TrustIoT Framework for Industry 4.0

"AI Model Inversion 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](#_Toc176334498)

[2. Purpose 4](#_Toc176334499)

[3. Scope 4](#_Toc176334500)

[4. Policy Statement 4](#_Toc176334501)

[4.1. Model Protection 4](#_Toc176334502)

[4.2. Access Control 4](#_Toc176334503)

[4.3. Monitoring and Detection 4](#_Toc176334504)

[4.4. Response and Mitigation 5](#_Toc176334505)

[5. Responsibilities 5](#_Toc176334506)

[6. Breaches of Policy 5](#_Toc176334507)

[7. Document Management 5](#_Toc176334508)

# Introduction

Artificial Intelligence (AI) models, particularly those deployed in Internet of Things (IoT) systems, can be targets of model inversion attacks. These attacks aim to extract sensitive information or reverse-engineer the model's underlying structure and parameters. Such breaches can lead to the compromise of confidential data, intellectual property theft, and the potential for malicious manipulation of the AI system's behaviour. This policy outlines the organisation's commitment to protecting against model inversion attacks and preserving the confidentiality and integrity of its AI models.

# Purpose

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

* Safeguard the confidentiality of AI models and the sensitive data they process.
* Prevent unauthorised access, extraction, or reverse engineering of AI models.
* Detect and respond to attempts at model inversion attacks.
* Maintain the integrity and trustworthiness of AI-driven decisions and actions.

# Scope

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

# Policy Statement

## Model Protection

* **Access Control:** Strict access controls shall be enforced to limit access to AI models, training data, and model outputs to authorised personnel and systems only.
* **Encryption:** Sensitive AI models and their associated data shall be encrypted at rest and in transit using strong encryption algorithms and secure key management practices.
* **Obfuscation and Hardening:** Code obfuscation and other model hardening techniques may be employed to make it more difficult for adversaries to reverse engineer or extract sensitive information from AI models.

## Access Control

* **Role-Based Access Control (RBAC):** Access to AI models, training data, and model outputs shall be granted based on user roles and responsibilities within the organisation.
* **Need-to-Know Basis:** Access shall be granted on a need-to-know basis, limiting exposure to sensitive information and minimising the risk of unauthorised model extraction.
* **Regular Reviews:** Access rights shall be periodically reviewed and updated to ensure they remain appropriate and aligned with current roles and responsibilities.

## Monitoring and Detection

* **Anomaly Detection:** Anomaly detection mechanisms shall be implemented to identify unusual or suspicious patterns of access or usage of AI models, potentially indicating model inversion attempts.
* **Logging and Auditing:** All access to and interactions with AI models shall be logged and audited for traceability and accountability.
* **Intrusion Detection Systems (IDS):** Network-based and host-based IDS may be employed to detect and alert on potential model inversion attacks.

## Response and Mitigation

* **Incident Response:** A well-defined incident response plan shall be in place to address suspected or confirmed model inversion attacks, including containment, eradication, and recovery procedures.
* **Model Updates and Redeployment:** In the event of a successful model inversion attack, affected models shall be updated or redeployed with enhanced security measures.
* **Lessons Learned:** Post-incident analysis shall be conducted to identify vulnerabilities and improve defences against future model inversion attempts.

# 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 with appropriate security measures to protect against model inversion attacks.
* **IT Department:** Responsible for providing secure infrastructure and tools for AI model management and protection.
* **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. Additionally, successful model inversion attacks may result in significant financial loss, reputational damage, and legal consequences for the organisation.

# 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