Ethics in Artificial Intelligence (Module 1)

Last update: 23 February 2025

Academic Year 2024 – 2025
Alma Mater Studiorum · University of Bologna

Contents

1 Trustworthy AI

The European Commission's vision for artificial intelligence is based on three pillars:

1. Increase public and private investments,
2. Prepare for socio-economic changes (e.g., protect who gets substituted with AI),
3. Ensure a proper ethical and legal framework to strengthen European values.

To achieve this, in 2018 the Commission established the **High-Level Expert Group on Artificial Intelligence (AI HLEG)**: an independent group tasked to draft:

High-Level Expert Group on Artificial Intelligence (AI HLEG)

- Guidelines for AI ethics,
- Policy and investments recommendations.

1.1 AI HLEG's AI Ethics Guidelines

Voluntary framework addressed to all AI stakeholders (from designers to end-users) that bases AI trustworthiness on three components:

Lawful

Lawful AI must adhere to laws and regulations. The main legal sources are:

1. EU primary law (i.e., EU Treaties and Fundamental Rights).
2. EU secondary law (e.g., GDPR, . . .).
3. International treaties (e.g., UN Human Rights treaties, Council of Europe conventions, . . .).
4. Member State laws.
5. Domain-specific laws (e.g., regulations for medical data, . . .)

> **Remark.** The guidelines do not provide legal guidance. Therefore, this component is not explicitly covered in the document.

Ethical

Ethical AI must be in line with ethical principles and values (i.e., moral AI) for which laws might be lacking or unsuited for the purpose.

Robust

Robust AI must be technically and socially robust in order to minimize intentional or unintentional harm.

> **Remark.** Each individual component is necessary but not sufficient. Ideally, they should all be respected. If in practice there are tensions between them, it is responsibility of the society to align them.

> **Remark** (Law vs ethics)**.**
>
> Law
>
> **Law** Norms adopted and enforced by institutional entities.
>
> Ethics
>
> **Ethics** Norms that guide what should be done (instead of what can be done). It is rooted in shared societal values.

Example (Ethical washing)**.** To pursue their interests, some entities push to avoid regulations (which must be enforced) and state to adhere to ethical values (which are not explicitly enforced).

Example (Brussels effect)**.** Extension of EU regulations to other countries due to economic reasons (e.g., it is economically more convenient to have a single system respecting the EU's GDPR instead of having two separate ones).

The document itself is composed of three chapters:

Foundations of trustworthy AI Describes the ethical principles an AI should respect.

Realization of trustworthy AI Describes the requirements to achieve trustworthiness.

Assessment of trustworthy AI Describes trustworthiness assessment methods.

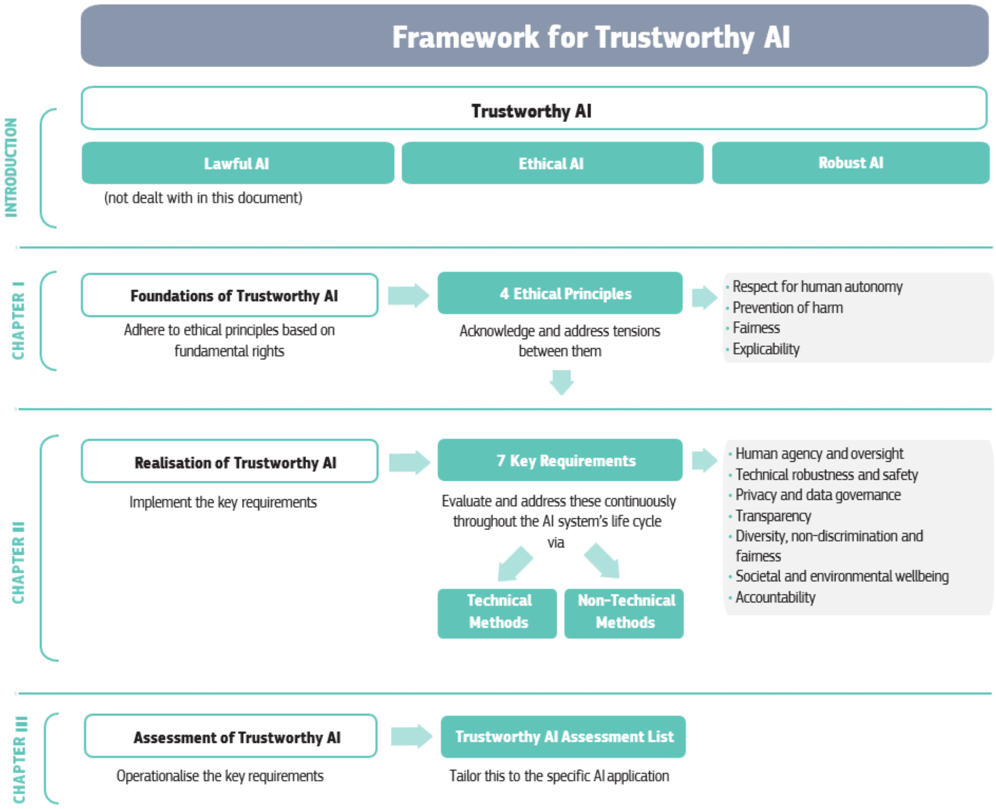
Framework for Trustworthy AI
INTRODUCTION
Trustworthy AI
Lawful AI
Ethical AI
Robust AI
(not dealt with in this document)
CHAPTER I
Foundations of Trustworthy AI
Adhere to ethical principles based on fundamental rights
4 Ethical Principles
Acknowledge and address tensions between them
Respect for human autonomy
Prevention of harm
Fairness
Explicability
CHAPTER II
Realisation of Trustworthy AI
Implement the key requirements
7 Key Requirements
Evaluate and address these continuously throughout the AI system's life cycle via
Technical Methods
Non-Technical Methods
Human agency and oversight
Technical robustness and safety
Privacy and data governance
Transparency
Diversity, non-discrimination and fairness
Societal and environmental wellbeing
Accountability
CHAPTER III
Assessment of Trustworthy AI
Operationalise the key requirements
Trustworthy AI Assessment List
Tailor this to the specific AI application

