Trustworthy AI

What makes an AI system trustworthy? (HLEG AI 2019)

trustworthy AI systems need to be:

(1) lawful: complying with all applicable laws and regulations

(2) ethical: ensuring adherence with ethical principles and values

(3) robust: from a technical and social perspective

Ethical principles for AI

- respect for autonomy
- prevention of harm
- fairness
- explicability

Additional ethical requirements

Pay particular attention to vulnerable groups, e.g. children, persons with disabilities, disadvantaged persons, situations of power asymmetries

Acknowledge potential risks even of beneficial systems, and the difficulties of correctly predicting them

Take into account risks even if they are difficult to measure, e.g. impacts on democracy, the human mind,

Adopt adequate measures to mitigate risks, proportionate to the magnitude of risk

Seven key requirements for Trustworthy AI

(1) human agency and oversight,

(2) technical robustness and safety,

(3) privacy and data governance,

(4) transparency,

(5) diversity, non-discrimination and fairness,

(6) environmental and societal well-being

(7) accountability

Figure 2: Interrelationship of the seven requirements: all are of equal importance, support each other, and should be implemented and evaluated throughout the AI system's lifecycle

1. Human agency and oversight

AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights.

At the same time, proper oversight mechanisms need to be ensured, which can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches

2. Technical robustness and safety

- AI systems need to be resilient and secure.
- They need to be safe, ensuring a fall back plan in case something goes wrong.
- They need to be accurate, reliable and reproducible.
- Unintentional harm needs to be minimized and prevented.

3. Privacy and data governance

- full respect for privacy and data protection needs to be ensured
- adequate data governance mechanisms must be ensured,
- the quality and integrity of the data needs to be taken into account
- only legitimised access to data should occur

4. Transparency

The data, system and AI business models should be transparent. Traceability mechanisms can help achieving this.

AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned.

Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations.

5. Diversity, non-discrimination and fairness

Unfair bias must be avoided, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination.

AI systems should be accessible to all, regardless of any disability

Relevant stakeholders should be involved throughout the entire life circle of AI systems

6. Societal and environmental well-being

- AI systems should benefit all human beings, including future generations.
- It must be ensured that they are sustainable and environmentally friendly
- They should take into account the environment, including other living beings,
- Their social and societal impact should be carefully considered.

7. Accountability

Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes.

Auditability, which enables the assessment of algorithms, data and design processes plays a key role therein, especially in critical applications.

An adequate and accessible redress should be ensured.

1. Additional implementation requirements: technology in society

- Consider technical and non-technical measures to ensure implementation
- Foster research & innovation to help assess AI systems and how they meet requirements
- Foster dissemination of knowledge to wider public
- Systematically train experts in AI ethics

