TrustIoT Framework for Industry 4.0

"AI Fairness and Bias Mitigation"

|  |  |
| --- | --- |
| 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](#_Toc176334444)

[2. Purpose 4](#_Toc176334445)

[3. Scope 4](#_Toc176334446)

[4. Policy Statement 4](#_Toc176334447)

[4.1. Fairness and Non-discrimination 4](#_Toc176334448)

[4.2. Transparency and Explainability 4](#_Toc176334449)

[4.3. Data Governance and Bias Mitigation 4](#_Toc176334450)

[4.4. Monitoring and Review 5](#_Toc176334451)

[5. Responsibilities 5](#_Toc176334452)

[6. Breaches of Policy 5](#_Toc176334453)

[7. Document Management 5](#_Toc176334454)

# Introduction

Artificial Intelligence (AI) models are increasingly integrated into IoT systems to enable advanced analytics, automation, and decision-making capabilities. However, AI models can inadvertently perpetuate or amplify biases present in training data, leading to discriminatory or unfair outcomes. This policy outlines the organisation's commitment to ensuring fairness and mitigating bias in the development, deployment, and use of AI within the IoT ecosystem.

# Purpose

The purpose of this policy is to establish a framework for promoting fairness and mitigating bias in AI algorithms and decision-making processes within the organisation's IoT infrastructure. This policy aims to:

* Ensure that AI-driven decisions are fair, unbiased, and do not discriminate against individuals or groups.
* Promote transparency and explainability in AI models to enable understanding and scrutiny of decision-making processes.
* Establish guidelines for data governance and model development to minimise the risk of bias.
* Foster trust and accountability in the use of AI within the IoT ecosystem.

# Scope

This policy applies to all AI models, algorithms, and applications developed, deployed, or utilised within the organisation's IoT environment.

# Policy Statement

## Fairness and Non-discrimination

* **Prohibited Bias:** AI systems shall not be designed or used in a manner that results in discriminatory or unfair outcomes based on protected characteristics such as race, gender, age, religion, or disability.
* **Fairness Testing:** AI models shall be regularly tested and evaluated to identify and address any potential biases or discriminatory impacts.
* **Impact Assessments:** Prior to deployment, AI systems shall undergo impact assessments to evaluate their potential societal and ethical implications, including the risk of bias or discrimination.

## Transparency and Explainability

* **Explainable AI (XAI):** The organisation shall strive to use explainable AI techniques to provide insights into the reasoning behind AI-driven decisions.
* **Documentation:** AI models and algorithms shall be documented, including their purpose, data sources, features used, and decision-making processes.
* **Communication:** Clear and understandable explanations of AI-driven outcomes shall be provided to affected individuals and stakeholders upon request.

## Data Governance and Bias Mitigation

* **Data Collection and Preparation:** Training data shall be collected and prepared in a manner that minimises the risk of bias, ensuring diversity and representativeness.
* **Data Quality and Integrity:** Data used for AI training shall be accurate, complete, and free from errors or inconsistencies that may introduce bias.
* **Bias Detection and Mitigation:** Techniques for detecting and mitigating bias in training data and AI models shall be employed, such as:
  + Regular audits and reviews of training data and model outputs
  + Use of diverse and inclusive development teams
  + Application of fairness metrics and algorithms

## Monitoring and Review

* **Ongoing Monitoring:** Deployed AI models shall be continuously monitored for signs of bias or discriminatory outcomes.
* **Regular Reviews:** Periodic reviews shall be conducted to assess the fairness and ethical implications of AI systems.
* **Feedback Mechanisms:** Mechanisms shall be in place to receive and address feedback from users and stakeholders regarding potential biases or concerns about AI-driven decisions.

# Responsibilities

* **AI Ethics Committee:** An AI Ethics Committee shall be established to oversee the ethical development and use of AI within the organisation, including the promotion of fairness and bias mitigation.
* **Data Scientists and AI Developers:** Responsible for developing and deploying AI models in accordance with this policy, ensuring fairness and minimizing bias.
* **IT Department:** Responsible for providing tools and infrastructure to support bias detection and mitigation efforts.
* **Management:** Responsible for fostering a culture of ethical 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.

# 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