TrustIoT Framework for Industry 4.0

"AI Specific Threats and Vulnerabilities"

|  |  |
| --- | --- |
| 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](#_Toc176334486)

[2. Purpose 4](#_Toc176334487)

[3. Scope 4](#_Toc176334488)

[4. Policy Statement 4](#_Toc176334489)

[4.1. Adversarial Attacks 4](#_Toc176334490)

[4.2. Data Poisoning 4](#_Toc176334491)

[4.3. Model Theft and Reverse Engineering 4](#_Toc176334492)

[4.4. Evasion Attacks 5](#_Toc176334493)

[4.5. Model Bias and Fairness 5](#_Toc176334494)

[5. Responsibilities 5](#_Toc176334495)

[6. Breaches of Policy 5](#_Toc176334496)

[7. Document Management 5](#_Toc176334497)

# Introduction

The integration of Artificial Intelligence (AI) into the Internet of Things (IoT) ecosystem introduces new and unique security challenges. AI models, while enabling advanced functionalities, can be susceptible to various threats and vulnerabilities that can compromise their integrity, confidentiality, and availability. This policy outlines the organisation's commitment to understanding and mitigating these AI-specific threats and vulnerabilities within the IoT environment.

# Purpose

The purpose of this policy is to establish a framework for identifying, assessing, and mitigating AI-specific threats and vulnerabilities within the organisation's IoT infrastructure. This policy aims to:

* Enhance the security and resilience of AI models deployed in IoT systems.
* Protect against adversarial attacks, data poisoning, and other malicious activities targeting AI models.
* Ensure the integrity and trustworthiness of AI-driven decisions and actions.

# Scope

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

# Policy Statement

## Adversarial Attacks

* **Threat Modelling:** Regular threat modelling exercises shall be conducted to identify and assess potential adversarial attacks targeting AI models, including evasion attacks, poisoning attacks, and model extraction attacks.
* **Adversarial Training:** AI models shall be trained using adversarial examples to improve their robustness against malicious inputs.
* **Input Validation and Sanitisation:** Input data provided to AI models shall be rigorously validated and sanitised to prevent the injection of adversarial examples or malicious data.

## Data Poisoning

* **Data Integrity:** Measures shall be implemented to ensure the integrity and authenticity of training data used for AI model development, including data provenance tracking and validation.
* **Secure Data Sources:** Training data shall be sourced from trusted and reliable sources, minimizing the risk of contamination or manipulation.
* **Anomaly Detection:** Anomaly detection techniques shall be employed to identify unusual or suspicious patterns in training data that may indicate data poisoning attempts.

## Model Theft and Reverse Engineering

* **Access Controls:** Strict access controls shall be enforced to protect AI models from unauthorised access, modification, or theft.
* **Encryption:** Sensitive AI models and their associated data shall be encrypted at rest and in transit.
* **Code Obfuscation:** Code obfuscation techniques may be employed to make it more difficult for adversaries to reverse engineer or extract sensitive information from AI models.

## Evasion Attacks

* **Robustness Testing:** AI models shall be subjected to rigorous testing to evaluate their resilience against evasion attacks, where adversaries attempt to craft inputs that bypass or deceive the model's detection capabilities.
* **Adversarial Example Generation:** Techniques for generating adversarial examples shall be utilised to identify potential weaknesses in AI models and improve their robustness.

## Model Bias and Fairness

* **Bias Assessment:** AI models shall be regularly assessed for potential biases that may lead to discriminatory or unfair outcomes.
* **Bias Mitigation:** Measures shall be taken to mitigate or address identified biases in training data or model algorithms.
* **Fairness:** AI models shall be designed and trained to promote fairness and avoid discriminatory outcomes.

# 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 that are resistant to adversarial attacks and free from bias.
* **IT Department:** Responsible for providing secure infrastructure and tools for AI model development, deployment, and monitoring.
* **Security Operations Centre (SOC):** Responsible for monitoring AI model activity, detecting anomalies, and responding to potential threats.

# 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