TrustIoT Framework for Industry 4.0

"AI Robustness against Adversarial Attacks"

|  |  |
| --- | --- |
| 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](#_Toc190905309)

[2. Purpose 4](#_Toc190905310)

[3. Scope 4](#_Toc190905311)

[4. Policy Statement 4](#_Toc190905312)

[4.1. Adversarial Threat Modelling 4](#_Toc190905313)

[4.2. Robust Model Development and Training 4](#_Toc190905314)

[4.3. Input Validation and Sanitisation 4](#_Toc190905315)

[4.4. Adversarial Detection and Defence 5](#_Toc190905316)

[4.5. Continuous Monitoring and Improvement 5](#_Toc190905317)

[5. Responsibilities 5](#_Toc190905318)

[6. Breaches of Policy 5](#_Toc190905319)

[7. Document Management 5](#_Toc190905320)

# Introduction

Artificial Intelligence (AI) models deployed in IoT systems are susceptible to adversarial attacks, where malicious actors craft inputs designed to deceive or manipulate the model's behaviour. These attacks can have severe consequences, compromising the integrity, availability, and confidentiality of IoT systems and data. This policy outlines the organisation's commitment to developing and implementing techniques to enhance the resilience of AI models against such attacks.

# Purpose

The purpose of this policy is to establish a framework for proactively addressing the risk of adversarial attacks on AI models within the organisation's IoT infrastructure. This policy aims to:

* Identify and assess potential adversarial threats to AI models.
* Develop and implement robust AI models that are resistant to adversarial manipulation.
* Employ input validation and sanitisation techniques to detect and mitigate adversarial inputs.
* Continuously monitor and improve the resilience of AI models against evolving attack techniques.

# 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

## Adversarial Threat Modelling

* **Threat Identification:** Potential adversarial threats and attack vectors targeting AI models shall be identified and analysed, considering the specific context and use case of each model.
* **Vulnerability Assessment:** Vulnerabilities in AI models and their associated data pipelines that could be exploited by adversaries shall be assessed.
* **Impact Analysis:** The potential impact of successful adversarial attacks on the organisation's operations, data, and reputation shall be evaluated.

## Robust Model Development and Training

* **Adversarial Training:** AI models shall be trained using adversarial examples to improve their robustness against malicious inputs.
* **Defensive Distillation:** Techniques such as defensive distillation may be employed to enhance model resilience against adversarial perturbations.
* **Diversity and Redundancy:** Ensembles of diverse models or redundant systems may be utilised to reduce the impact of successful attacks on individual models.

## Input Validation and Sanitisation

* **Input Validation:** Input data provided to AI models shall be rigorously validated against predefined schemas or expected ranges to detect and reject anomalous or potentially adversarial inputs.
* **Sanitisation:** Input data may be sanitised or pre-processed to remove or neutralise potential adversarial perturbations.

## Adversarial Detection and Defence

* **Anomaly Detection:** Anomaly detection mechanisms shall be implemented to identify inputs or model outputs that deviate significantly from expected patterns, potentially indicating adversarial activity.
* **Runtime Monitoring:** AI models shall be monitored during operation to detect any signs of adversarial manipulation or unexpected behaviour.
* **Active Defence:** Active defence mechanisms, such as adversarial example detection or input transformation, may be employed to proactively defend against adversarial attacks.

## Continuous Monitoring and Improvement

* **Ongoing Research and Development:** The organisation shall stay abreast of the latest research and developments in adversarial machine learning and adapt its defences accordingly.
* **Red Teaming:** Periodic red teaming exercises may be conducted to simulate adversarial attacks and evaluate the effectiveness of the organisation's defences.
* **Lessons Learned:** Lessons learned from adversarial attacks or red teaming exercises shall be incorporated into the development and deployment of future AI models.

# Responsibilities

* **Information Security Officer:** Responsible for overseeing the implementation and enforcement of this policy.
* **Data Scientists and AI Developers:** Responsible for developing and deploying robust AI models that are resistant to adversarial attacks.
* **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 adversarial attacks.

# 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