13.05.19

Kiel.AI

The European Commission's AI Ethics Guidelines

OPENCAMPUS.sh

MANDATE OF THE HIGH LEVEL EXPERT GROUP

The European Commission established the High-Level Expert Group on Artificial Intelligence (AI HLEG), with drafting two documents:

- ❑ AI Ethics Guidelines (published on April 8, 2019)
- ❑ Policy and Investment Recommendations (June?, 2019)

CONTENT

SPIRIT

"We […] want producers of AI systems to get a competitive advantage by embedding Trustworthy AI in their products and services." (p.4)

OPENCAMPUS.sh

[CHAPTER 1]
TRUSTWORTHY AI

Trustworthy AI has **three components,** which should be met throughout the system's entire life cycle, it should be

(1) **lawful**,

(2) **ethical**, and

(3) **robust**.

ETHICAL IMPERATIVES

Based on the EU Charter of Fundamental Fights, four ethical imperatives are formulated:

(1) Respect for human autonomy

(2) Prevention of harm

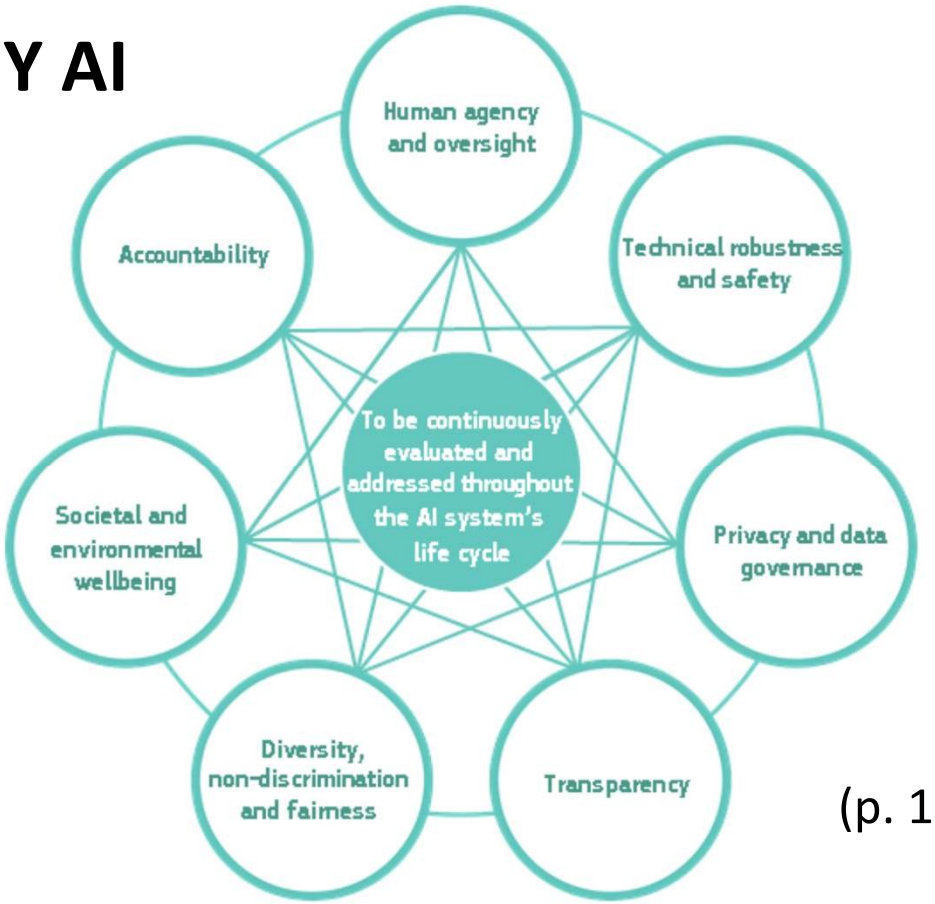
(3) Fairness

(4) Explicability

FAIRNESS

“The procedural dimension of fairness entails the 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.“ (p. 13)

REQUIREMENTS OF TRUSTWORTHY AI



(p. 15)

OPENCAMPUS.sh

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

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

Users “should be given the knowledge and tools to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system.” (p. 16)

TRANSPARENCY – COMMUNICATION

"Beyond this, the AI system's capabilities and limitations should be communicated to AI practitioners or end-users in a manner appropriate to the use case at hand. This could encompass communication of the AI system's level of accuracy, as well as its limitations." (p. 18)

TRANSPARENCY – EXPLAINABILITY

"Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings. Moreover, trade-offs might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability). Whenever an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned (e.g. layperson, regulator or researcher). In addition, explanations of the degree to which an AI system influences and shapes the organisational decision-making process, design choices of the system, and the rationale for deploying it, should be available (hence ensuring business model transparency)." (p. 18)

AVOIDANCE OF UNFAIR BIAS

”hiring from diverse backgrounds, cultures and disciplines can ensure diversity of opinions and should be encouraged.” (p. 18)

OPENCAMPUS.sh

ACCOUNTABILITY – AUDITABILITY

”AI systems should be able to be independently audited .” (p. 20)

OPENCAMPUS.sh

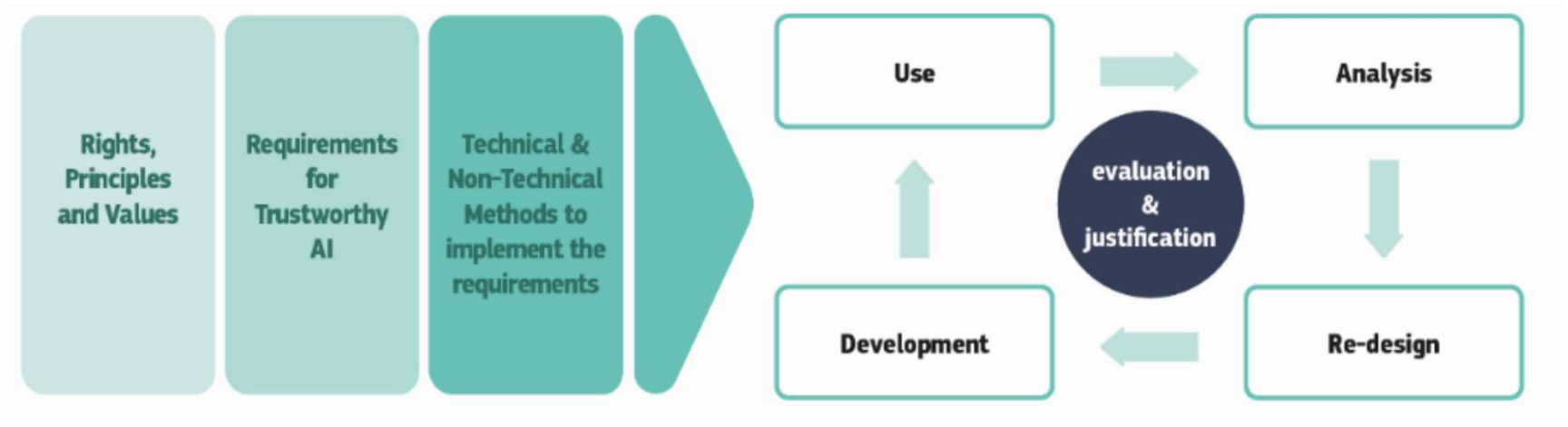
ACCOUNTABILITY – MINIMISATION AND REPORTING OF NEGATIVE IMPACTS

”Due protection must be available for whistle-blowers, NGOs, trade unions or other entities when reporting legitimate concerns about an AI system.” (p. 20)

ACCOUNTABILITY – TRADE OFFS

"the presence of an internal and/or external ethical (and sector specific) expert or board might be useful to highlight areas of potential conflict and suggest ways in which that conflict might best be resolved." (p. 20)

[CHAPTER 2]
REALIZATION OF TRUSTWORTHY AI



(p. 20)

TECHNICAL METHODS TO IMPLEMENT THE REQUIREMENTS

- ❑ Architectures for Trustworthy AI
- ❑ Ethics and rule of law by design (X-by-design)
- ❑ Explanation methods
- ❑ Testing and validating
- ❑ Quality of Service Indicators

NON-TECHNICAL METHODS TO IMPLEMENT THE REQUIREMENTS

- ❑ Regulation
- ❑ Codes of conduct
- ❑ Standardisation
- ❑ Certification
- ❑ Accountability via governance frameworks
- ❑ Education and awareness to foster an ethical mind-set
- ❑ Stakeholder participation and social dialogue
- ❑ Diversity and inclusive design teams

[CHAPTER 3]
OPERATIONALIZATION OF TRUSTWORTHY AI

- ❑ Governance
- ❑ Using the Trustworthy AI assessment list
- ❑ Relation to existing law and processes

GOVERNANCE

- ❑ Management and Board
- ❑ Compliance/Legal department/Corporate responsibility department
- ❑ Product and Service Development or equivalent
- ❑ Quality Assurance
- ❑ HR
- ❑ Procurement
- ❑ Day-to-day Operations

OPENCAMPUS.sh

AI ASSESSMENT LIST

1 Human agency and oversight
Including fundamental rights (2), human agency (1), and human oversight (2)

2 Technical robustness and safety
Including resilience to attack and security (4), fall back plan and general safety (4), accuracy (4), reliability and reproducibility (1)

3 Privacy and data governance
Including respect for privacy (6), quality and integrity of data (4), and access to data (1)

4 Transparency
Including traceability (4), explainability (3), and communication (4)

5 Diversity, non-discrimination and fairness
Including the avoidance of unfair bias (4), accessibility and universal design (2), and stakeholder participation (2)

6 Societal and environmental wellbeing
Including sustainability and environmental friendliness (2), social impact (2), society and democracy (1)

7 Accountability
Including auditability (2), minimisation and reporting of negative impact (4), trade-offs (2), and redress (2).

EXAMPLES OF TRUSTWORTHY AI'S OPPORTUNITIES

- Climate action and sustainable infrastructure
- Health and well-being
- Quality education and digital transformation

OPENCAMPUS.sh

EXAMPLES OF CRITICAL CONCERNS RAISED BY AI

- ❑ Identifying and tracking individuals with AI
- ❑ Covert AI systems
- ❑ AI enabled citizen scoring in violation of fundamental rights
- ❑ Lethal autonomous weapon systems (LAWS)
- ❑ Potential longer-term concerns

CRITIQUE

“To do this, 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.“ (p. 4; also, cf. p. 10)

OTHER AI ETHICS INITIATIVES

- Bertelsmann Foundation and iRights.Lab: **Algo.Rules**
  (Rules for the Design of Algorithmic Systems: https://algorules.org/en/home/)

- KI Bundesverband e.V.: **KI Gütesiegel** [AI Quality Seal]
  (Signing of a voluntary declaration: https://ki-verband.de/wp-content/uploads/2019/02/KIBV_Guetesiegel.pdf)

OPENCAMPUS.sh

INTERESTING READS

- Newsletter Algorithmenethik: https://algorithmenethik.de/
- Yoshua Bengio and why “the dangers of abuse are very real”: https://www.nature.com/articles/d41586-019-00505-2
- Sueddeutsche Zeitung, “Das Silicon Valley kauft sich ein Gewissen”: https://www.sueddeutsche.de/digital/silicon-valley-ethik-kommissionen-feigenblatt-1.4399509

OPENCAMPUS.sh