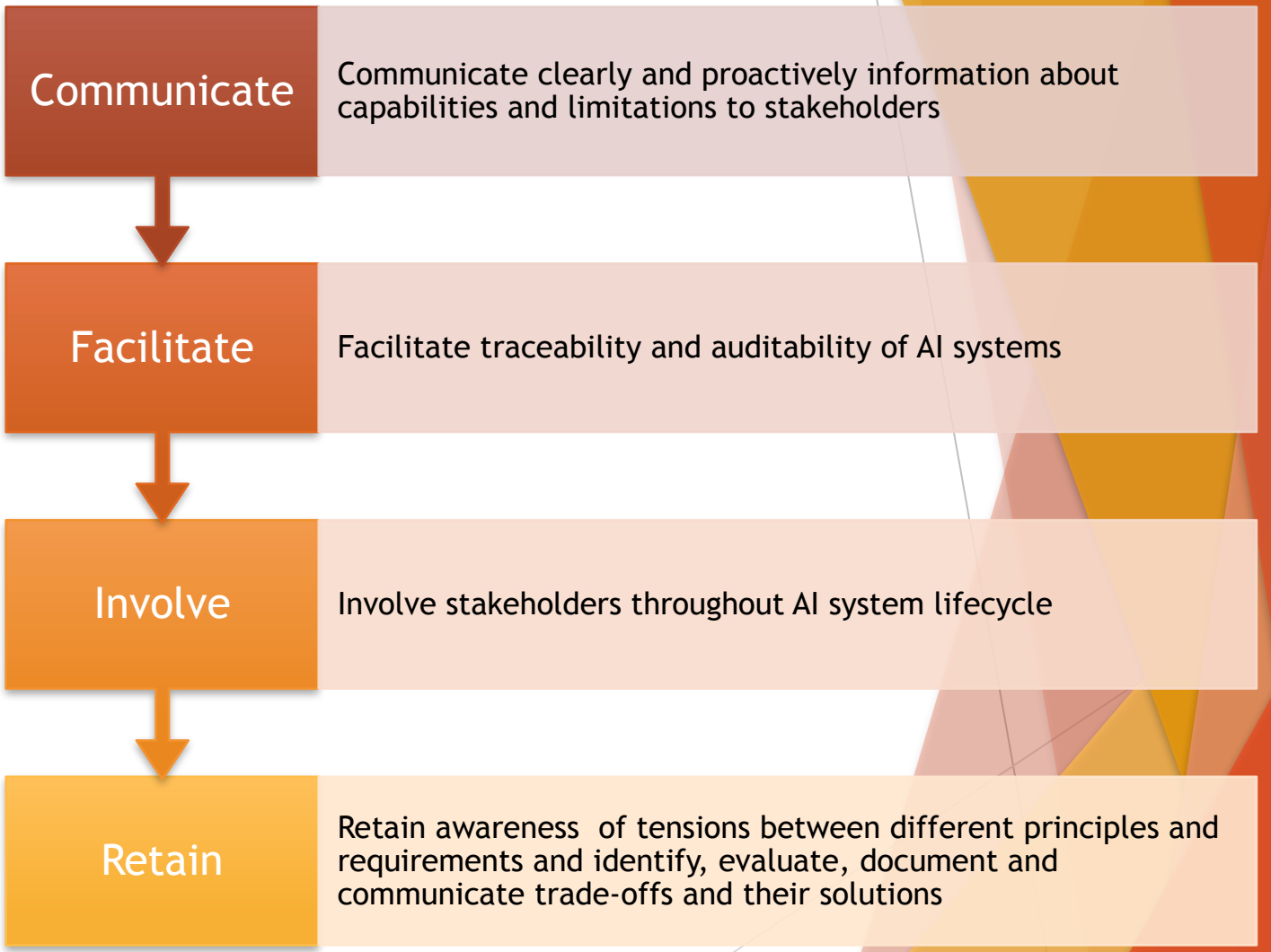
Technical methods

- Architectures for trustworthy AI scaffolding trustworthiness requirements
- Ethics and rule of law by design
- Explanation methods (XAI)
- Testing and validation
- Quality of service indicators

Non-technical methods

- Regulation
- Codes of conduct
- Standardisation
- Certification
- Accountability via governance frameworks
- Education and awareness for ethical mindset
- Stakeholder participation and social dialogue
- Diversity and inclusive design teams

