TrustIoT Framework for Industry 4.0

"AI Model Security"

|  |  |
| --- | --- |
| Document Classification: | Internal |
| Document Ref. | *TrustIoT Framework for Industry 4.0* |
| Version: | *1* |
| Document Author: | *Jibran Saleem* |
| Document Owner: |  |

**Revision History**

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Revision Author** | **Summary of Changes** |
|  |  |  |  |
|  |  |  |  |

**Distribution**

|  |  |
| --- | --- |
| **Name** | **Title** |
|  |  |
|  |  |
|  |  |

**Approval**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Position** | **Signature** | **Date** |
|  |  |  |  |

Table of Contents

[1. Introduction 4](#_Toc176334290)

[2. Purpose 4](#_Toc176334291)

[3. Scope 4](#_Toc176334292)

[4. Policy Statement 4](#_Toc176334293)

[4.1. Model Development and Training 4](#_Toc176334294)

[4.2. Model Deployment and Integration 4](#_Toc176334295)

[4.3. Model Monitoring and Maintenance 4](#_Toc176334296)

[4.4. Adversarial Attacks and Defences 5](#_Toc176334297)

[5. Responsibilities 5](#_Toc176334298)

[6. Breaches of Policy 5](#_Toc176334299)

[7. Document Management 5](#_Toc176334300)

# Introduction

Artificial Intelligence (AI) models are increasingly integrated into IoT systems to enable advanced analytics, automation, and decision-making. However, AI models can be vulnerable to various security threats, such as adversarial attacks, data poisoning, and model theft. This policy outlines the organisation's commitment to ensuring the security and integrity of AI models used within the IoT ecosystem.

# Purpose

The purpose of this policy is to establish a framework for the secure development, deployment, and management of AI models within the organisation’s IoT infrastructure. This policy aims to:

* Protect AI models from unauthorised access, modification, and theft.
* Ensure the integrity and reliability of AI model outputs.
* Mitigate the risk of adversarial attacks and other security threats targeting AI models.
* Maintain the confidentiality of sensitive data used for training and inference.

# Scope

This policy applies to all AI models developed, deployed, or utilised within the organisation's IoT environment, regardless of their specific purpose or application.

# Policy Statement

## Model Development and Training

* **Secure Development Practices:** AI models shall be developed using secure coding practices and robust software development lifecycle (SDLC) processes.
* **Data Security:** Data used for training AI models shall be protected against unauthorised access, modification, or exfiltration.
* **Data Provenance:** The origin and lineage of training data shall be tracked and documented to ensure its authenticity and integrity.
* **Bias and Fairness:** Measures shall be taken to identify and mitigate potential biases in training data and model outputs.

## Model Deployment and Integration

* **Secure Deployment:** AI models shall be deployed in secure environments, utilising appropriate access controls and encryption mechanisms to protect against unauthorised access or tampering.
* **Model Integrity:** Mechanisms shall be implemented to verify the integrity of deployed models and detect any unauthorised modifications.
* **Input Validation:** Input data provided to AI models shall be validated and sanitised to prevent adversarial attacks or injection of malicious data.

## Model Monitoring and Maintenance

* **Performance Monitoring:** The performance and accuracy of deployed AI models shall be continuously monitored to identify any degradation or anomalies.
* **Model Updates:** Models shall be periodically retrained or updated to incorporate new data and address potential biases or performance issues.
* **Version Control:** Version control shall be maintained for AI models to track changes and enable rollback to previous versions if necessary.

## Adversarial Attacks and Defences

* **Adversarial Robustness:** AI models shall be designed and tested for resilience against adversarial attacks, such as evasion attacks, poisoning attacks, and model extraction attacks.
* **Defence Mechanisms:** Appropriate defence mechanisms, such as adversarial training, input sanitisation, or model hardening, shall be implemented to mitigate the risk of adversarial attacks.

# Responsibilities

* **Information Security Officer:** Responsible for overseeing the implementation and enforcement of this policy.
* **Data Scientists and AI Developers:** Responsible for developing and deploying AI models in accordance with this policy.
* **IT Department:** Responsible for providing secure infrastructure and tools for AI model development, deployment, and monitoring.
* **Model Owners:** Responsible for ensuring the ongoing security and maintenance of their respective AI models.

# Breaches of Policy

Non-compliance with this policy may result in disciplinary action, up to and including termination of employment or contractual relationships.

# Document Management

This document is valid as of [dd/mm/yyyy].

This document is reviewed periodically and at least annually to ensure compliance with the following prescribed criteria.

* Compliant with the Internet of Things (IoT) Security Framework for Industry 4.0.
* Legislative requirements defined by law, where appropriate.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

[Name 1]

Manager