TrustIoT Framework for Industry 4.0

"AI Explainability and Transparency"

|  |  |
| --- | --- |
| 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](#_Toc176334338)

[2. Purpose 4](#_Toc176334339)

[3. Scope 4](#_Toc176334340)

[4. Policy Statement 4](#_Toc176334341)

[4.1. Explainable AI (XAI) Adoption 4](#_Toc176334342)

[4.2. Transparency and Interpretability 4](#_Toc176334343)

[4.3. Documentation and Communication 4](#_Toc176334344)

[4.4. User Empowerment and Control 5](#_Toc176334345)

[5. Responsibilities 5](#_Toc176334346)

[6. Breaches of Policy 5](#_Toc176334347)

[7. Document Management 5](#_Toc176334348)

# Introduction

Artificial Intelligence (AI) models are increasingly integrated into IoT systems to enable advanced analytics, automation, and decision-making. However, the complexity and "black-box" nature of many AI algorithms can raise concerns about transparency, accountability, and potential bias. This policy outlines the organisation's commitment to promoting the use of explainable AI techniques to ensure transparency and understandability of AI model decisions, particularly in critical industrial applications.

# Purpose

The purpose of this policy is to establish guidelines and requirements for the development, deployment, and use of explainable AI within the organisation's IoT infrastructure. This policy aims to:

* Enhance transparency and understanding of AI model decisions, particularly in high-stakes or critical applications.
* Enable users and stakeholders to comprehend the reasoning behind AI-driven actions and outcomes.
* Facilitate the identification and mitigation of potential biases in AI models.
* Foster trust and acceptance of AI technologies within the organisation and among its stakeholders.

# Scope

This policy applies to all AI models developed, deployed, or utilised within the organisation's IoT environment, particularly those used in critical industrial applications where the impact of decisions can have significant consequences.

# Policy Statement

## Explainable AI (XAI) Adoption

* **Prioritisation:** The organisation shall prioritise the use of explainable AI techniques and models whenever possible, especially in critical applications where transparency and understandability are essential.
* **Model Selection:** When selecting AI models, preference shall be given to those that offer inherent explainability or can be readily interpreted using XAI techniques.
* **XAI Development:** Where black-box models are necessary, efforts shall be made to develop or integrate XAI methods to provide explanations for their decisions.

## Transparency and Interpretability

* **Clear Explanations:** AI models shall provide clear and understandable explanations for their decisions, highlighting the key factors and features that influenced the outcome.
* **Visualisation:** Where appropriate, visual representations or explanations shall be provided to aid in understanding complex model behaviour.
* **User-Friendly Interfaces:** Explanations shall be presented in a user-friendly manner, accessible to both technical and non-technical stakeholders.

## Documentation and Communication

* **Model Documentation:** Comprehensive documentation shall be maintained for each AI model, including details about its architecture, training data, features used, and explanation mechanisms.
* **Explanation Communication:** Explanations for AI model decisions shall be communicated to relevant stakeholders in a clear and concise manner, tailored to their level of technical understanding.

## User Empowerment and Control

* **User Interaction:** Where feasible, mechanisms shall be provided for users to interact with AI models and request explanations for specific decisions.
* **Feedback Loop:** A feedback loop shall be established to allow users to provide input on the quality and understandability of explanations, driving continuous improvement.

# Responsibilities

* **Information Security Officer:** Responsible for overseeing the implementation and enforcement of this policy.
* **Data Scientists and AI Developers:** Responsible for developing and deploying explainable AI models and providing clear explanations for their decisions.
* **IT Department:** Responsible for providing tools and infrastructure to support XAI development and deployment.
* **Model Owners:** Responsible for ensuring that their AI models comply with this policy and provide adequate transparency and explainability.

# Breaches of Policy

Non-compliance with this policy may result in disciplinary action, up to and including termination of employment or contractual relationships. Additionally, the use of opaque or unexplainable AI models in critical applications may lead to operational risks, regulatory non-compliance, and reputational damage.

# 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