AI ITA Blueprint Guidelines

MDIA
Malta Digital Innovation Authority

AI-ITA Blueprint Guidelines

Digital Innovation is, by definition, a rapidly evolving sector. These guidelines are expected to be updated to keep abreast with technology, regulatory and operational developments.

Document Version: 03 October 2019

Contents

1. Definitions 4
2. Blueprint of the AI-ITA 5
2.1. Scope 5
2.2. High-level description 5
2.3. Technical specifications 6
Data Collection & Storage Process 7
Data Processing & Analysis 7
AI Engine 8
Application & Implementation Process 8
ITA Harness 8
Other Processes 8
2.4. Forensic Node 9
2.5. Information security requirements 9
2.6. TA powers of intervention 9
2.7. Alignment with 'Ethical and Trustworthy AI' framework 10
Terms of Service 10
2.8. Other information 11
3. AI-ITA Nomenclature 12

1. Definitions

"Applicant", within the context of this document, refers to an individual and/or legal organisation applying for Certification of an Innovative Technology Arrangement (ITA) with the Authority.

"Authority" refers to the Malta Digital Innovation Authority ('MDIA'), as defined by the Malta Digital Innovation Authority Act, 2018 ('MDIA Act).

"Blueprint" refers to a document that includes a description of the qualities, attributes, features, behaviours or aspects of an ITA as defined in the 'ITA Blueprint Guidelines'.

"Innovative Technology Arrangement", also referred to as 'ITA' within this document, as defined within the First Schedule of the Innovative Technology Arrangements and Services Act, 2018.

"AI-ITA" refers to Innovative Technology Arrangements that exhibit features or qualities of Artificial Intelligence as recognised by the Authority and described in the 'AI Innovative Technology Arrangements Guidelines'.

"ITAS Act" refers to the Innovative Technology Arrangements and Services Act, 2018.

"Systems Auditor" ('SA') as defined in the Innovative Technology Arrangements and Services Act, 2018, and in line with further guidance issued by the Authority within the 'Systems Auditor Guidelines'.

"Technical Administrator" ('TA') as defined in the *Innovative Technology Arrangements and Services Act, 2018*, and in line with further guidance issued by the Authority under Chapter 3 of the Guidance Notes.

"Forensic Node" as defined by the Authority within the 'Forensic Node Guidelines".

"Ethical and Trustworthy AI Framework" refers to the *Malta Towards Trustworthy AI: Malta Ethical AI Framework* guidelines, published by the *Malta.AI* Taskforce on https://malta.ai/

2. Blueprint of the AI-ITA

2.1. Scope

The Blueprint is a document which highlights all the critical and important features which an AI-ITA should include in the information submitted to the Authority during the application for the ITA certification. This document will also be used by the Systems Auditor to understand and verify the implementation of the control objectives as described in 'Chapter 1 – Systems Auditor Control Objectives.

The Blueprint is split into the following requirements, as documented in 'Chapter 2 – AI Innovative Technology Arrangements Guidelines':

- **Purposes**: The reasons for which the AI-ITA is being, or was, created
- **Qualities**: The specific characteristics that the AI-ITA offers to its users
- **Aspects**: The specific elements or boundaries of the AI-ITA that are subject to the certification
- **Features**: The distinctive functional capabilities of the AI-ITA
- **Attributes**: The inherent capabilities of the AI-ITA
- **Failure Modes**: The manner how the AI-ITA responds to unexpected processes and inputs
- **Verification**: The substantiation of the results produced by the AI-ITA
- **Limitations**: The technical and/or operational restrictions of the AI-ITA

The next sub-sections explain the level of detail that would typically be expected in the Blueprint documentation of the AI-ITA. Note that these sections reference information that may be further explained within the 'AI-ITA Nomenclature' guidelines.

2.2. High-level description

The Applicant needs to provide a high-level description regarding the scope and purpose that the AI-ITA fulfils. As a minimum, such a description should include:

- What the AI-ITA's objectives are.
- What user demands are addressed by the AI-ITA.
- How the objectives will be met by the AI-ITA.
- What the risks of the AI-ITA are and how they are mitigated.
- Known limitations by the AI-ITA in addressing the user demands.
- What benefits the user will derive from when making use of such an AI-ITA.
- What expected benefits will the Applicant derive from the AI-ITA.
- What type of audience is the AI-ITA designed to address.
- How was quality assurance practised during the development and/or implementation of the AI-ITA.
- What underlying technologies is the AI-ITA making use of (including architectural diagrams).

- How changes to the AI-ITA are addressed.
- How is the AI-ITA aligned to the 'Ethical and Trustworthy AI Framework'.

