MALTA

TOWARDS TRUSTWORTHY AI

MALTA'S ETHICAL AI FRAMEWORK | OCTOBER 2019

MALTA.AI
Towards a National AI Strategy

PARLIAMENTARY SECRETARIAT FOR FINANCIAL SERVICES,
DIGITAL ECONOMY AND INNOVATION
OFFICE OF THE PRIME MINISTER

CONTENTS

Malta.AI Taskforce 4

Introduction 5

CHAPTER 1 — Malta's Vision: Ethical & Trustworthy AI 6

CHAPTER 2 — Ethical AI Framework 11

CHAPTER 3 — Governance and Control Practices 17

CHAPTER 4 — AI Certification 33

MALTA.AI TASKFORCE

Wayne Grixti
Chair

Dr. Angelo Dalli
Member

Ing. Emanuel Darmanin
Member

Prof. Alexiei Dingli
Member

Dr. Abdalla Kablan
Member

Dr. Jackie Mallia
Member

Francois Piccione
Member

Dylan Seychell
Member

Ing. Antoine Sciberras
Member

Ing. Godfrey Vella
Member

Wilbert Tabone
Secretary

INTRODUCTION

Malta aspires to become the **Ultimate AI launchpad** – a place in which local and foreign companies, and entrepreneurs, can develop, prototype, test and scale AI. The ambition is to create the conditions for AI to springboard from Malta to the world. A necessary condition to achieve this ambition is for Malta to create a regulatory and innovation ecosystem that develops trustworthy AI.

A strong ethical AI framework as a supplement to the current legal and regulatory system is a core component of Malta's AI strategy to ensure that AI development is **ethically aligned**, **transparent** and **socially responsible**. The following chapters of the document outline the proposed ethical AI guiding principles and policy considerations that will form the basis of Malta's Ethical AI Framework.

A first draft of this document was published on 09 August 2019 for public consultation. The Malta.AI Taskforce and the Parliamentary Secretary for Financial Services, Digital Economy and Innovation within the Office of the Prime Minister would like to express deep gratitude to the members of the public, industry and academia who contributed feedback, which was considered in the development of this revised version.

MALTA'S VISION: ETHICAL & TRUSTWORTHY AI

CHAPTER 1

AI RAISES PROFOUND QUESTIONS ACROSS ETHICAL, LEGAL AND REGULATORY DOMAINS, FROM PROTECTING NATIONAL SECURITY AND CITIZENS' RIGHTS TO ADVANCING COMMERCIAL INTERESTS AND INTERNATIONAL STANDING.

As AI use cases proliferate it is important that these issues are addressed at the outset to mitigate risks and unintended outcomes. Without doing so, AI's development may be hindered by a lack of trust and low adoption by stakeholders.

The Government is developing the Malta Ethical AI Framework as it understands that for Malta to become the Ultimate AI Launchpad, the country needs to create an ecosystem that promotes the design and operation of trustworthy AI.

In developing the Malta Ethical AI Framework, the Government had the following four (4) objectives:

Build on a human-centric approach;

Respect for all applicable laws and regulations, human rights and democratic values;

Maximise the benefits of AI systems while preventing and minimising their risks;

Align with emerging international standards and norms around AI ethics.

In developing the Malta Ethical AI Framework, the Government recognises that developing trustworthy AI is a complex task that will require the Framework to intersect with various policy initiatives including existing laws and regulations, investments in tools and continuous monitoring mechanisms, skills and capabilities, innovation ecosystem and regulatory mechanisms. A National Technology Ethics Committee will be set up under the Malta Digital Innovation Authority (MDIA) to oversee the Ethical AI Framework and its intersection across these areas.

The purpose of the Malta Ethical AI Framework is to establish a set of guiding principles and trustworthy AI governance and control practices that can serve as the foundation for the design and implementation of these broader constructs, as a supplement to the legal and regulatory system.

Malta intends to continue its strong tradition of being an innovator and first mover in the space of innovative technologies. It is therefore in the process of creating the world's first national AI certification programme, which is being launched in October 2019. The certification process will largely be based on Malta's Ethical AI Framework, and provide applicants with valuable recognition in the marketplace that their AI systems have been developed in an ethically aligned, transparent and socially responsible manner. The ambition is to create the right frameworks to help trustworthy AI springboard from Malta to the world, in line with Malta's vision to become the **Ultimate AI Launchpad**.

Build on a human-centric approach

AI aims to replicate cognitive tasks, at scale, such as natural language processing, perception and emotional cognition and can be used in a wide-range of use cases that can benefit society. For example, 'AI for Good' has been coined to refer to AI applications that improve human well-being and our natural environment, such as enhancing healthcare, addressing climate change and eradicating poverty.

AI can also potentially be very disruptive and become a destabilising force, reshaping the nature of work and employment, and creating scenarios that thwart existing data protections. Ultimately, the outcomes from AI will be dependent upon the objectives and goals it is given and the safeguards it must operate within.

To ensure that AI is used to increase individual and societal well-being and be used for good, "AI systems need to be human-centric, resting on a commitment to their use in the service of humanity and the common good, with the goal of improving human welfare and freedom."[1] This requires that goals and objectives set for the AI system being developed give consideration to creating a positive impact on human well-being, rights and freedoms, as well as ensuring that the AI system is designed with the intent to minimise harm.

In developing a human-centric approach to AI, the Government has identified the following conditions which are aligned with the European industrial policy on artificial intelligence and robotics: [2]

1. The user's needs, wishes and experiences need to be the starting point of the design for AI;
2. AI should be designed and deployed in a manner that preserves the dignity, autonomy and self-determination of an impacted individual;
3. The development and deployment of AI must always be based on the 'man operates machine' principle of responsibility;
4. An inclusive approach to the development of AI will facilitate greater benefits for society, enhance the quality of AI systems, improve user experience and more effectively addressing the challenges presented by AI; and
5. AI should be designed and deployed in a manner that is equitable and mitigates bias to the greatest extent possible (see page 13 for further detail on on this point).

We anticipate that adherence to the above human-centric approach will become a minimum expectation in the development of AI, as Malta supports the view that through trust, designers and operators of AI will obtain a "social license to operate" AI in their given field.

Respect for all applicable laws and regulations, human rights and democratic values

A cornerstone of the Malta Ethical AI Framework is to define the "ethical code" for which AI will be governed. A challenge in developing an ethical code, is that values and social norms can differ from one individual to another. In alignment with the European Union (EU), the Government has adopted an approach to AI ethics

[1] European Commission Independent High-Level Expert Group on *Artificial Intelligence*. (2019). *Ethics Guidelines for Trustworthy AI.*

[2] European Parliament (2019). *P8_TA-PROV (2019)0081*. A comprehensive *European industrial policy on artificial intelligence and robotics*. European Parliament resolution of 12 February 2019 on a comprehensive European industrial policy on artificial intelligence and robotics (2018/2088(INI))

that is based on the fundamental rights enshrined in the EU Treaties and the EU Charter. This approach is consistent with the ethical framework adopted for previous technologies, including Distributed Ledger Technology (DLT), and provides a common foundation in uniting the respect for human rights with a human-centric approach.

