics toolkit for monitoring language models.	Open-source	LLM	Helps monitor and evaluate LLM performance, safety, and security.	Limited to LLMs, may not cover broader AI system governance.
AI Verify (IMDA)[18] AI governance testing framework and software toolkit that validates the performance of AI systems through standardised tests.	Open-source	Various	A comprehensive tool designed for AI governance and responsible AI practices. It offers a range of features to support organisations in managing and evaluating their AI systems throughout their lifecycle. Provides guidance on bias detection and mitigation, fairness assessments, and stakeholder engagement.	Does not cover LLM.

[14] https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment
[15] https://www.oecd.org/digital/ieconomy/artificial-intelligence-machine-learning-and-big-data/trusted-ai-systems/
[16] https://github.com/moohax/Charcuterie
[17] https://github.com/whylabs/langkit
18 https://aiverifyfoundation.sg/what-is-ai-verify/

Project Moonshot[19] (IMDA) An LLM Evaluation Toolkit designed to integrate benchmarking, red teaming, and testing baselines. It helps developers, compliance teams, and AI system owners manage LLM deployment risks by providing a seamless way to evaluate their applications' performance, both pre- and post-deployment. This open-source tool is hosted on GitHub and is currently in beta.	Open-source	LLM	Moonshot provides intuitive results, so testing unveils the quality and safety of a model or application in an easily understood manner, even for a non-technical user	Does not cover LLM system security.
threat-composer (AWS labs) A simple threat modelling tool to help humans to reduce time-to-value when threat modelling	Open-source	Various	Identify security issues in the context of own AI system. Provides insights on how to improve.	

CSA does not endorse any commercial product or service. CSA does not attest to the suitability or effectiveness of these services and resources for any particular use case. Any reference to specific commercial products, processes, or services by service mark, trademark, manufacturer, or otherwise, does not constitute or imply their endorsement, recommendation, or favouring by CSA.

[19] https://aiverifyfoundation.sg/project-moonshot/

REFERENCES

Articles

1. LinkedIn: How can you design test AI Systems Safely? [20]
2. Elinext: How to test your medical AI for safety[21]
3. Mathworks: The Road to AI Certification: The importance of Verification and Validation in AI[22]
4. Techforgood Institute: AI Verify Foundation: Shaping the AI landscape of tomorrow[23]
5. FPF.Org: Explaining the Crosswalk Between Singapore's AI Verify Testing Framework and The U.S. NIST AI Risk Management Framework[24]
6. FPF.Org: AIVerify: Singapore's AI Governance Testing Initiative Explained[25]
7. Data Protection Report: Singapore proposes Governance Framework for Generative AI[26]

Standard / Regulatory Bodies

8. NIST: Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence[27]
9. ETSI: Securing Artificial Intelligence Introduction[28]
10. CSA: GUIDELINES FOR AUDITING CRITICAL INFORMATION INFRASTRUCTURE JANUARY 2020 [29]
11. IMDA: Singapore launches AI Verify Foundation to shape the future of international AI standards through collaboration[30]

[20] https://www.linkedin.com/advice/1/how-can-you-design-test-ai-systems-safety
[21] https://www.elinext.com/industries/healthcare/trends/step-by-step-guide-how-to-test-your-medical-ai-for-safety
[22] https://blogs.mathworks.com/deep-learning/2023/07/11/the-road-to-ai-certification-the-importance-of-verification-and-validation-in-ai/
[23] https://techforgoodinstitute.org/blog/articles/ai-verify-foundation-shaping-the-ai-landscape-of-tomorrow/
[24] https://fpf.org/blog/explaining-the-crosswalk-between-singapores-ai-verify-testing-framework-and-the-u-s-nist-ai-risk-management-framework/
[25] https://fpf.org/blog/ai-verify-singapores-ai-governance-testing-initiative-explained/
[26] https://www.dataprotectionreport.com/2024/02/singapore-proposes-governance-framework-for-generative-ai/
[27] https://www.nist.gov/artificial-intelligence/executive-order-safe-secure-and-trustworthy-artificial-intelligence/test
[28] https://portal.etsi.org/Portals/0/TBpages/SAI/Docs/2021-12-ETSI_SAI_Introduction.pdf
[29] https://www.csa.gov.sg/docs/default-source/csa/documents/legislation_supplementary_references/guidelines_for_auditing_critical_information_infrastructure.pdf?sfvrsn=8fe3dab7_0
[30] https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/singapore-launches-ai-verify-foundation-to-shape-the-future-of-international-ai-standards-through-collaboration

ANNEX B

AI Security Defences and their trade-offs

As AI becomes a cornerstone of innovation and national security, protecting its core components becomes paramount. Implementing a multi-layered, dynamically adaptive approach that combines technical safeguards (encryption, air gapping) with robust security protocols (access control, monitoring) and a culture of cyber awareness within organisations is crucial to safeguarding these new "crown jewels" of the digital age.

DEFENDING AI MODELS

Importantly, the models themselves are “fragile” and can be easily attacked using image or text adversarial robustness attacks, or the LLMs could be attacked using malicious prompts.

The table below gives a short summary on the techniques to defend AI systems (non LLM) from examples of adversarial attack.

Defence	Description
Adversarial Training	Train AI model using adversarial samples
Ensemble Models	Utilise blended models to perform a task, and compare their results
Defensive Distillation	Train AI model using class probabilities, instead of discrete class labels, to learn more information about data
Adversarial Detection • Compression • Blurring	Attempt to identify whether an input is an adversarial sample Counter Image Attacks
Explainability	Identify which part of the input had the highest impact in producing the resulting classification. To discover how and why the attack is happening and what makes it work?

Table C1: Countermeasures with description

ADVERSARIAL TRAINING

The most viable method is to introduce adversarial training into the training dataset and retrain the system, i.e., to simply generate and then incorporate adversarial examples into the training process. There are toolsets to do this. In addition, some of the latest image object recognition algorithms e.g. Yolo5, would incorporate adversarial training within this workflow when running training. This will improve model robustness but may not eliminate it.

Hence, the main goal of Adversarial Training is to make a model more robust against adversarial attacks by adding adversarial samples into the model's training dataset. Like adding augmented samples, such as mirrored or cropped images, into the training dataset to improve generalisation. An existing attack algorithm is used to generate these adversarial samples, and there are several variants that utilise different algorithms to generate the adversarial samples for training. Adversarial Training can also be thought of as a brute-force approach, which aims to widen the input distribution of the model so that the boundaries between classes become more accurate.

LIMITATIONS OF ADVERSARIAL TRAINING

Adversarial Training requires additional time to train the model using adversarial samples. Iterative attack algorithms such as Projected Gradient Descent (PGD) requires a much larger time cost, making it difficult to be used for training with massive datasets.

Adversarial Training is mainly effective against the adversarial samples the model was trained against. To note that models, even with Adversarial Training, are susceptible to black-box attacks that utilise a locally trained substitute model to generate adversarial samples. Another technique proposed in "Ensemble Adversarial Training: Attacks and Defences" by Tramèr et. Al.[31] adds random perturbations to an input before running the adversarial attacks on the perturbed input, successfully bypassing the Adversarial Training defence.

[31] https://arxiv.org/abs/1705.07204

ENSEMBLE MODELS

Another intuitive approach to enhance model robustness would be to use multiple models (best to be handling different aspects of the recognition problem) to either detect an attack or to prevent a bypass attack. For example, as depicted in the diagram below, if there were a second AI head detector, the person detector even though fooled by the physical logo on the attacker's shirt, the head detector would not be fooled. Additionally, if there are multiple recognition models, the summation results of different AI systems could still be functional, despite one model being successfully attacked.

LIMITATIONS OF ENSEMBLE MODEL

As multiple models are used on each input, the use of ensemble methods will require additional resources, more memory and computational power for each classification. Ensembles of models may also require more time for development and be more difficult to be used in scenarios where fast, real-time predictions may be required.

DEFENSIVE DISTILLATION

Distillation, also known as Teacher-Student Models, is a procedure which utilises knowledge obtained from a trained 'teacher' Deep Neural Network (DNN) to train a second 'student' DNN. The classes of the labelled training data are known as hard labels, and the output classifications of the 'teacher' DNN are known as soft labels which captures probability distributions indicating how confident the model is for each class. The 'student' DNN is trained using soft labels and the softer predictions make it harder to fool the student which has learnt a more nuanced representation of the dataset. This makes the DNN more robust to adversarial attacks.

LIMITATIONS OF DEFENSIVE DISTILLATION

However, defensively distilled models are still vulnerable to various black-box attacks, due to the strong transferability of adversarial samples generated by these attacks. Modified versions of existing attack algorithms, such as the modified Papernot's attack, have also successfully bypassed defensive distillation.

UTILISING EXPLAINABILITY

A different approach to Adversarial Detection involves the incorporation of Explainable AI (XAI) techniques, which 'explain' the reasons that led to the AI model's prediction. XAI is an emerging field in machine learning that aims to explain predictions made by AI models to improve accuracy, fairness and at the same time aid in the detection of possible anomalies or adversarial attacks. In order to understand the complex black boxes that are AI models, XAI is expected to provide explanations interpretable by humans with clear and simple visualisations.

The main strength of this method is its ability to gain insights into weaknesses present in the model, such as when the reasons leading to the resultant prediction are incorrect. A local interpreter is built to explain the factors that cause adversarial samples to be wrongly classified by the target model.

Furthermore, adversarial samples that exploit these weaknesses can then be generated for use in adversarial training, allowing the model to overcome them. In addition, as the interpretation technique is general to all classifiers, this method can be applied to improve any type of model that supports XAI techniques.

Finally, AI Explainability techniques can be applied to the suspected adversarial inputs, providing visualisations to human operators explaining why these inputs are potentially malicious. The operators can then find out if the detection was a false positive and work on improving the detection model. Otherwise, if the detection was accurate, problems with the defended model can potentially be identified, and the appropriate countermeasures can be applied.

A COMBINATION OF TECHNIQUES

Multiple countermeasures can be used to complement one another, creating a defence-in-depth approach as a higher level of using ensemble defences with differently configured AI models. to would ensure even stronger robustness against adversarial attacks.

DEFENDING YOUR AI SYSTEMS BEYOND THE MODELS

After defending the AI models, since it is still possible to subvert, poison and tamper with the AI system, enhanced infrastructural security measures would have to be added to counter the offensive TTPs that were identified during the risk assessment. The key areas to focus on include:

Continuous monitoring and threat intelligence: Staying informed about the latest threats and vulnerabilities through threat intelligence feeds and security monitoring tools.

Implementing security best practices: This includes basic hygiene measures like patching vulnerabilities, using strong passwords, and implementing multi-factor authentication. Increase system segregation and isolation using containers, VMs, air gaps, firewalls etc.

User awareness training: Educating employees about social engineering tactics and how to identify and avoid phishing attacks.

Security testing and vulnerability assessments: Regularly testing systems for vulnerabilities and implementing security controls to mitigate risks.

Investing in Security Automation: Utilise automation tools to streamline security processes and improve efficiency.

By staying proactive and adapting to the evolving threat landscape that heralds powerful AI-armed APT intruders, organisations can build stronger AI crown jewel defences and mitigate the impact of cyberattacks. Remember, cybersecurity is an ongoing process, not a one-time fix.

AI SECURITY DEFENCES AND THEIR TRADE-OFFS

It is prudent to start implementing countermeasures to protect AI models against attacks early, even as there remain unknowns.

- No one method or countermeasure can reliably defend against all attacks
- Limited awareness and know-how in understanding and operationalising adversarial countermeasures, exacerbated by the complexity of AI models that also makes it difficult to prove how and which defence method will work against some subset of attacks
- As with other changes to the AI model and system, modifications to the model to enhance defences can have impact on model/ system performance

Regardless, traditional security practices continue to be relevant, and provide a good foundation for securing cutting-edge technologies like AI, even as work in this space continues to evolve.