2.3. Technical specifications

The Applicant needs to provide a detailed description of the AI-ITA that is being submitted for registration with the Authority. The information provided needs to be in-line with the process-map described in the related *AI-ITA Nomenclature* document. The Authority is providing further guidance below on what information should be provided for each of the processes within the process-map whenever applicable to be answered by the applicant of the AI-ITA. Due to the wide spectrum which AI-ITAs may fall within, the processes or questions should not be considered exhaustive or applicable to all systems, and the Applicant must add any other information deemed necessary for the Authority and Systems Auditor to gain a sufficient understanding of the system. The Authority and Systems Auditor reserve the right to ask for further clarification on the blueprint wherever they deem this to be required.

For the entire scope of the AI-ITA, explain:

- What the architecture of the AI-ITA is: The Applicant is requested to provide a high-level process-flow diagram of the AI-ITA showing the components that make up the different processes and how they interact with one another.
- What underlying technologies the AI-ITA is specifically making use of: The Applicant is requested to provide a high-level diagram of the AI-ITA showing the underlying technologies and how they may interact with one another.
- How the AI-ITA handles data: The AI-ITA should provide a high-level data flow diagram showing how data flows between the various components of the AI-ITA, including any techniques being used, such as but not limited to anonymisation techniques and application of bias to underlying datasets.
- Describe where humans are involved in the ITA and what capabilities of overriding the system they have, including notifications of abnormal activity, and kill-switch functionality.
- Describe how the Technical Administrator's power-of-intervention mechanism is implemented.
- Describe how the AI-ITA takes into consideration and implements the objectives outlined in the 'Ethical and Trustworthy AI Framework' guidelines.
- Specify any security measures being applied.

For each of the processes described below and in additional to the process-specific requirements, explain:

- How fidelity to the claimed functionality is being addressed.
- How privacy, integrity and confidentiality is being addressed.
- Describe any access control mechanisms in place.
- Describe what programming language was used and/or technologies, such as platforms and libraries, which the AI-ITA process is dependent on.

- Describe any coding practices and standards used.
- Describe what external interface the AI-ITA process offers.
- Describe any external dependency the AI-ITA process requires.
- Describe the architecture of the AI-ITA process: The Applicant is requested to provide a detailed process-flow diagram showing the components that make up the process and how the components interact with one another.
- Describe the underlying technologies the AI-ITA process is specifically making use of.
- Describe how the AI-ITA process handles data: The Application is requested to provide a high-level data flow diagram showing how data flows between the various components, including any techniques being used, such as but not limited to anonymisation techniques and application of bias to underlying datasets.
- Describe where humans are or may be involved in the process.
- Specify any security measures being applied at the AI-ITA process level.
- Describe the levels of autonomy that the AI-ITA has and mitigating factors.
- Describe how the process takes into consideration and implements the considerations outlined in the 'Ethical and Trustworthy AI Framework'.
- Describe any logging that takes place at the process level (this should be further detailed in 2.4 Forensic Node).
- Describe compliance with laws and regulations through in-built functionalities within the AI-ITA process.
- Describe safety mechanisms in place to detect and handle abnormal results through the use of the ITA Harness.

Data Collection & Storage Process

- Identify the data that is collected.
- Why the data is needed.
- Where it is collected from.
- Where and how it is stored.
- How it is processed, including how any data may be anonymized.
- How long will it be retained for.
- Describe any feedback-loop mechanisms.
- Categorize data by sensitivity (e.g. differentiate between private data, publicly sourced open data).
- Governance controls around different types of data (e.g. who can access personal data).
- Define how the AI-ITA achieves resilience to attacks (e.g. Data pollution attacks).

Data Processing & Analysis

- How the data is processed, transformed and/or combined and at which stage.
- How the AI-ITA is utilising the data it has access to.
- Describe the metrics and/or outcomes that are being extracted from the datasets.
- Describe the explainability mechanisms in place.

- Describe any safety mechanisms in place to detect and handle abnormal results.

AI Engine

- Describe the initial state of the AI-ITA.
- How and what data was used to set the initial state of the AI-ITA and how this is maintained (e.g. via Feedback-loop mechanisms).
- Identify what AI algorithms are in use by the AI-ITA and provide a detailed description of them including their behaviours, process and data flows, and known limitations.
- Describe safety mechanisms in place to detect and handle abnormal results.
- The inputs (actions) to be provided by users or other dependencies and the boundaries within which they are expected in, as well as how any exceptions will be managed (through the involvement of the AI-ITA Harness).
- The outputs (reactions) that may be returned by the system and the boundaries within which they are expected in, as well as how any exceptions will be managed (through the involvement of the ITA Harness).
- Describe any other components that may be in use and define their purpose, behaviour and safety mechanisms.

Application & Implementation Process

- How it is implemented (e.g. System Architecture, programming languages and infrastructure relied upon).
- How the AI-ITA is accessed by end users (e.g. website, mobile app).
- How it fits within the system described by the AI-ITA at large.

ITA Harness

- What the boundaries of the AI-ITA are.
- How the ITA Harness detects and handles abnormal results at process level. Note that this is to be detailed at component level within each respective AI-ITA process.
- How the ITA Harness interacts with the Forensic Node and what information it reports for auditability and traceability purposes.

The Authority understands that there may be instances where a harness is limited or not applicable to an AI-ITA as it may not be deemed required, feasible or desirable. In any of these cases the Blueprint must include a detailed justification to ensure that any operating risks of the AI-ITA are well defined, monitored and contained separately.

Other Processes

- Define any other processes in use by the AI-ITA.
- Why the process is required.
- What the process does, its components, and its behaviour.
- Who and what other processes interact with the process and to which extent.
- Where the process resides in the process-map and architecture of the system.

2.4. Forensic Node

For details on the Forensic Node, please refer to separate MDIA guidelines dedicated to this topic. The Forensic Node guidelines are available on www.mdia.gov.mt in Guidelines section of the website.

2.5. Information security requirements

A risk assessment should be undertaken and documented on the AI-ITA. Such a risk assessment may include the following information:

- Risk reference
- Risk description
- Risk type, examples of which include:
    - Bias in the data set
    - Behaviours outside expected operating range and their impacts (e.g. Financial)
    - Infrastructural issues (e.g. Network failures)
- Scenario
- Threat actor
- Affected assets
- Risk owner
- Original risk rating
- Implemented mitigation controls
- Effectiveness of control
- Likelihood of happening
- Inherent risk
- Residual risk
- Financial impact value of risk

The Authority requires all Applicants to provide information based on the following:

- Describe how a security risk assessment plan was undertaken and is maintained.
- How information security is being addressed
- Provide information on what information security algorithms are being deployed and implemented.

2.6. TA powers of intervention

In line with Article 8(4)(d)(iii) of the ITAS Act, the AI-ITA needs to document within the Blueprint the powers and the technology features proposed to enable the Technical Administrator, or the Authority (in the case of unjustifiable failure by the Technology Administrator) to intervene in the event of:

- A material cause of loss to any user
- A material breach of law

Such intervention would need to be conducted in a transparent and effective manner. It must address the cause of loss or the breach of law such that it does not occur or re-occur.

The Authority requires the Applicant to:

- Provide information on the possibility for someone to intervene during the AI-ITA's activity. If there is no power to intervene, the Applicant needs to explain in detail the reason why intervention on the AI-ITA's activity is not technically feasible.
- Provide the documented procedure required to intervene upon the AI-ITA.
- Provide details of the authorised persons or entities who can intervene upon an AI-ITA.

2.7. Alignment with 'Ethical and Trustworthy AI' framework

The Authority requires the applicant to define its position in relation to the 'Ethical and Trustworthy AI Framework' both at the organisational level, as well as the specific AI-ITA aspect. Wherever alignment is not possible due to conflicts with the scope of the AI-ITA or technical reasons, adequate justification must be given. Alignment with additional ethical guidelines and standards is encouraged and should also be specified. These include IEEE's *Ethically Aligned Design*[1], OECD Principles on AI[2], and the EC's Ethics guidelines for trustworthy AI[3].

Terms of Service

In addition to defining the adherence to these guidelines in the blueprint, the below aspects from the above framework (at a minimum) need to be disclosed to direct users as well as anyone who may be affected by the AI-ITA. The disclosure should avoid technical language, can be within the Terms of Service, or in a dedicated online section on the organisations or AI-ITAs website, and must be in the English language. The Authority reserves the right to request that the AI-ITAs position in relation to a particular guideline objective be made public.

The terms of service should cover:

- How the solution respects all four pillars of the 'Ethical and Trustworthy AI Framework': Human Autonomy, Prevention of Harm, Fairness and Explicability.
- Any limitations, conditions or justifications in relation to bias inherent in the system.
- How explainable the results expressed by the AI-ITA are.
- Potential impact on workforce of an organisation and considerations to be taken to mitigate this should the AI-ITA get deployed within an organisation.
- Expected accuracy of the end-results, and whether there is any possible decision variability that may occur under the same conditions.
- Specify mechanisms by which the AI-ITA can deal with unexpected situations (Failure Modes).

---

[1]IEEE – Ethically Aligned Design, First Edition, https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

[2] OECD – OECD Principles on AI, https://www.oecd.org/going-digital/ai/principles/

[3]European Comission – Ethics guidelines for trustworthy AI, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai,

- What personal/sensitive data is collected and how it is utilised.

2.8. Other information

The Authority requires other information that is not of a technical nature.

- Provide information on the governance structure of the AI-ITA.
- Describe any limitations or restrictions on the operational boundaries of the system.
- Describe any specific features that distinguish the AI-ITA from other AI-ITA projects of a similar nature.
- Is there any expected end-of-life date or event for the AI-ITA?
- Are there any risks that may cause the AI-ITA to reach its end-of-life prematurely?

3. AI-ITA Nomenclature

The Authority is issuing, as a separate document, the 'AI-ITA Nomenclature' guidelines that are meant to complement this document by providing the use-case of a generic AI system, and is intended to guide the Applicant in understanding further the requirements outlined in this document, as well as when writing the Blueprint.