In drawing from the EU Treaties, the below listed fundamental rights are relevant to the design and deployment of AI:

- **Respect for human dignity** — AI systems should be designed and operated in a manner that respects, serves and protects humans' physical and mental integrity, personal and cultural sense of identity, and satisfaction of their essential needs.
- **Freedom of the individual** — AI systems should be designed and operated in a manner that respects a human being's freedom to make life decisions, and reduces the risk of coercion. They should also be protected from unjustified surveillance, deception and unfair manipulation.
- **Respect for democracy, justice and the rule of law** — AI systems should serve to maintain and foster democratic processes and respect the plurality of values and life choices of individuals. AI systems should also not undermine the foundation of the justice system, to ensure due process before the law.
- **Equality, non-discrimination and solidarity** — AI operations can not generate unfairly biased outcomes, and the benefits and opportunities of AI should be equitably available to all.
- **Citizen's rights** — AI design and operation should safeguard citizens' rights. [3]

While Malta, as an EU Member State, is legally obliged to respect and enforce the fundamental rights prescribed in the EU Treaties and EU Charter, it is important to note that these fundamental rights may not be sufficient to provide an ethical framework for every case in which AI may operate. In this case, its important that the governance and controls of an AI system is assessed in relation to all relevant obligations including rights, policies, laws, regulations, contracts, code of conduct, organisational commitments and stakeholder expectations. As Malta considers whether new AI-specific regulation is required, it will first conduct a robust assessment to determine the extent to which existing laws and regulations apply.

Maximise the benefits of AI systems while preventing and minimising their risks

In order for Malta to develop a "top 10" national AI programme, it will require an AI ecosystem that promotes an acceleration in the achievement of the benefits of AI, while minimising its risks. This will require that Malta maintains a fine balance between opposing factors including:

- Creating an AI ecosystem that promotes innovation and risk mitigation;
- Maintaining a dual role as a disruptor and protector; and
- Developing a regulatory framework that balances prescribed rules with agility.

[3] The list of fundamental rights associated to AI, and the accompanying description is in alignment with those provided in the *Ethics Guidelines for Trustworthy AI*, European Commission Independent High-Level Expert Group on Artificial Intelligence.

The development of the Malta Ethical AI Framework alongside Malta's National AI Strategy will allow Malta to navigate these dualities from the outset and pilot AI projects within the ethical principles and leading control practices.

The risks posed by AI can differ from one use case to another, however they can be generally summarised in the following categories:

Bias — unfair and discriminatory outcomes;

Transparency — lack of transparency and/or informed consent in the collection, storage or use of data for profiling and/or automated decision making;

Performance — the AI system does not perform with the desired level of precision and consistency in achieving its desired objectives;

Explainability — AI's training methods and decision criteria may not be understood and may not be readily available for challenge and validation by a human operator; and

Resilience — AI system is susceptible to corruption or adversarial attack.

The risks created by AI are both general and case specific, and as a result will require the adoption of a robust risk management practice and continuous monitoring techniques to identify and mitigate AI risks. In Chapter 3 additional information is provided on the core components of a robust governance and control system for AI.

Aligned with emerging international standards and norms around AI ethics

In developing the Malta Ethical AI Framework, the Government has strived to formulate AI guidance that is consistent with emerging international standards and guidance around AI ethics, including those established in the European Union, whilst also given due regard to those set out by the OECD.

In developing the framework, particular attention was given to ensuring that that Ethical AI Framework was aligned with the following publications:

Ethics Guidelines for Trustworthy AI published on 8 April 2019 by the High-Level Expert Group on Artificial Intelligence set up by the European Commission.

Recommendations of the Council on Artificial Intelligence adopted on 21 May 2019 by the OECD countries and a number of non-member adherents.

Although the above two publications were used as the foundation for developing Malta's Ethical AI Framework, the Government considered the work of various other international organisations such as the Asilomar AI Principles, the Institute of Electrical and Electronics Engineers' (IEEE) Ethically Aligned Design, the Montréal Declaration, and the ethical AI frameworks released by other governments and technology companies.

The Malta Ethical AI Framework will be updated as new guidance materials are released and as changes are identified through AI project pilots.

ETHICAL AI FRAMEWORK

CHAPTER 2

AI HAS THE POTENTIAL TO POSITIVELY IMPACT THE WELL-BEING OF INDIVIDUALS, COMMUNITIES AND BROADER SOCIETY IN MANY WAYS, BUT FOR THIS TO HAPPEN IT MUST BE DESIGNED, TRAINED AND OPERATED IN A MANNER THAT CAN BE TRUSTED. WITHOUT TRUST, THE ADOPTION AND USE OF AI WILL BE STALLED AND THE MANY BENEFITS IT CAN PROVIDE WILL GO UNREALISED.

For an AI system to be trusted, ethics, governance and strong control practices must be central to its design and deployment. Organisations and governments which effectively incorporate ethical considerations into their AI projects, supported by a robust risk management system and monitoring mechanisms can expect to gain a competitive advantage over their peers.

In developing the Malta Ethical AI Framework, the Government's ambition is to create a practical and workable framework that can serve as a guide and enabler for AI practitioners to create trustworthy AI in Malta and beyond. The intention is for the Malta Ethical AI Framework to support AI practitioners in identifying and managing the potential risks of AI, while also serving to identify opportunities to encode into AI a higher ethical standard.

In developing the Malta Ethical AI Framework, the Government recognises that the field of AI is continually evolving and that new considerations will evolve over time.

The Government will continue to develop the framework as the experiences of working with AI are expanded and the field of AI and ethics matures.

Ethical AI Principles

In developing the Malta Ethical AI Framework, the Government has established four (4) Ethical AI Principles for establishing trustworthy AI which are in alignment with the EU Ethics Guidelines for Trustworthy AI.

- **Human autonomy** — humans interacting with AI systems must be able to keep full and effective **self-determination** over themselves;
- **Prevent harm** — AI systems must **not cause harm** at any stage of their lifecycle to humans, the natural environment or other living beings;
- **Fairness** — the development, deployment, use and operation of AI systems must be **fair**; and
- **Explicability** — end-users and other members of the public should be able to **understand** and **challenge** the operation of AI systems, as required for the particular use case.

The achievement of the above objectives is already encoded, in part, in existing legal and regulatory requirements and therefore they should be considered in relation to mandatory compliance required as a function of laws and regulations, as well as enhanced ethical expectations by stakeholders.