REFERENCES

Standards / Regulatory Bodies

1. **NIST**
    a. NIST: NIST Secure Software Development Framework for Generative AI and for Dual Use Foundation Models Virtual Workshop [32]

NIST: Executive Order 14110 on Safe, Secure, and Trustworthy Artificial Intelligence (October 2023)[33]

b. NIST: AI RMF Knowledge Base[34]
c. NIST: A USAISI Workshop: Collaboration to Enable Safe and Trustworthy AI[35]
d. NIST: USAISI Workshop Slides[36]
e. NIST: Artificial Intelligence[37]
f. NIST: Biden-Harris Administration Announces first ever consortium dedicated to AI Safety[38]
g. NIST: NIST Secure Software Development Framework for Generative AI and for Dual Use Foundation Models Virtual Workshop[39]

2. **ENISA**
    a. ENISA: Artificial Intelligence[40]
    b. ENISA: EU Elections at Risk with Rise of AI-Enabled Information Manipulation [41]
    c. ENISA: Multilayer Framework for Good Cybersecurity Practices for AI [42]
    d. ENISA: Cybersecurity and AI and Standardisation Report[43]
    e. ENISA: Artificial Intelligence Cybersecurity Challenges [44]
    f. ENISA: Is Secure and Trusted AI Possible? The EU Leads the Way [45]
    g. ENISA: Cybersecurity and privacy in AI - Medical imaging diagnosis [46]
    b. ENISA: Cybersecurity and privacy in AI - Forecasting demand on electricity grids

3. **NCSC**
    a. NCSC: Guidelines for secure AI system development [4748]

4. **NSA**
    a. Deploying AI Systems Securely: Best Practices for Deploying Secure and Resilient AI Systems[49]

5. **Singapore**
    a. CSA: Codes of Practice[50]
    b. PDPC: Primer for 2nd Edition of AI Gov Framework [51]
    c. Gov.SG ISAGO[52]
    d. PDPC: Advisory Guidelines On use of Personal Data In AI Recommendation and Decision Systems[53]

6. **Standards and Guides**
    a. ISO/IEC 42001:2023 Information Technology: Artificial Intelligence: Management System[54]
    b. ISO/IEC 23894:2023 Information Technology: Artificial Intelligence: Guidance on Risk Management[55]
    c. OWASP AI Security and Privacy Guide[56]

7. **Others**
    a. PartnershiponAI[57]
    b. AJL[58]
    c. International Telecommunication Union (ITU)[59]
    d. OECD Artificial Intelligence [60]

8. **GitHub Repositories**
   a. Privacy Library of Threats 4 Artificial Intelligence[61]
   b. Guardrails.AI[62]
   c. PyDP: Differential Privacy[63]
   d. IBM Differential Privacy Library[64]
   e. TenSEAL: Encrypting Tensors with Microsoft SEAL[65]
   f. SyMPC: Extends Pysft with SMPC Support[66]
   g. PyVertical: privacy-preserving, vertical federated learning using syft [67]

9. **Articles**
   a. CSO: NIST releases expanded 2.0 version of the Cybersecurity Framework[68]
   b. Technologylawdispatch.com: ENISA Releases Comprehensive Framework
   f. Kim & Chang: South Korea: Legislation on Artificial Intelligence to Make Significant Progress [73]
   g. MetaNews: South Korean Government Says No Copyright for AI Content[74]
   h. East Asia Forum: The future of AI policy in China [75]
   i. Reuters: China approves over 40 AI models for public use in past six months[76]