Figure 1.1: General overview of the document

1.1.1 Chapter I: Foundations of trustworthy AI

The concept of AI ethics presented in the framework is rooted to the fundamental rights described in the EU Treaties, EU Charter, and international human rights laws.

Remark (Fundamental rights)**.**

- Respect human dignity as moral subjects rather than objects in the pipeline of the system. AI systems should protect humans' physical and mental integrity, personal and cultural identity, and essential needs.
- Guarantee individual's freedom such as freedom of business, of the arts and science,

of expression, of assembly, and the right of privacy. AI systems should be mitigated for coercion, threats, surveillance, deception, ...

- Guarantee equality, non-discrimination, and solidarity. The output of an AI system should not be biased. Vulnerable groups that risk exclusion should be respected.
- Respect for democracy and citizen's rights. AI systems should not undermine democratic processes or citizen's rights such as the right to vote, to access public documents, to petition, ...

Remark. Seen as legally enforceable rights, fundamental rights can be considered as part of the LAWFUL AI component. Seen as the rights of everyone, from a moral status, they fall within the ETHICAL AI component.

This chapter describes four ethical principle for trustworthy AI based on fundamental rights:

Principle of respect for human autonomy

Principle of respect for human autonomy AI users should keep full self-determination. AI systems should be human-centric leaving room for human choices and they should not manipulate them.

Principle of prevention of harm

Principle of prevention of harm AI systems should operate in technically robust and safe environments. Attention must be paid to groups vulnerable to exclusion and to those subject to power asymmetries (e.g., employer-employee).

Principle of fairness

Principle of fairness The concept of fairness is described in a substantive and procedural dimension. The substantive dimension implies unbiased outputs and an equal distribution between benefits and costs. The procedural dimension involves the ability to contest and correct decisions made by AI systems and by humans using them.

Principle of explicability

Principle of explicability AI systems need to be transparent, their capabilities and purpose should be communicated, and their decisions should be as explainable as possible. For black box algorithms, alternative explicability measures might be needed (e.g., traceability, auditability, and communication of capabilities). Also, the degree of explicability that is required is dependent on the context and the use case.

Remark. There might be tensions between these principles (e.g., between prevention of harm and human autonomy in predictive policing) and methods to deal with them have to be established. Overall, the benefits of AI systems should exceed the risks. Practitioners should study these trade-offs in a reasoned and evidence-based way and not solely based on intuition.

1.1.2 Chapter II: Realization of trustworthy AI

This chapter defines concrete requirements from the principles of Chapter I. Stakeholders that these requirements involve are:

Developers

Developers Who research, design, and develop AI systems. They should concretely apply these requirements.

Deployers

Deployers Who use AI systems in their business processes and offer products or services to others. They should ensure that the systems they use meet the requirements.

End-users

End-users Who use the final AI system. They should be informed of these requirements and can request that they are respected.

The main requirements the framework defines are:

Human agency and oversight

Human agency and oversight AI systems should enhance human autonomy and decision-making (principle of respect for human autonomy):

- If there is the risk of violating fundamental rights, a study of the impacts should be conducted to justify it. External feedback should also be considered.
- Users should be provided with the necessary knowledge and tools to comprehend and interact with AI systems.
- Users have the right to not be subject to only automatic decisions if this significantly affects them.
- There should be oversight mechanisms (of varying degrees depending on the risk) to prevent AI systems from undermining human autonomy:
  - Human-in-the-loop (HITL): human intervention in every decision.
  - Human-on-the-loop (HOTL): human intervention in the design cycle and monitoring of the system's operation.
  - Human-in-command (HIC): human to decide if, when, and how to use an AI system in any particular situation.