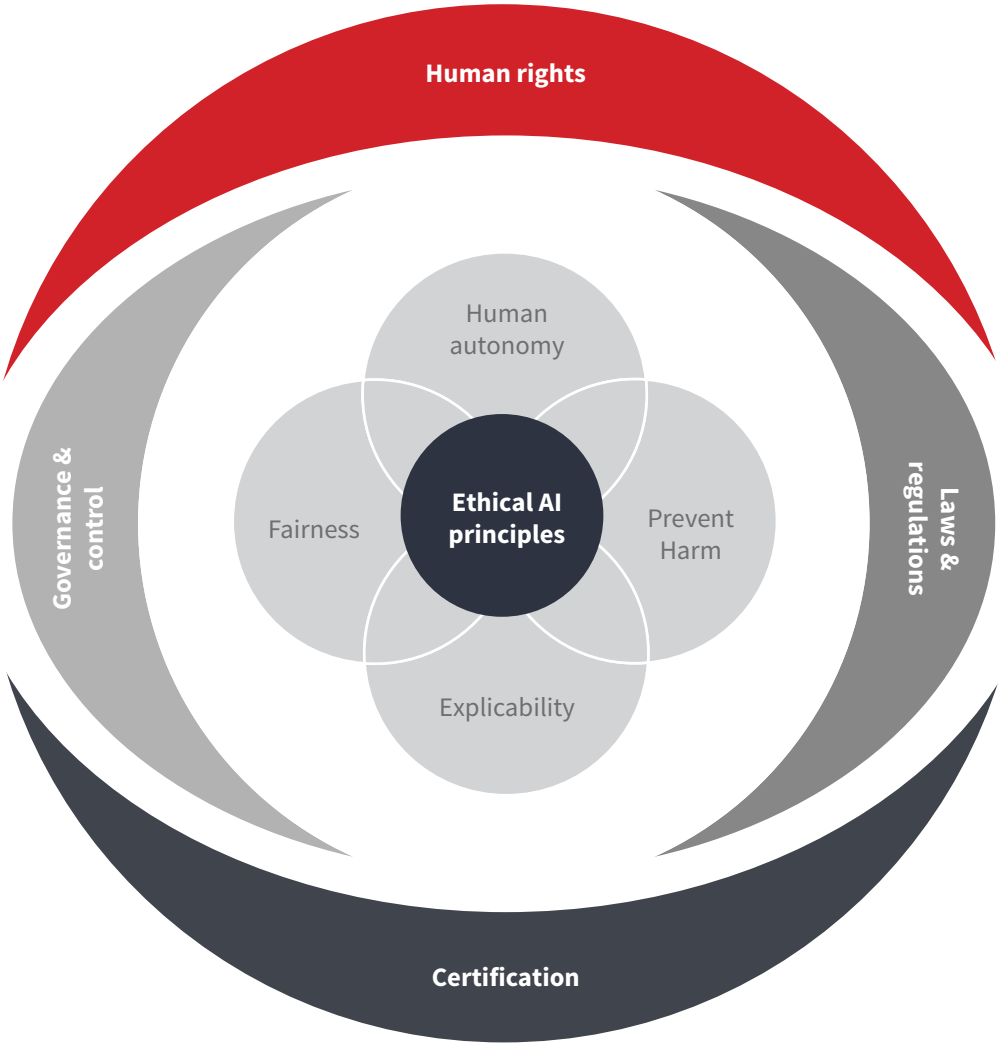

Figure 1: Ethical AI Framework

The intersection of the Ethical AI Principles (inner circles) and their supporting legal and regulatory constructs (outer circles) to achieve lawful and ethical AI are depicted in the above diagram.

Achieving the Ethical AI Principles will require AI practitioners to embed them into their system requirements from the outset, and to continually monitor them throughout the AI lifecycle. It will also require a thoughtful and in-depth evaluation of the applicability of each principle and how each is manifested in a particular AI use case and context.

For example, the elimination of bias with an objective of achieving neutral impartiality may not necessarily be fair due to the removal of key characteristics important to a prediction or decision. In fact, the elimination of bias may decrease the overall information value of an AI outcome.

For each of the AI principles, we have listed on the next pages a number of conditions that should be met when designing and operating an AI system in accordance with the Ethical AI Principles.

Human autonomy

Objective: Humans interacting with AI systems must be able to keep full and effective self-determination over themselves.

Design and operating considerations:

- AI systems should not **unjustifiably subordinate**, coerce, deceive, manipulate condition or nudge humans.
- AI systems should be designed to **augment, complement and empower** human cognitive, social and cultural skills.
- The allocation of functions between human and AI systems should follow **human-centric design** principles and leave meaningful opportunity for human choice.
- There should be meaningful and appropriate **human oversight** over work processes performed by AI.

Prevention of harm

Objective: AI systems must not cause harm at any stage of their lifecycle to humans, the natural environment or other living beings.

Design and operating considerations:

- AI systems and the environments they operate in must be **safe and secure**.
- AI systems should be **technically robust** and not open to malicious use.
- Due care should be taken to prevent unintended harms to **particularly vulnerable groups** and people subject to power imbalances.
- The **environmental impact** of AI development and use should be minimised as much as possible.

Fairness

*Objective: The development, deployment, use and operation of AI systems must be **fair**.*

Design and operating considerations:

Substantive:

- Data collection, design and development processes should **involve diverse individuals**, representative of likely end-users and other affected groups.
- AI systems should not (directly or indirectly) produce **discriminatory or biased outcomes** or exacerbate existing biases and inequities.
- AI systems should not be used to deceive people or unjustifiably impair their freedom of choice.
- Clear processes should be in place to ensure there is always a **human who can be held accountable** for the operation of an AI system.
- Accessible **complaints resolution processes** should be implemented to ensure effective redress for individuals harmed by AI systems.

Explicability

Objective: End-users and other members of the public should be able to **understand** *and* **challenge** *the operation of AI systems.*

Design and operating considerations:

- End-users and other affected individuals should be provided **clearly expressed information** on the operation, capabilities and key risks of AI systems. This may include information on the reasons, criteria and relative weighting of the AI system's outcomes.
- Information provided should be **appropriate for the likely user group**, taking particular care with AI systems likely to be used by children, the elderly, persons with disabilities, persons with particular cultural backgrounds and economically vulnerable individuals.
- Information provided should be sufficient to enable affected individuals to **understand and challenge decisions** made by or based on AI systems. Heightened transparency should be ensured where decisions present **particularly severe consequences** for individuals' lives.
- Where this is not possible for technical reasons (in so-called "black box" situations), **alternative transparency mechanisms** should be in place (e.g. traceability, auditability, or providing information on system capabilities).
- Individuals should always be aware when they are **interacting with an AI system** rather than a human.

It is expected that for specific AI use cases, ethical dilemmas and tensions may arise in meeting the full intent of the above ethical AI principles. The AI designer or operator is responsible to develop governance and control practices to identify and evaluate the potential impacts and trade-offs and to determine the best course of action. It should also be expected that in interpreting stakeholder expectations and ethical considerations, evaluations of what is trustworthy will be subjective and involve judgements that will vary across individuals, cultures and regions.

In applying the principles it is expected that there will be ambiguity in their definition and application, and that further clarity will be required over and above existing reference documents. Some examples include:

- Who decides what is justified and what is unjustified, particularly when the AI objective may subvert human autonomy?
- How is bias and fairness being defined and measured?
- How do you resolve a conflict when the preference of an individual is not lawful?
- How do you achieve fair and equitable AI outcomes between majority and minority sub-classes?
- How are minimum requirements determined across cross-cultural groups with different social norms and laws?
- How do you assess the accuracy and reasonableness of a system-generated explanation and rationale for a particular prediction outcome or decision?

In these situations, it is recommended that a process of reasoned, ethical-based reflection involving a wide group of stakeholders be undertaken to arrive at an equitable and lawful solution.

Specific requirements of Trustworthy AI

To achieve Trustworthy AI, the Ethical AI Principles discussed above must be translated into specific requirements for AI systems that can be measured and evaluated. These requirements are applicable across the full lifecycle of an AI system, however different stakeholders in the design and deployment of an AI system have a different role to play in ensuring that the requirements are met.

- **Designers and trainers** of AI need to incorporate these requirements into the design and objectives of the AI system, and include them as a fundamental part of the system design requirements of the AI system.

- **Deployers and operators** of AI need to ensure these requirements are met, and to put a continuous monitoring system in place to monitor the AI system's performance against these requirements.
- **End-users and broader society** should be informed of these conditions and given proof, upon request, that they are being met.

Although provided on the next page separately, the specific Trustworthy AI Requirements are inter-related and should be continuously evaluated and addressed throughout the AI system's lifecycle. The relevance and importance of each of the requirements will differ based on the particular context and stakeholder impact of an AI system. In addition, it is common for tensions to exist between one or more of the requirements which will need to be addressed.

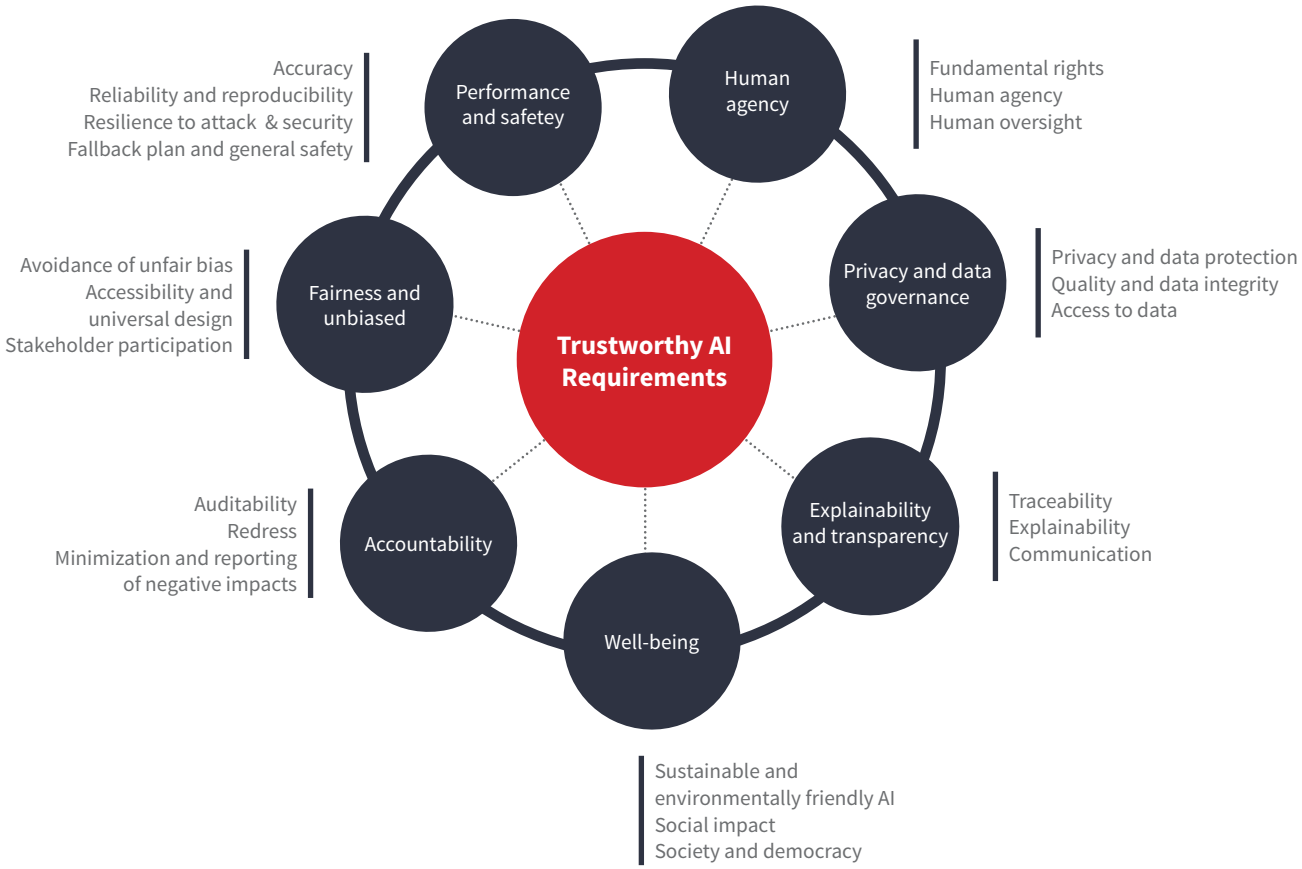

Figure 2: Trustworthy AI Requirements

The diagram above lists the key conditions to be met for each Trustworthy AI Requirement. The relevance and level of importance of each of the requirements will depend on the nature of each use case, and will need to take into consideration the objectives, data sets, functional components, level of autonomy, environmental conditions and impacts of each particular AI system. It is also important to note that some of the above requirements are already reflected in existing laws.

The above list of requirements is not exhaustive and further considerations may be required to meet the needs and expectations of stakeholders when designing and deploying an AI system.

In the next chapter, the above Trustworthy AI Requirements are further explained in relationship to the governance and control practices required to meet each requirement.

GOVERNANCE AND CONTROL PRACTICES

CHAPTER 3

AS ORGANISATIONS RACE TO BE FIRST IN THE DEVELOPMENT AND ADOPTION OF AI, IT IS IMPORTANT THAT GOVERNANCE AND CONTROL PRACTICES ARE DEVELOPED AT A COMPARABLE PACE.

It is common for early technology developers to prioritise innovation and benefit realisation, over controls. But with AI, as it can develop its decision framework through adaptive learning rather than pre-programmed code, its important that ethical, legal and regulatory requirements are considered as part of the training conditions for an AI system from the outset.

To achieve the Ethical AI Guidelines and Trustworthy AI Requirements outlined in Chapter 2, it is important that AI practitioners leverage existing control practices, while also developing new control practices that address the unique trust conditions necessary for AI.

Existing governance and IT general controls have a lot of control practices that can be leveraged for AI but will need to be modified to address the adaptive learning nature of AI. For example, change management controls will need to incorporate monitoring mechanisms for changes to decision frameworks that occur after an AI system is put into production and data ingestion practices will need to provide reliable and trusted data sources and consider the key role that data plays in training an AI system (e.g. garbage in, garbage out).

Due to the unique nature of AI and the quasi-cognitive nature of the activities it is performing, new control constructs will also be required to augment those historically used for legacy technology. The development of a robust system of controls for an AI system will require the retrofit of control practices over human-based processes to machine based-decisions, and the codification of ethical and trust objectives into system objectives and monitoring practices.

Academics and technicians are working hard to develop robust control practices to meet the needs of AI practitioners, however the field is still in its infancy and there is still a lot of work to be done. As the efficacy and adequacy of AI control practices mature, organisations designing or operating AI systems need to adopt a robust governance framework to proactively identify, measure and respond to the risks of AI, whilst at the same time implementing domain-specific control practices to ensure achievement of the Trustworthy AI Requirements.

GUIDANCE ON THE APPLICATION OF CONTROL PRACTICES FOR ETHICAL AND TRUSTWORTHY AI

Provided on the following pages are illustrative leading control practices for AI, first at the governance-level and then for each of the Trustworthy AI Requirements. They are intended to provide directional guidance to AI practitioners in developing a robust control framework for ethical and trustworthy AI.

They do not address all potential control practices regarding the governance of AI in Malta. Due to the diverse AI technologies being deployed and expansive use cases being considered, a one size fits all approach to addressing the ethical issues associated with AI will not be sufficient.

In addition to considering the following suggested controls, AI practitioners should also consider the unique risks of their AI systems and innovations in control practices over time and relevant international and domestic laws. It should be anticipated that there will be at times conflicts between one or more of the trust principles and/or suggested control practices. There may also be justifiable exceptions or a misalignment between obligations set through laws and regulations, and broader stakeholder expectations.

AI practitioners should use their judgement to understand their suitability and applicability to their organisations and AI systems.

When considering the sufficiency of governance and control practices over an AI system, an organisation should employ a risk-based approach which considers the full spectrum of obligations, risks and stakeholder impact of the AI system. It should be recognised that AI systems will be used in a broad set of uses and contexts, and have varied levels of complexity, stakeholder impact and control maturity which will need to be considered in determining the appropriate controls to put into place. As the risk profile and stakeholder impact of the AI system increases however, so should the governance and control practices. For example, the greater the impact of an AI prediction, decision or action on an affected stakeholder, the greater the obligation of a designer or operator will be to be transparent about its use, decision framework and impact.

As both AI systems and their governance and control mechanisms will continue to evolve, frequent re-validations should be performed to assess the adequacy of mitigation responses for existing or newly identified risks. Due to the anticipated fluidity of AI systems and their governance and control practices, organisations should be transparent with users as to any limitations, boundary conditions or risks that may adversely impact the user and for which they may need to compensate.

Trustworthiness in any technology but particularly AI will evolve over time as awareness, understanding and familiarity with AI systems evolves.

Leading governance control practices over AI

The following list of illustrative controls represent leading governance control practices for AI systems. It is expected that these controls would be applicable across an organisation's AI programme.

A. Internal governance processes and mechanisms (where developing AI)

Governance processes and mechanisms to establish oversight and incorporate a values into AI design, development and use

Staff	
1	Create mechanisms for staff to flag issues relating to: • bias/discrimination; • privacy/data protection; and • poor performance of the AI system. A communications and disclosure process should be in place to assess for issues if there is a requirement to disclose the details and impact of the issue to an end-user, and if so, on what timeframe and reporting mechanism.
2	Provide training and education on the AI governance, control and design framework to develop internal accountability practices, including on the ethical, legal frameworks and regulations applicable to the AI system.
3	Involve any relevant Data Processing Officers as early as possible in data collection and processing.
4	Assess whether the team involved in building the AI system is representative of the target user audience as well as the wider population, considering also other groups who might tangentially be impacted. If the team is found to not be representative of the target population, consider creating focus and test groups from the target user group.
5	Where seeking to implement AI in the workplace, educate and involve impacted workers and their representatives in advance.
Governance	
6	Introduce ethical AI considerations as corporate values.
7	Ensure clear roles and responsibilities for the ethical deployment of AI.
8	Establish the role of an Ethics Officer or consider assigning responsibility and accountability to a senior executive of the organisation for the lawful and ethical design, operation and use of AI systems in relation to ethical and legal obligations. This individual should also have responsibility for establishing an effective governance and control structure over the AI systems, which include compliance to all relevant laws, regulations and guidelines. This role could also be outsourced to an expert in the field for additional accountability, or instances where the appropriate expertise is not present in the organisation.

9	Consider establishing a cross-disciplinary AI advisory board or similar mechanism to monitor and assist an organisation in developing appropriate responses to AI ethical and legal obligations.
10	Consider bringing in external guidance or establishing auditing processes to oversee ethics and accountability, in addition to internal initiatives.
11	Taking out an insurance policy as a risk mitigation measure against potential damage from the AI system. As an alternative, consider a self-insurance mechanism where the AI designer or operator segregates a portion of funds as insurance.

B. Operations management

Framework for minimising risk and creating appropriate decision-making model (including Issues to consider when developing, selecting and maintaining AI models)

Processes, protocols, measures, procedures	
12	Establish a mechanism to identify, document and justify interests and values implicated by the AI system and potential trade-offs between them.
13	Implement process to ensure quality and integrity of data, including: • maintaining a data provenance record that allows the organisation to determine quality of data, trace potential sources of errors, update data and attribute data to their sources; • regular review and updating of datasets (including training, testing and validation datasets); • verification that datasets have not been compromised or hacked; • monitor characteristics and type of new incoming data for shift in underlying data that may adversely impact the outcomes; and • assessment of the extent to which quality of external data sources is controllable.
14	Implement protocols, processes or procedures to ensure: • only appropriately authorised and qualified persons can access personal data and only in appropriate circumstances; and • traceability of who has accessed personal information, when where, for how long and for what purpose.
15	Establish measures (such as an audit trail) to ensure traceability, including in relation to: • programming methods or how the model is built (for rule-based AI systems); • training methods, including which input data is collected and selected and how (for learning-based AI systems); • scenarios or cases used to test and validate (for rule-based AI systems); • detail on data used to test and validate (for learning-based AI systems); and • outcomes or decisions that could be made by or based on the algorithm. This may include implementation of a black box records that captures all input data streams.

16	Establish mechanisms to measure the environmental impact of the AI system's development, deployment and use (e.g. the amount of data used by the data centres).
17	Establish a strategy or set of procedures to avoid creating or reinforcing unfair bias in the AI system, in terms of both input data and algorithmic design, including: • assessment of limitations arising from dataset composition; • ensuring diversity and representativeness of users in the data and testing for specific populations or problematic test cases; • researching and using available technical tools to improve understanding of the data, model and performance; and • implementing processes to test and monitor for potential biases during development, deployment and use phases.
18	Establish mechanisms to: • measure whether the AI system is making an unacceptable amount of inaccurate predictions within pre-defined tolerance levels; • increase the AI system's accuracy; and • verify what harm may be caused if the AI system makes inaccurate predictions.
19	Assess what level and definition of accuracy is required in the context of the particular AI system and use case, including: • determining how accuracy will be measured and assured; • establishing measures to ensure that data used is comprehensive and up to date; and • establishing measures to assess whether there is a need for additional data (e.g. to improve accuracy or mitigate bias).
20	Adopt an adequate working definition of "fairness" aligned with industry and globally agreed upon definitions, as well as non-discriminatory laws, to use when designing AI systems and ensure a quantitative analysis or metrics to measure and test this applied definition of fairness.
21	Assess whether there is any possible decision variability that can occur under the same conditions. If so, consider what the possible causes of this could be.
22	Implement a strategy to monitor and test whether the AI system is meeting its goals, purposes and intended applications, including reliability and reproducibility.
23	Assess potential forms of attack to which the AI system could be vulnerable (including vulnerabilities such as data pollution, physical infrastructure and cyberattacks).
24	Establish measures or systems to ensure the integrity and resilience of the AI system against potential attacks.
25	In testing and validation subject the AI system to sensitivity and stress testing beyond its business-as-usual boundary conditions to determine how the AI system behaves in unexpected situations and environments.
26	Consider the degree to which the AI system could be dual-use and take appropriate preventative measures (e.g. non-publication of the research or non-deployment of the system).
27	Perform regular model tuning to reflect changes in market behaviour and preferences, incorporating new data into training sets where appropriate.

Assessments	
28	Assess whether the AI system encourages humans to develop attachment and empathy towards the system.
29	Assess and take steps to counteract negative social impacts of the AI system (e.g. risk of job loss or de-skilling of the workforce).
30	Assess whether the AI system may interfere with human decision-making in unintended ways and build in appropriate safeguards.
31	Assess the broader societal impacts of the AI system's use beyond the individual user (e.g. effects on indirectly affected stakeholders).
Design considerations	
32	Align system with relevant standards (e.g. ISO and IEEE) or widely adopted protocols for daily data management and governance.
33	Consider ways to develop the AI system or train the model with minimal use of potentially sensitive or personal data, including use of encryption, anonymisation or aggregation
34	Establish mechanisms that facilitate the system's auditability, such as ensuring traceability and logging of the AI system's processes and outcomes.
35	Design AI systems with explainability in mind from the outset, including by: • researching and attempting to use the simplest and most interpretable model possible for the application in question; • assessing whether it is possible to analyse, change and update training and testing data; and • assessing whether interpretability can be examined after the model's training and development, or whether the model's internal workflow can be accessed.
36	Ensure that personal data is processed only in accordance with the organisation's privacy or data protection policy, as well as all applicable legal requirements.
37	Ensure measures to reduce the environmental impact of the AI system's life cycle.
38	Assess and verify whether the AI system accommodates a wide range of individual preferences and abilities, including: • whether the AI system is usable by those with special needs/disabilities or those at risk of exclusion; and • whether particular persons or groups might be disproportionately affected by negative implications.
39	Ensure that information about the AI system is accessible for to users of assistive technologies.
40	Ensure that AI system has a sufficient fallback plan for adversarial attacks and other unexpected situations (e.g. technical switching procedures or asking for a human operator before proceeding).

41	As part of the risk assessment, assess the sufficiency of control mechanisms if an AI system fails or produces erroneous outcomes including the technical feasibility of a fallback to previous versions, "kill-switch" or human override.

C. End-user and third-party processes and mechanisms

42	Ensure users are made aware, in clear and easily understandable language: • that relevant decisions, content, advice or outcomes are the result of an algorithmic decision (unless there are compelling reasons in the public interest or the individual's interests weighing against this); • for chatbots and similar conversational systems, that they are interacting with a non-human agent (e.g. through a label or disclaimer). The communication should include sufficient information for the end-user to understand any inherent limitations, boundary conditions or potential risks of the AI system, including inherent bias.
43	Ensure individuals can exercise appropriate levels of control over their personal data (e.g. mechanisms for giving and revoking valid consent to different types of processing).
44	Ensure that, commensurate with the explainability requirement of an AI system, an explanation can be provided in a way which is understandable by all users as to why the AI system made a particular decision.
45	Ensure the AI system clearly indicates that its social interaction is simulated and that it has no capacities of "understanding" or "feeling".
46	Establish redress mechanisms for or other adverse impacts caused by the AI system.
47	Establish mechanisms to provide information to users and third-parties about opportunities for redress.
48	Establish procedures for third parties (e.g. suppliers, consumers, distributors, vendors) or workers to report potential vulnerabilities, risks or biases in the AI system.
49	During design and development phases, involve or consult individuals and groups likely to use or otherwise be affected by the AI system.
50	Establish communication mechanisms to assure users about the AI system's reliability.

D. Impact and risk assessments

51	Conduct privacy impact assessment.
52	Conduct a risk or impact assessment of the AI system that considers direct and indirect effects on different stakeholders.

53	Conduct a risk assessment of the AI system causing harm or damage to users or third parties, including: • assessment of the likelihood, nature and severity of different types of harm or damage; • consideration of potential impact or safety risk to the environment, natural resources and biodiversity; • consideration potential safety risks or damage caused by security or network problems such as cybersecurity hazards due to unintentional behaviour of the AI system; and • consideration of liability and consumer protection rules.
54	Identify potential safety risks of foreseeable uses of the AI system, including accidental or malicious misuse and create a plan to measure/assess and mitigate/manage these risks.
55	Estimate the likely impact of a failure of the AI system when it provides wrong results, becomes unavailable, or provides societally unacceptable results (e.g. discrimination).
56	Test and implement governance procedures to trigger fallback plans, including definition of relevant thresholds.

E. Internal processes and mechanisms (where implementing/using AI)

57	Before deploying AI systems, decide on the commercial objectives of using AI and weigh them against the risks of using AI in decision-making, in line with the organisation's corporate values.
58	When implementing an AI system in an organisation, assess: • how understandable the AI system's decisions and outcomes are; • the degree to which the AI system influences the organisation's decision-making processes; • the purpose for which the AI system is deployed in a particular area; and • what the particular AI system's business model is (e.g. how it creates value for the organisation).
59	Where operating in multiple countries, consider differences in laws, societal norms and values.
60	Where AI systems are implemented in work and labour processes, ensure that: • safeguards are in place to prevent overconfidence or overreliance on the AI system; and • task allocation between human and automated processes enhances or augments human capabilities.
61	Implement mechanisms and measures to: • ensure a level of human control appropriate for the particular AI system and use case as determined in the requirements definition and risk assessment (ensuring particularly robust control and oversight for self-learning or autonomous AI systems); and • audit and remedy issues relating to AI autonomy.
62	Ensure a stop button or procedure to safely abort a procedure or delegate control to a human.
63	Establish a mechanism to identify, document and justify interests and values implicated by use of the AI system and potential trade-offs between them.

Leading control practices to achieve Trustworthy AI Requirements

The applicability of the Trustworthy AI Requirements will vary depending on the objectives set for an AI system, its functional design and capabilities, data collected and processed, and the ultimate impact to users. In implementing the leading control practices described above, an organisation may need to adapt the nature and design of the control to address the specific Trustworthy AI Requirements relevant to each AI system.

Provided below, we have re-listed the leading control practices to demonstrate their applicability to each requirement area. In some cases, the full control is described and in others, more granular guidance is provided on how the control activity could be operated to meet the stated Trustworthy AI Requirement.

1. Human agency

Sub-requirement	*Control*
a) Fundamental rights *Ensure that AI development and use does not breach fundamental rights recognised at international and domestic law*	i. Conduct **human rights impact assessment**, identifying and documenting potential trade-offs between different principles and rights. ii. Actively consider whether the AI system **may interfere with human decision-making** in unintended ways and build in appropriate safeguards. iii. Provide measures to make users aware: A. that relevant decisions, content, advice or **outcomes are the result of an algorithmic decision**; and B. for chatbots and similar conversational systems, that **they are interacting with a non-human** agent.
b) Human agency *Ensure appropriate level of human engagement with AI*	i. Where AI systems are implemented in work and labour processes, ensure that: A. safeguards are in place to **prevent overconfidence** in or overreliance on the AI system (e.g. automation bias) ; and B. task allocation between human and automated processes **enhances or augments human capabilities**.
c) Human oversight *Ensure appropriate level of human oversight of AI*	i. Implement mechanisms and measures to: A. ensure a level of human control appropriate for the particular AI system and use case; B. audit and remedy issues relating to AI autonomy. ii. Ensure particularly robust control and oversight (including detection and response mechanisms) for self-learning or autonomous AI systems. iii. Ensure a stop button or procedure to safely abort a procedure or delegate control to a human.

2. Privacy and data governance

Sub-requirement	*Control*
a) Privacy and data protection *Ensure protection of individuals' privacy rights, including compliance with all relevant data processing laws*	***NOTE** that these controls are not an exhaustive statement of, or substitute for compliance with, existing legal requirements on data protection, such as the General Data Protection Regulation and related domestic Maltese laws.* i. Determine the type and scope of data to be used in AI development. ii. Conduct a Privacy Impact Assessment and ensure compliance with all applicable legislative requirements relating to data protection, including: A. **transparency:** notifying individuals that you are collecting their personal information, the purposes for which you will process it, to whom (if anyone) you will disclose it, how you will store the information, and other key information; B. **lawful basis for processing:** ensure that you have a lawful basis for the intended processing of that data, including obtaining valid consent from the individual where appropriate. iii. Ensure individuals can exercise appropriate levels of control over their personal data (e.g. mechanisms for giving and revoking valid consent to different types of processing). iv. Ensure that personal data is processed only in accordance with the organisation's privacy or data protection policy, as well as all applicable legal requirements. v. Involve any relevant Data Processing Officers as early as possible in data collection and processing. vi. Implement an internal mechanism for individuals to flag privacy issues related to data collection and processing. vii. Consider ways to develop the AI system or train the model with minimal use of potentially sensitive or personal data, including use of encryption, anonymisation or aggregation.
b) Quality and data integrity *Ensure quality and integrity of data used in AI design, development and training*	i. Align system with **relevant standards (e.g. ISO and IEEE)** or widely adopted protocols for daily data management and governance. ii. Establish **oversight mechanisms** for data collection, storage, processing and use. iii. Establish **oversight mechanism**s for data collection, storage, processing and use. A. verification that datasets have not been compromised or hacked; and B. assessment of the extent to which quality of external data sources is controllable.
c) Access to data *Protect against unauthorised access to data*	i. Implement protocols, processes or procedures to ensure: A. only appropriately authorised and qualified persons can access personal data and only in appropriate circumstances; B. traceability of who has accessed personal information, when where, for how long and for what purpose.

3. Explainability and transparency

Sub-requirement	*Control*
a) Traceability *Ensure traceability of processes used and decisions made in AI design and development*	Establish measures to ensure traceability, including: ***Design and development phase*** i. Programming methods or how the model is built (for rule-based AI systems) ii. **Training** methods, including which input data is collected and selected and how (for learning-based AI systems) ***Testing and validation phase*** iii. **Scenarios or cases used** to test and validate (for rule-based AI systems) iv. Detail on **data used** to test and validate (for learning-based AI systems) ***Outcomes of the algorithmic system*** v. **Outcomes or decisions** that could be made by or based on the algorithm.
b) Explainability *Ensure that end-users and other affected individuals can understand the operation of the AI system*	i. When implementing an AI system in an organisation, assess: A. how understandable the AI system's decisions and outcomes are; B. the degree to which the AI system influences the organisation's decision-making processes; C. the purpose for which the AI system is deployed in a particular area; and D. what the particular AI system's business model is (e.g. how it creates value for the organisation). ii. Ensure that an explanation can be provided in a way which is understandable by all users as to why the AI system made a particular decision. iii. Design AI systems with explainability in mind from the outset, including by: A. researching and attempting to use the simplest and most interpretable model possible for the application in question; B. assessing whether it is possible to analyse, change and update training and testing data; C. assessing whether interpretability can be examined after the model's training and development, or whether the model's internal workflow can be accessed.
c) Communication *Provide appropriate communicate to end-users*	i. Communicate to end-users that they are interacting with an AI system rather than a human (e.g. by way of a label or disclaimer)

4. Wellbeing

Sub-requirement	*Control*
a) Sustainable and environmentally friendly AI *Ensure negative environmental impacts of AI development and use are minimised*	i. Establish mechanisms to **measure the environmental impact** of the AI system's development, deployment and use (e.g. the amount of data used by the data centres). ii. Ensure measures to **reduce the environmental impact** of the AI system's life cycle.
b) Social impact *Ensure negative social impacts of AI development and use are minimised*	i. Assess whether the AI system **encourages humans to develop attachment** and empathy towards the system. ii. Ensure the AI system clearly indicates that its **social interaction is simulated** and that it has no capacities of "understanding" or "feeling". iii. Assess and take steps to counteract **negative social impacts** of the AI system (e.g. risk of job loss or de-skilling of the workforce).
c) Society and democracy *Ensure indirect negative social impacts of AI development and use are minimised*	i. Assess the broader societal impacts of the AI system's use beyond the individual user (e.g. effects on indirectly affected stakeholders).

5. Accountability

Sub-requirement	*Control*
a) Auditability *Ensure the AI system is auditable*	i. Establish mechanisms that facilitate the system's auditability, such as ensuring **traceability and logging** of the AI system's processes and outcomes. ii. In applications affecting fundamental rights, ensure that the AI system can be **audited independently**.
b) Redress *Ensure individuals can seek redress for harms cause by AI*	i. Establish redress mechanisms for or other adverse impacts caused by the AI system and provide clear information on these to users and other affected individuals.
c) Minimisation and reporting of negative impacts *Minimise negative impacts of the AI system*	i. Conduct a risk or impact assessment of the AI system that considers direct and indirect effects on different stakeholders. ii. Provide **training and education** to develop accountability practices, possibly including the legal framework applicable to the AI system. iii. Consider establishing an **"ethical AI board"** or similar mechanism to discuss overall accountability and ethics practices, including potentially grey areas. iv. Consider bringing in **external guidance** or establishing auditing processes to oversee ethics and accountability, in addition to internal initiatives. v. Establish procedures for third parties (e.g. suppliers, consumers, distributors, vendors) or workers to **report potential vulnerabilities**, risks or biases in the AI system.
d) Documenting trade-offs	i. Establish a mechanism to identify, document and **justify interests and values** implicated by the AI system and potential trade-offs between them.