---

[32] https://www.nist.gov/news-events/events/nist-secure-software-development-framework-generative-ai-and-dual-use-foundation
[33] https://www.nist.gov/artificial-intelligence/executive-order-safe-secure-and-trustworthy-artificial-intelligence
[34] https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF
[35] https://www.nist.gov/news-events/events/usaisi-workshop-collaboration-enable-safe-and-trustworthy-ai
[36] https://www.nist.gov/system/files/documents/noindex/2023/11/20/USAISI-workshop-slides%20%28combined%20final%29.pdf
[37] https://www.nist.gov/artificial-intelligence/artificial-intelligence-safety-institute
[38] https://www.nist.gov/news-events/news/2024/02/biden-harris-administration-announces-first-ever-consortium-dedicated-ai
[39] https://www.nist.gov/news-events/events/nist-secure-software-development-framework-generative-ai-and-dual-use-foundation
[40] https://www.enisa.europa.eu/topics/iot-and-smart-infrastructures/artificial_intelligence
[41] https://www.enisa.europa.eu/news/eu-elections-at-risk-with-rise-of-ai-enabled-information-manipulation
[42] https://www.enisa.europa.eu/publications/multilayer-framework-for-good-cybersecurity-practices-for-ai
[43] https://www.enisa.europa.eu/publications/cybersecurity-of-ai-and-standardisation/@@download/fullReport
[44] https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges
[45] https://www.enisa.europa.eu/news/is-secure-and-trusted-ai-possible-the-eu-leads-the-way
[46] https://www.enisa.europa.eu/publications/cybersecurity-and-privacy-in-ai-medical-imaging-diagnosis
[47] https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development
[48] https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf
[49] https://www.nsa.gov/Press-Room/Press-Releases-Statements/Press-Release-View/Article/3741371/nsa-publishes-guidance-for-strengthening-ai-system-security/
[50] https://www.csa.gov.sg/legislation/Codes-of-Practice
[51] https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/primer-for-2nd-edition-of-ai-gov-framework.pdf
[52] http://go.gov.sg/isago
[53] https://www.pdpc.gov.sg/guidelines-and-consultation/2024/02/advisory-guidelines-on-use-of-personal-data-in-ai-recommendation-and-decision-systems
[54] https://www.iso.org/standard/81230.html
[55] https://www.iso.org/standard/77304.html
[56] https://owasp.org/www-project-ai-security-and-privacy-guide/
[57] https://partnershiponai.org/
[58] https://www.ajl.org
[59] https://www.itu.int/
[60] https://www.oecd.org/digital/artificial-intelligence/
[61] https://plot4.ai/
[62] https://github.com/guardrails-ai/guardrails
[63] https://github.com/OpenMined/PyDP
[64] https://github.com/IBM/differential-privacy-library
[65] https://github.com/OpenMined/TenSEAL
[66] https://github.com/OpenMined/SyMPC
[67] https://github.com/OpenMined/PyVertical
[68] https://www.csoonline.com/article/1310046/nist-releases-expanded-2-0-version-of-the-cybersecurity-framework.html
[73] https://www.kimchang.com/en/insights/detail.kc?sch_section=4&idx=26935
[74] https://metanews.com/south-korean-government-says-no-copyright-for-ai-content/
[75] https://eastasiaforum.org/2023/09/27/the-future-of-ai-policy-in-china/
[76] https://www.reuters.com/technology/china-approves-over-40-ai-models-public-use-past-six-months-2024-01-29/

for Ensuring Cybersecurity in the Lifecycle of AI Systems [69]

c. DataGuidance: ENISA releases four reports on AI and Cybersecurity [70]
d. KoreaTimes: Korea issues first AI ethics checklist [71]
e. Dig.watch: South Korea to boost trust in AI with watermarking initiative[72]
j. DataNami: Artificial Intelligence Leaders Partner with Cloud Security Alliance to Launch the AI Safety Initiative [77]
k. World Economic Forum: Why we need to care about responsible AI in the age of the algorithm [78]
l. What is Confidential Computing? | Data Security in Cloud Computing (Anjuna)[79]
m. What is Confidential Computing? (NVIDIA Blog)[80]

[69] https://www.technologylawdispatch.com/2023/06/data-cyber-security/enisa-releases-comprehensive-framework-for-ensuring-cybersecurity-in-the-lifecycle-of-ai-systems/
[70] https://www.dataguidance.com/news/eu-enisa-releases-four-reports-ai-and-cybersecurity
[71] https://m.koreatimes.co.kr/pages/article.asp?newsIdx=352971
[72] https://dig.watch/updates/south-korea-to-boost-trust-in-ai-with-watermarking-initiative
[77] https://www.datanami.com/this-just-in/artificial-intelligence-leaders-partner-with-cloud-security-alliance-to-launch-the-ai-safety-initiative/
[78] https://www.weforum.org/agenda/2023/03/why-businesses-should-commit-to-responsible-ai/
[79] https://www.anjuna.io/blog/confidential-computing-a-new-paradigm-for-complete-cloud-security
[80] https://docs.nvidia.com/nvtrust/index.html

Advisory and Cloud Providers

10. **Google**
    a. Google: Introducing Google's Secure AI Framework[81]
    b. Google: OCISO Securing AI Similar or Different? [82]

11. **Microsoft**
    a. Microsoft: Azure Platform[83]
    b. Microsoft: Introduction to Azure Security[84]
    c. Microsoft: Responsible AI[85]
    d. Microsoft: Responsible AI Principles and Approach[86]
    e. Microsoft: AI Fairness Checklist[87]
    f. Microsoft: AI Lab project: Responsible AI dashboard[88]
    g. Microsoft: Our commitments to advance safe, secure, and trustworthy AI[89]

12. **Amazon Web Services**
    a. Secure approach to generative AI[90]

[81] https://blog.google/technology/safety-security/introducing-googles-secure-ai-framework/
[82] https://services.google.com/fh/files/misc/ociso_securing_ai_different_similar.pdf
[83] https://www.microsoft.com/en-us/ai/ai-platform
[84] https://learn.microsoft.com/en-us/azure/security/fundamentals/overview
[85] https://www.microsoft.com/en-us/ai/responsible-ai
[86] https://www.microsoft.com/en-us/ai/principles-and-approach
[87] https://www.microsoft.com/en-us/research/project/ai-fairness-checklist/
[88] https://www.microsoft.com/en-us/ai/ai-lab-responsible-ai-dashboard
[89] https://blogs.microsoft.com/on-the-issues/2023/07/21/commitment-safe-secure-ai/
[90] https://aws.amazon.com/ai/generative-ai/security/

Consultancies

13. **Deloitte**
    a. Deloitte: Trustworthy AI[91]
    b. Deloitte: Omnia AI[92]
    c. Deloitte AI Institute: The State of Generative AI in the Enterprise: Now Decides Next[93]

14. **EY**
    a. EY: How to navigate generative AI use at work[94]
    b. EY: EY's commitment to developing and using AI ethically and responsibly[95]
    c. EY: Making Artificial Intelligence and Machine Learning trustworthy and ethical[96]

15. **KPMG**
    a. KPMG: AI security framework design[97]
    b. KPMG: Trust in Artificial Intelligence[98]

16. **PwC**
    a. PwC: Balancing Power and Protection: AI in Cybersecurity and Cybersecurity in AI [99]
    b. PWC: What is Responsible AI[100]

17. **Research Papers**
    a. TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI[101]
    b. China Academy of Information and Communications Technology: Whitepaper on Trustworthy Artificial Intelligence[102]
    c. Trustworthy AI: From Principles to Practices [103]

[91] https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-ai-framework.html
[92] https://www2.deloitte.com/ca/en/pages/deloitte-analytics/articles/omnia-artificial-intelligence.html
[93] https://www2.deloitte.com/us/en/pages/deloitte-analytics/articles/advancing-human-ai-collaboration.html#
[94] https://www.ey.com/en_us/consulting/video-how-to-navigate-generative-ai-use-at-work
[95] https://www.ey.com/en_gl/ai/principles-for-ethical-and-responsible-ai
[96] https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/consulting/ey-making-artificial-intelligence-and-machine-learning-trustworthy-and-ethical.pdf
[97] https://kpmg.com/us/en/capabilities-services/advisory-services/cyber-security-services/cyber-strategy-governance/security-framework.html
[98] https://kpmg.com/xx/en/home/insights/2023/09/trust-in-artificial-intelligence.html
[99] https://www.pwc.com/m1/en/publications/balancing-power-protection-ai-cybersecurity.html
[100] https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai.html
[101] https://arxiv.org/abs/2306.06924
[102] http://www.caict.ac.cn/english/research/whitepapers/202110/P020211014399666967457.pdf
[103] https://arxiv.org/abs/2110.01167