TrustIoT Framework for Industry 4.0

"Human Oversight and AI Control"

|  |  |
| --- | --- |
| 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](#_Toc176334473)

[2. Purpose 4](#_Toc176334474)

[3. Scope 4](#_Toc176334475)

[4. Policy Statement 4](#_Toc176334476)

[4.1. Critical AI Systems Identification 4](#_Toc176334477)

[4.2. Human-in-the-Loop Processes 4](#_Toc176334478)

[4.3. Override Mechanisms 4](#_Toc176334479)

[4.4. Accountability and Auditability 5](#_Toc176334480)

[5. Responsibilities 5](#_Toc176334481)

[6. Breaches of Policy 5](#_Toc176334482)

[7. Document Management 5](#_Toc176334483)

# Introduction

Artificial Intelligence (AI) systems are increasingly being integrated into various aspects of the organisation's operations, including IoT devices and applications. While AI offers numerous benefits in terms of efficiency, automation, and decision-making, it is crucial to maintain human oversight and control, especially in critical systems where the consequences of errors or unintended actions can be significant. This policy outlines the principles and procedures for ensuring human involvement and control in the operation of AI systems within the organisation.

# Purpose

The purpose of this policy is to establish a framework for maintaining human oversight and control over critical AI systems within the organisation's IoT infrastructure. This policy aims to:

* Ensure that AI systems operate within defined ethical and legal boundaries.
* Prevent unintended consequences and mitigate risks associated with AI decision-making.
* Enable human intervention and override in critical situations.
* Promote transparency and accountability in the use of AI.

# Scope

This policy applies to all AI models, algorithms, and applications deployed or utilised within the organisation's IoT environment, particularly those involved in critical decision-making processes or operations that can impact safety, security, or significant business outcomes.

# Policy Statement

## Critical AI Systems Identification

* **Risk Assessment:** A risk assessment shall be conducted to identify critical AI systems based on their potential impact on the organisation, individuals, or society.
* **Classification:** Critical AI systems shall be classified and documented, with clear criteria for determining their criticality level.

## Human-in-the-Loop Processes

* **Human Review and Approval:** Critical AI-driven decisions or actions shall be subject to human review and approval before execution.
* **Exception Handling:** Procedures shall be established for handling exceptions and escalating decisions to higher levels of human oversight when necessary.
* **Training and Competence:** Personnel involved in human-in-the-loop processes shall receive adequate training to understand the AI system's capabilities, limitations, and potential risks.

## Override Mechanisms

* **Manual Override:** Mechanisms shall be implemented to allow authorised personnel to manually override or intervene in AI systems in case of errors, malfunctions, or ethical concerns.
* **Emergency Stop:** In critical situations, an emergency stop mechanism shall be available to immediately halt the operation of AI systems.
* **Clear Documentation:** Override procedures and authorisation levels shall be clearly documented and communicated to relevant personnel.

## Accountability and Auditability

* **Clear Responsibility:** Clear lines of responsibility and accountability shall be established for AI-driven decisions and actions.
* **Audit Trails:** Audit trails shall be maintained to track and record AI system activities, including decisions made, actions taken, and any human interventions.
* **Regular Reviews:** AI systems and their outputs shall be periodically reviewed by designated personnel to ensure their continued alignment with ethical principles and organisational objectives.

# Responsibilities

* **AI Ethics Committee:** Responsible for overseeing the ethical development and use of AI within the organisation, including the implementation of human oversight and control mechanisms.
* **Information Security Officer:** Responsible for ensuring that AI systems comply with this policy and relevant security standards.
* **Data Scientists and AI Developers:** Responsible for designing and implementing AI systems with appropriate human oversight and control mechanisms.
* **Management:** Responsible for fostering a culture of responsible AI use and providing necessary resources and support.

# 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 lack of human oversight and control in critical AI systems may lead to operational risks, safety hazards, 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