TrustIoT Framework for Industry 4.0

"Membership Inference 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](#_Toc176334513)

[2. Purpose 4](#_Toc176334514)

[3. Scope 4](#_Toc176334515)

[4. Policy Statement 4](#_Toc176334516)

[4.1. Model Protection and Privacy 4](#_Toc176334517)

[4.2. Differential Privacy 4](#_Toc176334518)

[4.3. Access Control and Monitoring 4](#_Toc176334519)

[5. Responsibilities 5](#_Toc176334520)

[6. Breaches of Policy 5](#_Toc176334521)

[7. Document Management 5](#_Toc176334522)

# Introduction

Machine learning models, particularly those deployed in Internet of Things (IoT) systems, can be vulnerable to membership inference attacks. These attacks aim to determine whether a specific data record was part of the training dataset used to develop the model. Successful attacks can compromise the privacy of individuals whose data was used for training, potentially leading to data breaches and regulatory non-compliance. This policy outlines the organisation's commitment to protecting against membership inference attacks and preserving the privacy of training data.

# Purpose

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

* Safeguard the privacy of individuals whose data is used for AI model training.
* Prevent unauthorised inference of membership in training datasets.
* Ensure compliance with data protection regulations and ethical considerations.
* Maintain the trustworthiness and confidentiality of AI model training processes.

# Scope

This policy applies to all AI and ML models developed, deployed, or utilised within the organisation's IoT environment, particularly those trained on sensitive or personal data.

# Policy Statement

## Model Protection and Privacy

* **Privacy-Preserving Techniques:** AI models shall be developed and trained using privacy-preserving techniques, such as differential privacy, to minimise the risk of membership inference attacks.
* **Model Output Control:** Access to model outputs and predictions shall be carefully controlled to prevent adversaries from using them to infer membership in the training dataset.

## Differential Privacy

* **Implementation:** Where applicable, differential privacy techniques shall be incorporated into the AI model training process to add controlled noise to model parameters or outputs, making it difficult to infer the presence of specific data points in the training set.
* **Privacy Budget:** A privacy budget shall be defined and managed to control the level of noise added and balance privacy protection with model utility.

## Access Control and Monitoring

* **Restricted Access:** Access to AI models, training data, and model outputs shall be limited to authorised personnel and systems only.
* **Strong Authentication:** Robust authentication mechanisms shall be implemented to verify the identity of users and systems accessing AI models or data.
* **Logging and Auditing:** All access to and interactions with AI models and data shall be logged and audited for traceability and accountability.
* **Anomaly Detection:** Anomaly detection mechanisms shall be employed to identify unusual or suspicious patterns of access or usage that may indicate potential membership inference attacks.

**4.4 Training Data Management**

* **Data Minimisation:** The collection and retention of training data shall be limited to only what is necessary for the specific AI model development purposes.
* **Data Anonymisation and Pseudonymisation:** Where feasible, data anonymisation or pseudonymisation techniques shall be applied to training data to reduce the risk of re-identification.

# Responsibilities

* **Information Security Officer:** Responsible for overseeing the implementation and enforcement of this policy.
* **Data Protection Officer:** Responsible for ensuring compliance with data protection regulations and advising on privacy-preserving techniques.
* **Data Scientists and AI Developers:** Responsible for developing and deploying AI models in accordance with this policy, incorporating privacy-preserving techniques and mitigating the risk of membership inference attacks.
* **IT Department:** Responsible for providing secure infrastructure and tools for AI model development, deployment, and monitoring.

# 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 membership inference attacks may result in privacy breaches, 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