6. Fairness and lack of bias

Sub-requirement	*Control*
a) Avoidance of unfair bias *Ensure the AI system does not create or perpetuate unfair bias*	i. Establish a strategy or set of procedures to a**void creating or reinforcing unfair bias** in the AI system, in terms of both input data and algorithmic design, including: A. assessment of limitations arising from dataset composition; B. ensuring diversity and the appropriate representation of users in the data and testing for specific populations or problematic test cases; C. researching and using available technical tools to improve understanding of the data, model and performance; and D. implementing processes to test and monitor for potential biases during development, deployment and use phases. ii. Implement a mechanism for **individuals to flag issues about bias**, discrimination or poor performance of the AI system (in relation to end users as well as indirectly affected individuals). Clearly communicate to relevant stakeholders how and with whom they can raise such issues. iii. Adopt an adequate **working definition of "fairness"** to use when designing AI systems, after considering a variety of definitions and how commonly used they are. iv. Ensure a quantitative analysis or metrics to **measure and test this applied definition** of fairness. v. Establish mechanisms to ensure fairness in AI systems. vi. Assess whether there is any possible decision variability that can occur under the same conditions. If so, consider what the possible causes and/or effects of this could be. vii. Establish a measurement or assessment mechanism of the potential impact of such variability on fundamental rights.
b) Accessibility and universal design *Ensure the AI system caters for users with special needs*	i. Assess and verify whether the AI system accommodates a wide range of individual preferences and abilities, including: A. whether the AI system is usable by those with special needs/ disabilities or those at risk of exclusion; and B. whether particular persons or groups might be disproportionately affected by negative implications. ii. Ensure that information about the AI system is accessible for to **users of assistive technologies**. iii. Involve or **consult these communities** during the development phase of the AI system. iv. Assess whether the team involved in building the AI system is **representative of the target user audience** as well as the wider population, considering also other groups who might tangentially be impacted.

c) Stakeholder participation *Ensure participation of diverse individuals in AI design, development and deployment*	i. **Identify stakeholders that could be impacted by the AI** system and define the potential impacts and likelihood. ii. Include the **participation of different stakeholders** in the AI system's development and use through focus groups, surveys and / or feedback mechanisms. Where seeking to implement AI in the workplace, inform and **involve impacted workers** and their representatives in advance.

7. Performance and safety

Sub-requirement	*Control*
a) Accuracy *Ensure accuracy of AI system's outputs*	i. Assess what level and definition of accuracy is required in the context of the particular AI system and use case, including: A. determining **how** accuracy will be measured and assured; B. establishing measures to ensure **that data used is comprehensive and up to date**; C. establishing measures to assess whether there is a **need for additional data** (e.g. to improve accuracy or mitigate bias). ii. Establish mechanisms to: A. measure whether the AI system is making an **unacceptable amount of inaccurate predictions** based on pre-defined tolerance levels and quality assurance activities; B. continuously monitor the AI system's performance against pre-defined tolerance levels; and C. increase the AI system's accuracy. iii. Verify **what harms may be caused** if the AI system makes inaccurate predictions.
b) Reliability and reproducibility *Ensure reliability of the AI system*	i. Implement a strategy to monitor and test **whether the AI system is meeting its goals**, purposes and intended applications, including reliability and reproducibility. ii. Establish communication mechanisms to assure users about the AI system's reliability.

c) Resilience to attack and security *Mitigate the AI system's vulnerabilities*	i. Assess **potential forms of attack** to which the AI system could be vulnerable (including vulnerabilities such as data pollution, physical infrastructure and cyberattacks). ii. Establish measures or systems to ensure the integrity and resilience of the AI system against potential attacks. iii. Verify how the **AI system behaves in unexpected situations** and environments. iv. Consider the degree to which the **AI system could have built-in quality, protection and safety mechanisms** and take appropriate preventative measures (e.g. non-publication of the research or non-deployment of the system).
d) Fallback plan and general safety *Ensure the AI system is developed and used safely*	i. Ensure that AI system has a sufficient **fallback plan** for adversarial attacks and other unexpected situations (e.g. technical switching procedures or asking for a human operator before proceeding). ii. If there is a risk to human physical integrity, **provide necessary information** to all relevant individuals. iii. Consider taking out an **insurance policy** to deal with potential damage from the AI system. iv. **Identify potential safety risks** of foreseeable uses of the AI system, including accidental or malicious misuse and create a plan to measure/assess and mitigate/manage these risks. v. **Conduct a risk assessment** of the AI system causing harm or damage to users or third parties, including: A. assessment of the likelihood, nature and severity of different types of harm or damage; B. consideration of potential impact or safety risk to the environment, natural resources and biodiversity; C. consideration potential safety risks or damage caused by security or network problems such as cybersecurity hazards due to unintentional behaviour of the AI system; and D. consideration of liability and consumer protection rules. vi. **Estimate the likely impact** of a failure of the AI system when it provides wrong results, becomes unavailable, or provides societally unacceptable results (e.g. discrimination). vii. **Test and implement governance procedures** to trigger fallback plans, including definition of relevant thresholds.

AI CERTIFICATION

CHAPTER 4

MALTA HAS TAKEN A GLOBAL LEAD IN DEVELOPING A REGULATORY AND CERTIFICATION FRAMEWORK FOR INNOVATIVE TECHNOLOGY ARRANGEMENTS (ITAS) THROUGH THE SET-UP OF THE MALTA DIGITAL INNOVATION AUTHORITY AND THE CREATION OF THE INNOVATIVE TECHNOLOGIES AND SERVICES ACT (ITAS ACT).

The MDIA is the primary Authority responsible for promoting all governmental policies that promote Malta as the centre for excellence for technological innovation, while setting and enforcing standards that ensure compliance with any other international obligations.

The MDIA has developed a certification framework which aims to provide a standard mechanism to build trust and transparency amongst users, consumers and wider stakeholders in ITAs.

In support of the Malta National AI Strategy and the achievement of the Malta Ethical AI Guidelines, the MDIA is in the process of expanding the ITA definition to include a certification framework for AI-based solutions, which will include AI-specific Control Objectives and Evaluation Criteria.

It will be the world's first national AI certification programme and aims to give a platform to AI solutions that have been developed in an ethically aligned, transparent and socially responsible manner. The ambition is to create the conditions for AI to springboard from Malta to the world, in line with Malta's vision to become the **Ultimate AI Launchpad**.

The AI certification programme will launch in October 2019. The MDIA recognises that it is working at the forefront of technology and regulatory framework in terms of the innovative technologies reviewed by it. The main challenge is to find the right balance between the MDIA's regulatory function, whilst not hindering innovation in the field of AI through doing so.

The MDIA's intention in developing a certification framework for AI-based technologies is to provide a structured approach in the production of detailed, concise, practical AI-guidelines that can aid an AI practitioner in the development of trustworthy AI.

The AI-based ITA certification framework will be developed in the format of existing MDIA guidelines to facilitate an integrated approach for adherents across their full technology portfolio.

The AI-based ITA certification framework will align to the Malta Ethical AI Framework, including the leading control practices identified in the previous chapter. The system control objectives developed will be based on a design that are built for the AI of today but scalable to the AI of tomorrow.

The AI-based systems guidelines will provide a delineation between enterprise-wide governance and IT general-control objectives and AI-specific objectives. This construct will enable Malta to adopt a more agile methodology to update the ITAs as AI technology and use cases evolve.

For further information visit: **www.mdia.gov.mt**

MALTA.AI
Towards a National AI Strategy

PARLIAMENTARY SECRETARIAT FOR FINANCIAL SERVICES,
DIGITAL ECONOMY AND INNOVATION
OFFICE OF THE PRIME MINISTER