2. Additional implementation requirements: stakeholders‘ reflections



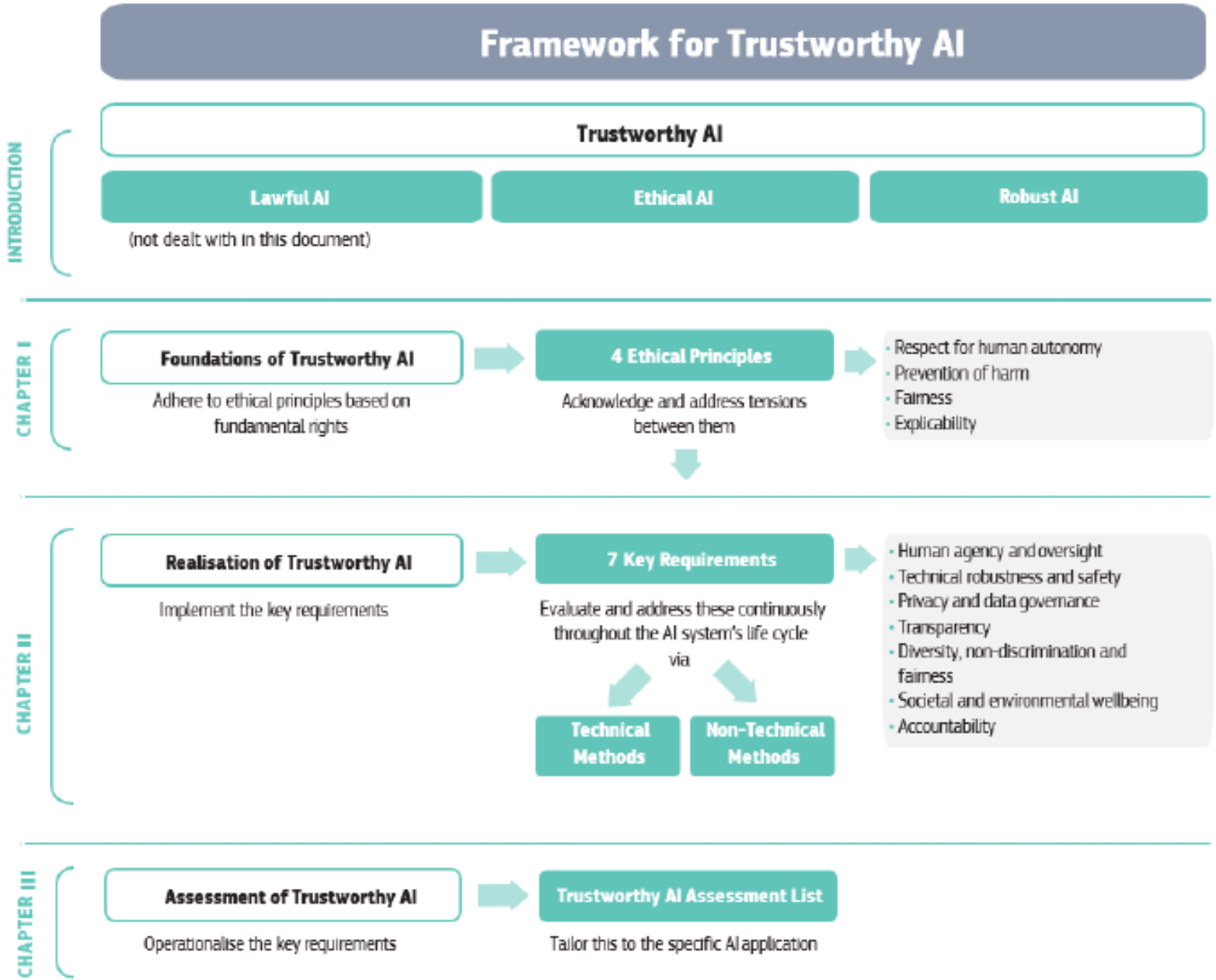


Figure 1: The Guidelines as a framework for Trustworthy AI

Assessment by comprehensive checklist

TRUSTWORTHY AI ASSESSMENT LIST (PILOT VERSION)

1. Human agency and oversight

Fundamental rights:

- ✓ Did you carry out a fundamental rights impact assessment where there could be a negative impact on fundamental rights? Did you identify and document potential trade-offs made between the different principles and rights?
- ✓ Does the AI system interact with decisions by human (end) users (e.g. recommended actions or decisions to take, presenting of options)?
  - Could the AI system affect human autonomy by interfering with the (end) user's decision-making process in an unintended way?
  - Did you consider whether the AI system should communicate to (end) users that a decision, content, advice or outcome is the result of an algorithmic decision?
  - In case of a chat bot or other conversational system, are the human end users made aware that they are interacting with a non-human agent?

Human agency:

- ✓ Is the AI system implemented in work and labour process? If so, did you consider the task allocation between the AI system and humans for meaningful interactions and appropriate human oversight and control?
  - Does the AI system enhance or augment human capabilities?
  - Did you take safeguards to prevent overconfidence in or overreliance on the AI system for work processes?

Human oversight: