MAI4CAREU Master programmes in Artificial Intelligence 4 Careers in Europe

ETHICS GUIDELINES FOR TRUSTWORTHY AI By the High-Level Expert Group on Artificial Intelligence

Giovanni Sartor

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

The document

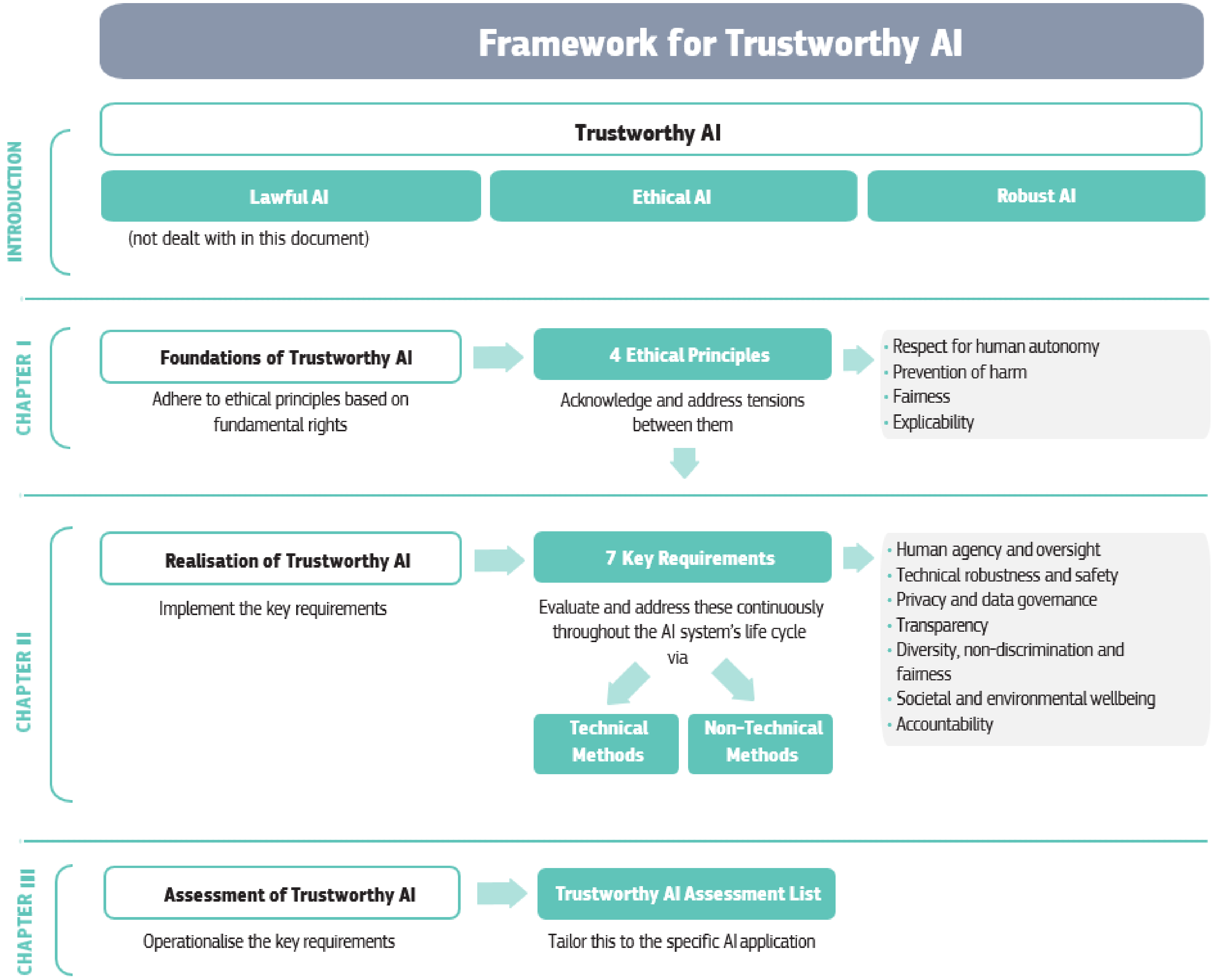
- Prepared by the High-Level Expert Group on Artificial Intelligence set up by the European Commission in June 2018.
- made public on 8 April 2019.
- available online (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).

- It is a good example of the many documents on ethics of AI published so far

The idea of trustworthy AI

- AI should be
    - Lawful, complying with all applicable laws and regulations
    - Ethical, ensuring adherence to ethical principles and values
    - Robust, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm

- This requirements should be met throughout the system's entire life cycle
- Question. Can you think of examples of unlawful, unethical or non-robust uses of AI?



Framework for Trustworthy AI

INTRODUCTION

Trustworthy AI

Lawful AI

(not dealt with in this document)

Ethical AI

Robust AI

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

Chapter 1: Ethical principles

- Develop, deploy and use AI systems in a way that adheres to ethical principle :
    - respect for human autonomy,
    - prevention of harm,
    - fairness and
    - explicability.
- Acknowledge and address the potential tensions between these principles.
- Pay particular attention to
    - situations involving more vulnerable groups such as children, persons with disabilities and others that have historically been disadvantaged or are at risk of exclusion, and
    - situations which are characterised by asymmetries of power or information, such as between employers and workers, or between businesses and consumers.
- Acknowledge that, while bringing substantial benefits to individuals and society,
    - AI systems also pose certain risks and may have a negative impact including impacts which may be difficult to anticipate, identify or measure (e.g. on democracy, the rule of law and distributive justice, or on the human mind itself.)
    - Adopt adequate measures to mitigate these risks when appropriate, and proportionately to the magnitude of the risk.

Chapter II: guidance of realisation trustworthy AI

- Ensure that the development, deployment and use of AI systems meets the 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 and
  - (7) accountability.
- Consider technical and non-technical methods to ensure the implementation of those requirements.

Chapter II: guidance of realisation trustworthy AI (continues)

- Foster research and innovation
  - to help assess AI systems and to further the achievement of the requirements; disseminate results and open questions to the wider public, and systematically train a new generation of experts in AI ethics.
- Communicate, in a clear and proactive manner, information to stakeholders about the AI system’s capabilities and limitations,
  - enabling realistic expectation setting, and about the manner in which the requirements are implemented. Be transparent about the fact that they are dealing with an AI system.
- Facilitate the traceability and auditability of AI systems
  - , particularly in critical contexts or situations.
- Involve stakeholders throughout the AI system’s life cycle.
  - Foster training and education so that all stakeholders are aware of and trained in Trustworthy AI.
- Be mindful that there might be fundamental tensions between different principles and requirements.
  - Continuously identify, evaluate, document and communicate these trade-offs and their solutions.

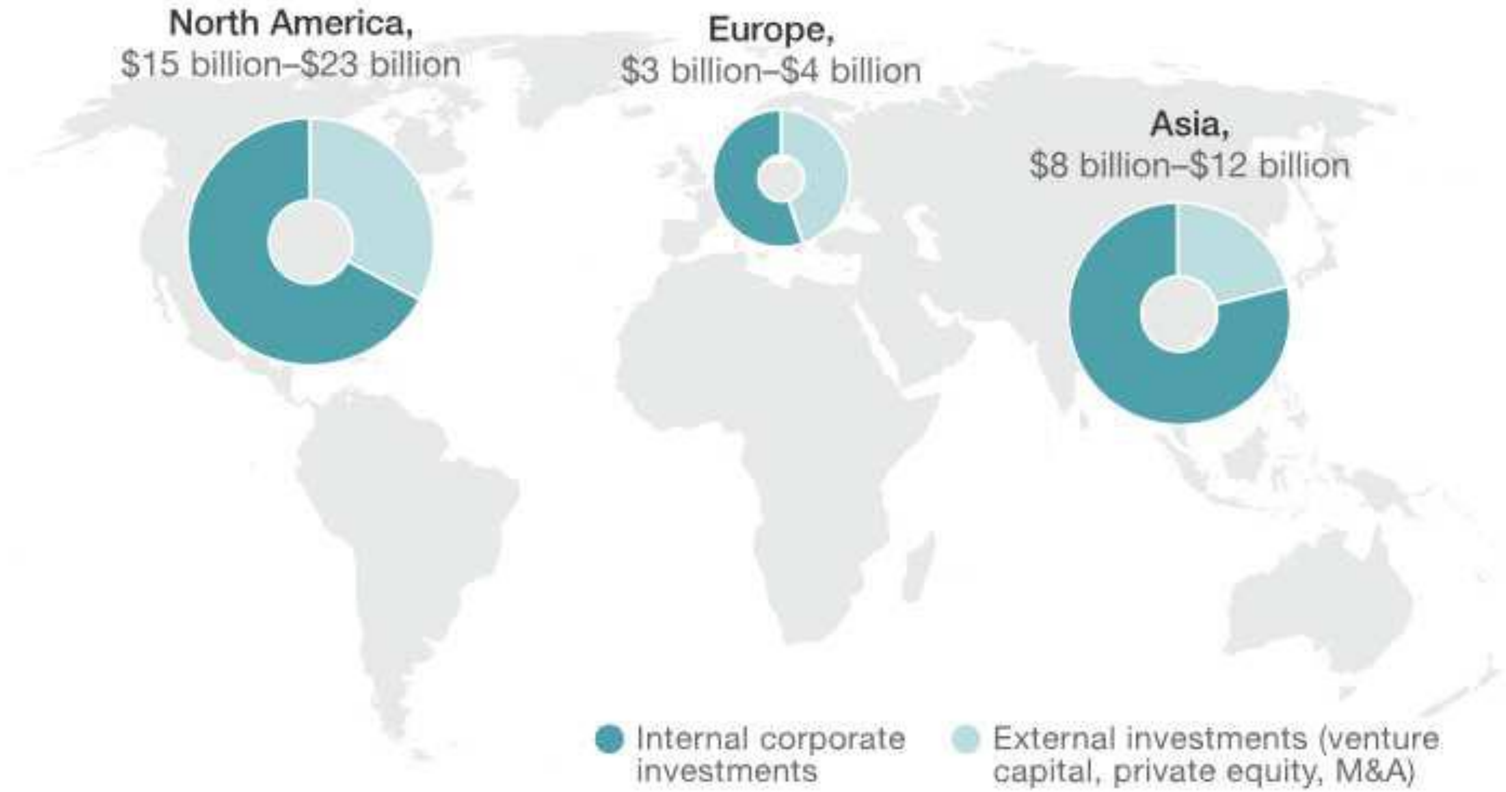
Chapter III: Trustworthy AI assessment

- Adopt a Trustworthy AI assessment list
  - when developing, deploying or using AI systems, and adapt it to the specific use case in which the system is being applied.
- Keep in mind that such an assessment list will never be exhaustive.
  - Ensuring Trustworthy AI is not about ticking boxes, but about continuously identifying and implementing requirements, evaluating solutions, ensuring improved outcomes throughout the AI system's lifecycle, and involving stakeholders in this.

The Commission’s approach to AI

- Communications 25 April 2018 and 7 December 2018 (COM(2018)237 and COM(2018)795). Three pillars:
  - (i) increasing public and private investments in AI to boost its uptake
  - (ii) preparing for socio-economic changes, and
  - (iii) ensuring an appropriate ethical and legal framework to strengthen European values.
- https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-237-F1-EN-MAIN-PART-1.PDF
- https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-795-F1-EN-MAIN-PART-1.PDF

Issue: are we really able to match US and China?



- https://ec.europa.eu/growth/tools-databases/dem/monitor/content/usa-china-eu-plans-ai-where-do-we-stand

Human-centric AI

- commitment to the use of AI in the service of humanity and the common good, with the goal of improving human welfare and freedom.
- Maximise the benefits of AI systems while at the same time preventing and minimising their risks.

Ethics vs law

- Ethics: norms indicating what should be done, with regard to all interests at stake
  - Positive ethics: norms shared in a society (possibly including ideas of social hierarchy, gender roles, etc.)
  - Critical ethics: norms that are viewed as most appropriate, or rational
- Law: norms that adopted through institutional processes and coercively enforced.

The Guidelines for Trustworthy AI as a (critical) ethics?

- Stakeholders committed towards achieving Trustworthy AI can *voluntarily* opt to use these Guidelines as a method to operationalise their commitment,
- The guidelines are addressed to all AI stakeholders designing, developing, deploying, implementing, using or being affected by AI,
  - including but not limited to companies, organisations, researchers, public services, government agencies, institutions, civil society organisations, individuals, workers and consumers.
- “Nothing in this document shall create legal rights nor impose legal obligations towards third parties. We however recall that it is the duty of any natural or legal person to comply with laws – whether applicable today or adopted in the future according to the development of AI.”

- What is the role of ethics, relatively to law in the AI domain?

AI should be lawful

- It should comply with
  - EU primary law (the Treaties of the European Union and its Charter of Fundamental Rights),
  - EU secondary law (regulations and directives, such as the General Data Protection Regulation, the Product Liability Directive, the Regulation on the Free Flow of Non-Personal Data, anti-discrimination Directives, consumer law and Safety and Health at Work Directives),
  - UN Human Rights treaties and the Council of Europe conventions (such as the European Convention on Human Rights),
  - Laws of EU Member State laws (Italian law).
- Laws can be horizontal of domain-specific rules (e.g., on medical devices)
- Issue: Can you think of a horizontal law covering all AI applications?

Foundations of trustworthy AI

- AI ethics is a sub-field of applied ethics,
  - focusing on the ethical issues raised by the development, deployment and use of AI.
  - Its central concern is to identify how AI can advance or raise concerns to the good life of individuals, whether in terms of quality of life, or human autonomy and freedom necessary for a democratic society.

SUSTAINABLE DEVELOPMENT GOALS
17 GOALS TO TRANSFORM OUR WORLD
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
SUSTAINABLE DEVELOPMENT GOALS

Foundation: (Ethical) fundamental rights

- Respect for human dignity. Human dignity encompasses the idea that every human being possesses an “intrinsic worth”
- Freedom of the individual. Human beings should remain free to make life decisions for themselves: including (among other rights) protection of the freedom to conduct a business, the freedom of the arts and science, freedom of expression, the right to private life and privacy, and freedom of assembly and association.

Foundation: (Ethical) fundamental rights

- Respect for democracy, justice and the rule of law. AI systems must not undermine democratic processes, human deliberation or democratic voting systems, due process and equality before the law
- Equality, non-discrimination and solidarity - including the rights of persons at risk of exclusion. In an AI context, equality entails that the system's operations cannot generate unfairly biased outputs. (GS: we need to understand what this means)
- Other citizens' rights the right to vote, the right to good administration or access to public documents, and the right to petition the administration

Ethical principles (based on human rights)

- (i) Respect for human autonomy
- (ii) Prevention of harm
- (iii) Fairness
- (iv) Explicability

Respect for human autonomy

- Humans interacting with AI systems must be able to keep full and effective self-determination over themselves, and be able to partake in the democratic process.
  - AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans.
  - they should be designed to augment, complement and empower human cognitive, social and cultural skills.
  - The allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity for human choice. T
  - This means securing human oversight over work processes in AI systems, supporting humans in the working environment, and aiming for the creation of meaningful work.

The principle of prevention of harm

- AI systems should neither cause nor exacerbate harm or otherwise adversely affect human beings.
    - This entails the protection of human dignity as well as mental and physical integrity.
    - AI systems and the environments in which they operate must be safe and secure.

The principle of fairness

- Substantive dimension
  - ensuring equal and just distribution of both benefits and costs, and
  - ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation.
  - Promoting equal opportunity in terms of access to education, goods, services and technology.
  - Never leading to people being deceived or unjustifiably impaired in their freedom of choice.
  - AI practitioners should respect the principle of proportionality between means and ends, and consider carefully how to balance competing interests and objectives
- Procedural dimension.
  - ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them
    - In order to do so, the entity accountable for the decision must be identifiable, and the decision-making processes should be explicable.

The principle of explicability

- To ensure contestability
  - processes need to be transparent,
  - the capabilities and purpose of AI systems openly communicated, and
  - decisions – to the extent possible – explainable to those directly and indirectly affected.

- An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible.
  - other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities) may be required, provided that the system as a whole respects fundamental rights.
  - thee degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate.3

Tensions between the principles

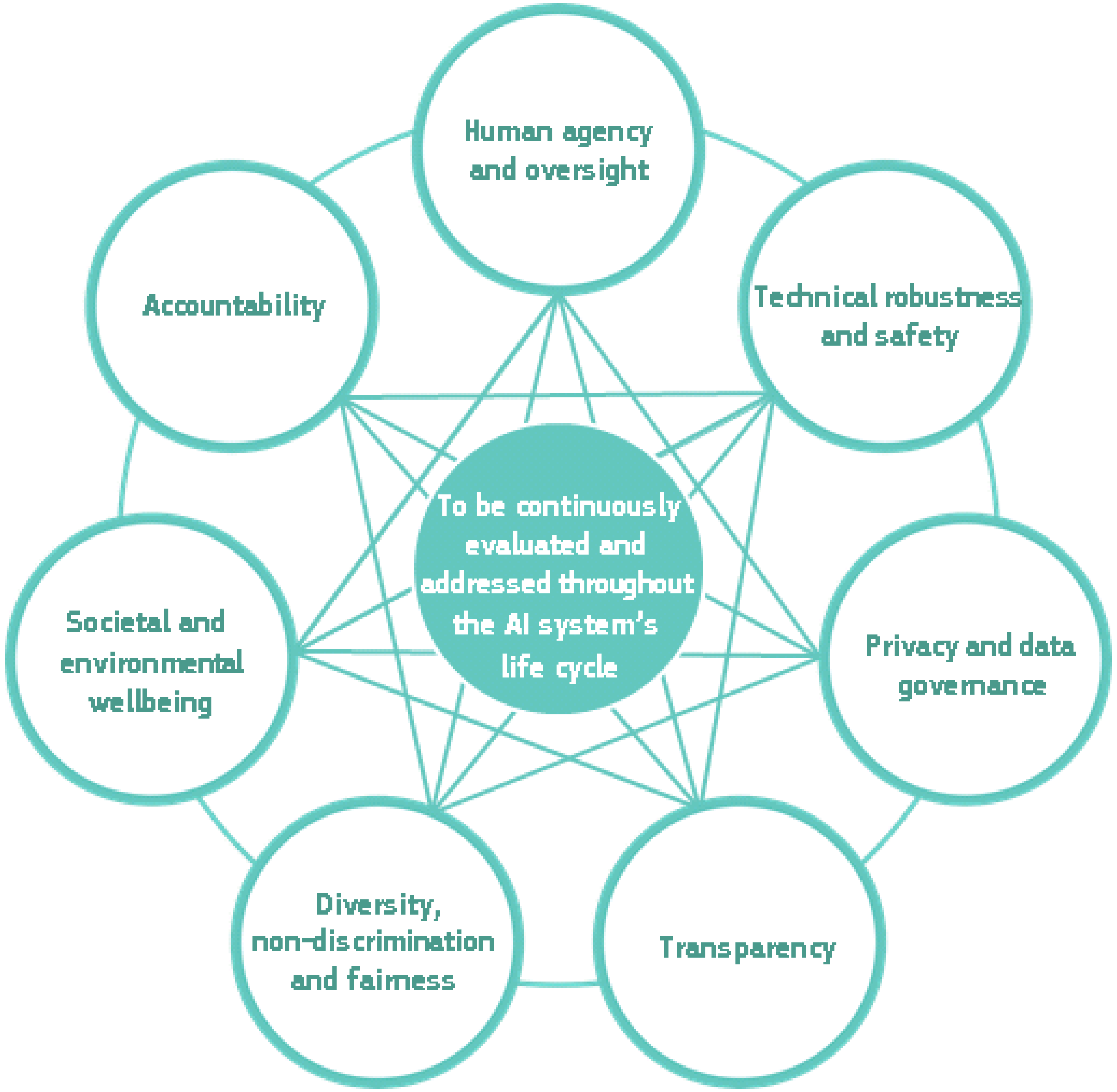
- Methods of accountable deliberation to deal with such tensions should be established.
  - Conflicts between prevention of harm and human autonomy
  - Also between welfare and security?

Requirements of Trustworthy AI

- 1. Human agency and oversight
  - Including fundamental rights, human agency and human oversight
- 2 Technical robustness and safety
  - Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility
- 3 Privacy and data governance
  - Including respect for privacy, quality and integrity of data, and access to data
- 4 Transparency
  - Including traceability, explainability and communication

Requirements of Trustworthy AI (continues)

- 5 Diversity, non-discrimination and fairness
  - Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation
- 6 Societal and environmental wellbeing
  - Including sustainability and environmental friendliness, social impact, society and democracy
- 7 Accountability
  - Including auditability, minimisation and reporting of negative impact, trade-offs and redress.



Human agency and oversight
Technical robustness and safety
Privacy and data governance
Transparency
Diversity, non-discrimination and fairness
Societal and environmental wellbeing
Accountability
To be continuously evaluated and addressed throughout the AI system's life cycle

Human agency and oversight

- AI systems should support human autonomy and decision-making. Therefore they should support
  - Fundamental rights
    - Human rights assessment
  - Human agency.
    - Users should be able to make informed autonomous decisions regarding AI systems.
  - Human oversight.
    - Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects (human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach + public controls)
  - Technical robustness and safety
    - AI systems be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimising unintentional and unexpected harm, and preventing unacceptable harm.

Human agency and oversight (continues)

- Resilience to attack and security
  - AI systems, should be protected against vulnerabilities that can allow them to be exploited by adversaries
- Fallback plan and general safety
  - AI systems should have safeguards that enable a fallback plan in case of problems
- Accuracy
  - AI systems should have the ability to make correct judgements, for example to correctly classify information into the proper categories, or its ability to make correct predictions, recommendations, or decisions based on data or models.
- Reliability and Reproducibility .
  - The results of AI systems should be reproducible, as well as reliable.

Privacy and data governance

- Prevention of harm necessitates privacy and data governance:
  - Privacy and data protection.
    - AI systems must guarantee privacy and data protection throughout a system’s entire lifecycle.
  - Quality and integrity of data
    - The data used to train a systems should not contain socially constructed biases, inaccuracies, errors and mistakes, malicious data should not be added
  - Access to data
    - Data protocols governing data access should be put in place.

Transparency

- This requirement is closely linked with the principle of explicability
  - Traceability.
    - The data sets and the processes that yield the AI system’s decision, should be documented
  - Explainability.
    - The technical processes of an AI system and the related human decisions should be explainable
  - Communication.
    - Humans have the right to be informed that they are interacting with an AI system.

Diversity, non-discrimination and fairness

- We must enable inclusion and diversity throughout the entire AI system's life cycle
  - Avoidance of unfair bias
    - Prevent unintended (in)direct prejudice and discrimination against certain groups or people, potentially exacerbating prejudice and marginalisation, due to data or algorithms
  - Accessibility and universal design.
    - AI systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, abilities or characteristics
  - Stakeholder Participation.
    - Open discussion and the involvement of social partners and stakeholders, including the general public
  - Diversity and inclusive design teams
    - the teams that design, develop, test and maintain, deploy and procure these systems reflect the diversity of users and of society in general

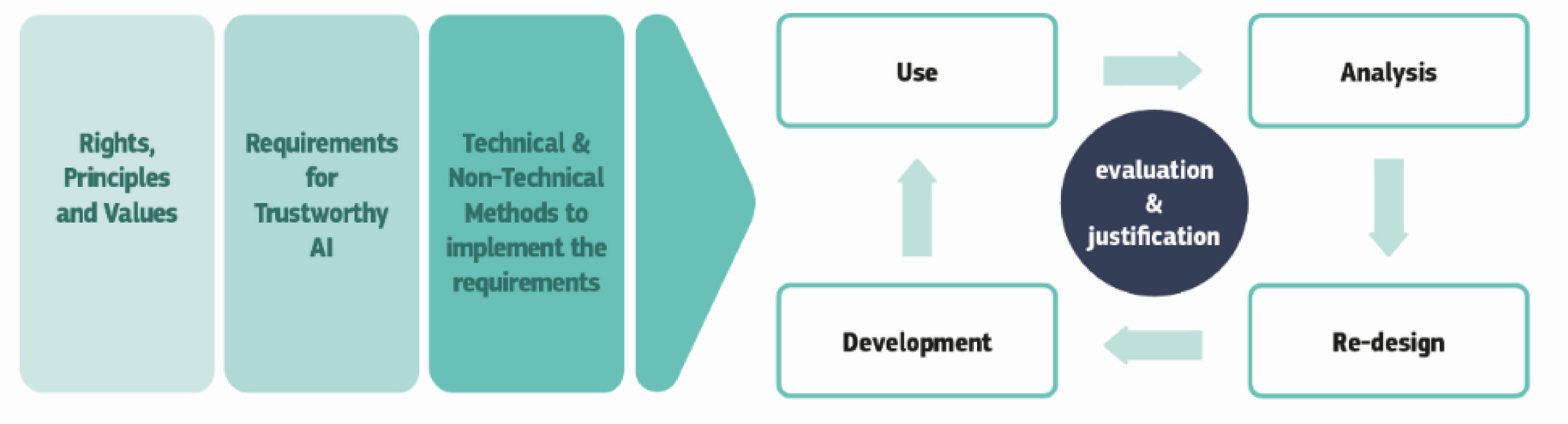
Societal and environmental well-being

- The broader society, other sentient beings and the environment should be also considered as stakeholders throughout the AI system’s life cycle.
    - Sustainable and environmentally friendly AI
        - Measures securing the environmental friendliness of AI systems’ entire supply chain should be encouraged.
    - Social impact.
        - The effects of these systems on individuals, groups and society must therefore be carefully monitored and considered.
    - Society and Democracy.
        - Take into account AI’s effect on institutions, democracy and society at large

Accountability

- Ensure responsibility and accountability for AI systems and their outcomes
    - Auditability
        - Enablement of the assessment of algorithms, data and design processes
    - Minimisation and reporting of negative impacts
        - The ability to report on actions or decisions that contribute to a certain system outcome, and to respond to the consequences of such an outcome, must be ensured.
    - Trade-offs
        - Trade-offs should be addressed in a rational and methodological manner within the state of the art
    - Redress.
        - Accessible mechanisms should be foreseen that ensure adequate redress

Technical and non-technical methods to realise Trustworthy AI


Rights, Principles and Values
Requirements for Trustworthy AI
Technical & Non-Technical Methods to implement the requirements
Use
Analysis
evaluation & justification
Re-design
Development

Questions and suggestions

- Questions
  - Has the Trustworthy AI document provided you with useful indications?
  - Do you think that they are concretely applicable?
  - Are ethical guidelines that are not legally binding really useful?
  - Any specific criticism?

- Suggestion
  - Read all the document!
  - Read also

Thanks for your attention

- Giovanni.sartor@Unibo.it

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423