  Public enforcers should also have the ability to exercise oversight with proper authorizations.

Technical robustness and safety

Technical robustness and safety There should be preventative measures to minimize unintentional harm (principle of prevention of harm):

- AI systems should be protected against vulnerabilities and attacks that target the data (data poisoning), the model (model leakage), or the infrastructure.
- Possible unintended uses or abuse of the system should be taken into account and mitigated.
- There should be fallback plans in case of problems (e.g., switching from a statistical to a rule-based algorithm, asking a human, . . .).
- There should be an explicit evaluation process to assess the accuracy of the AI system and determine its error rate.
- The output of an AI system should be reliable (robust to a wide range of inputs) and reproducible.

Privacy and data governance

Privacy and data governance Quality and security of the data should be guaranteed through the lifecycle of the AI system (principle of prevention of harm):

- Data provided by the user and derived from it should be protected and not used unlawfully or unfairly.
- Datasets should be cleared from biases, inaccuracies, and errors before training.
- The integrity of the datasets must be ensured to prevent malicious attacks.
- Processes and datasets should be tested and documented.

Transparency

Transparency There should be transparency in all the elements of an AI system (principle of explicability):

- The construction process of the dataset and the processes that lead to the AI system's decision should be documented.

- Decisions made by an AI system should be understandable and traceable by a human.
- The reason to use an AI system and the degree to which it influences decision-making and design choices should be stated.
- AI systems should not present themselves as humans and users have the right to be informed if they are interacting with an AI system. Depending on the use case, there should be the option to interact with a human.
- Capabilities and limitations of an AI system should be communicated to practitioners or end-users.

Diversity, non-discrimination, and fairness

Diversity, non-discrimination, and fairness Inclusion and diversity should be considered in the entire lifecycle of an AI system (principle of fairness):

- Biases should be removed from the data during the collection phase. Oversight processes should be put in place.
- AI systems should be user-centric and designed to be accessible by all people, regardless of disabilities.
- Stakeholders who might be affected by the AI system should be consulted.

Societal and environmental well-being

Societal and environmental well-being The impact of AI systems should also consider society in general and the environment (principles of fairness and prevention of harm):

- The environmental impact of the lifecycle of an AI system should be assessed.
- The effects of AI systems on people's physical and mental well-being, as well as institutions, democracy, and society should be assessed and monitored.

Accountability

Accountability Clear responsibilities should be defined for decisions made by AI systems (principle of fairness):

- Internal or external auditors should assess algorithms, data, and design processes.
- Potential negative impacts of AI systems should be identified, assessed, documented, and minimized.
- When there is tension between some of these requirements, trade-offs should be studied methodologically.
- There should be a redress mechanism for unjust decisions made by AI systems.

The chapter also describes some technical and non-technical methods to ensure trustworthy AI:

Technical methods

Technical methods

Architecture for trustworthy AI Embed trustworthiness requirements into the AI system as procedures or constraints.

Ethics and rule of law by design Methods to provide some properties by design.

Explanation methods Use techniques to understand the underlying mechanisms.

Testing and validating Define tests and validate the system in its entire lifecycle.

Quality of service indicators Use indicators to set the baseline for a trustworthy AI.

Non-technical methods

Non-technical methods

- **Regulation** Revise, adapt, or introduce regulations.
- **Codes of conduct** Describe how the organization intends to use AI systems.
- **Standardization** Define standards for a trustworthy system.
- **Certification** Create organizations to attest that an AI system is trustworthy.
- **Accountability via governance frameworks** Organizations should appoint a person or a board for decisions regarding ethics.
- **Education and awareness** Educate, and train involved stakeholders.
- **Stakeholder participation and social dialogue** Ensure open discussions between stakeholders and involve the general public.
- **Diversity and inclusive design teams** The team working on an AI system should reflect the diversity of users and society.

1.1.3 Chapter III Assessment of trustworthy AI

This chapter defines a generic assessment list to implement the requirements of Chapter II. The list has been devised by first taking feedback from a small selection of companies, organizations, and institutions that implemented it. Then, it was extended to all stakeholders and another round of feedback was taken.

Assessment list Steps to concretely assess the trustworthiness of an AI system. The main considerations to take into account are that:

Assessment list

- It should be tailored based on the specific use case.
- It can be integrated into existing governance mechanisms.
- It is continuously improved.

> **Remark.** In its pilot version, the list is composed of a series of questions for each requirement described in Chapter II.