Ethical Implication of AI: Assessing Trustworthy AI in Practice
Introduction

Course Coordinators:

Prof. Roberto V. Zicari, SNU
Prof. Heejin Kim, SNU

Teaching Assistants:

Sarah Hyojin, SNU **and Dennis Vetter**, Goethe University Frankfurt

SNU Data Science, *Seoul National University*
September 15, 2022

Artificial Intelligence (AI)

"Everything we love about civilization is a product of intelligence, so amplifying our human intelligence with artificial intelligence has the potential of helping civilization flourish like never before – as long as we manage to keep the technology ***beneficial****."*

Max Tegmark, President of the Future of Life Institute

Benefits vs. **Risks**

How do we "know" what are the **Benefits** vs. **Risks** of using an AI system?

Trustworthy AI

- Applications based on Machine Learning and/or Deep Learning carry specific (mostly unintentional) **risks** that are considered within **AI ethics**.
- As a consequence, the quest **for trustworthy AI** has become a central issue for governance and technology impact assessment efforts, and has increased in the last four years, with focus on identifying both ethical and legal principles.

What is Ethics?

- A branch of philosophy
  - ➨An appraisal of what is right (good) and wrong (bad)
  - ➨An assessment of human actions

Assignment: Watch

Source: **The Ethics of Artificial Intelligence (AI) (Dr. Emmanuel Goffi)**

YouTube Video:
https://www.youtube.com/watch?v=uuh7spVdf0c&t=7s

Trustworthy AI

- As AI capabilities have grown exponentially, it has become increasingly difficult to determine whether their model outputs or system behaviors protect the rights and interests of an ever-wider group of stakeholders –let alone evaluate them as ethical or legal, or meeting goals of improving human welfare and freedom.

Trustworthy AI

- For example, **what if decisions made using an AI-driven algorithm benefit some socially salient groups more than others?**

- And **what if we fail to identify and prevent these inequalities because we cannot explain how decisions were derived?**

How to assess Trustworthy AI systems in practice

- Moreover, we also need to consider how the adoption of these new algorithms and the lack of knowledge and control over their inner workings may impact those in charge of making decisions.

- This course will help students to assess Trustworthy AI systems in practice by using the **Z-Inspection® process.**

How to assess Trustworthy AI systems in practice

- The Z-Inspection® process is the result of 4.5 years of applied research of the Z-Inspection® initiative, a network of high class world experts lead by Prof. Roberto V. Zicari.
- Z-Inspection® is a holistic process used to evaluate the trustworthiness of AI-based technologies at different stages of the AI lifecycle. It focuses, in particular, on the identification and discussion of **ethical issues and tensions** through the elaboration of **socio-technical scenarios**. It uses the general European Union's High-Level Expert Group's (EU HLEG) guidelines for trustworthy AI.

Course Description

i) Learn Basic Principles

- Students will learn the **ethical implications of the use of Artificial Intelligence** (AI). What are the consequences for society? For human beings / individuals? Does AI serve human kind?
- Discussion and debate of ethical issues is an essential part of professional development.

Course Description

ii) Apply Principles in Practice

- Students will learn how to assess an AI system in practice, identifying possible risks and “issues”.
- Students will work in small groups of three assessing real AI systems.

Syllabus

1. **Fundamental rights as moral and legal entitlements**

Fundamental rights encompass rights such as human dignity and non-discrimination, as well as rights in relation to data protection and privacy, to name just some examples.

2. **From fundamental rights to ethical principles**

Syllabus

3. **EU guidelines for Trustworthy AI**

3.1. **Trustworthy AI Ethical principles:**

– *Respect for human autonomy*

– *Prevention of harm*

– *Fairness*

– *Explicability*

3.2 **Ethical Tensions and Trade offs**

Syllabus

3.3. **Seven Requirements (+ sub-requirements)**
for Trustworthy AI

– **Human agency and over**sight. ***Including fundamental rights****, human agency and human oversight*

– **Technical robustness and safety**. *Including resilience to attack and security, fall back plan and general safety*

– **Privacy and data governance**. *Including respect for privacy, quality and integrity of data, and access to data*

Syllabus

– **Transparency**. *Including traceability, explainability and communication*

– **Diversity, non-discrimination and fairness**. *Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation*

– **Societal and environmental wellbeing**. *Including sustainability and environmental friendliness, social impact, society and democracy*

– **Accountability**. *Including auditability, minimisation and reporting of negative impact, trade-offs and redress.*

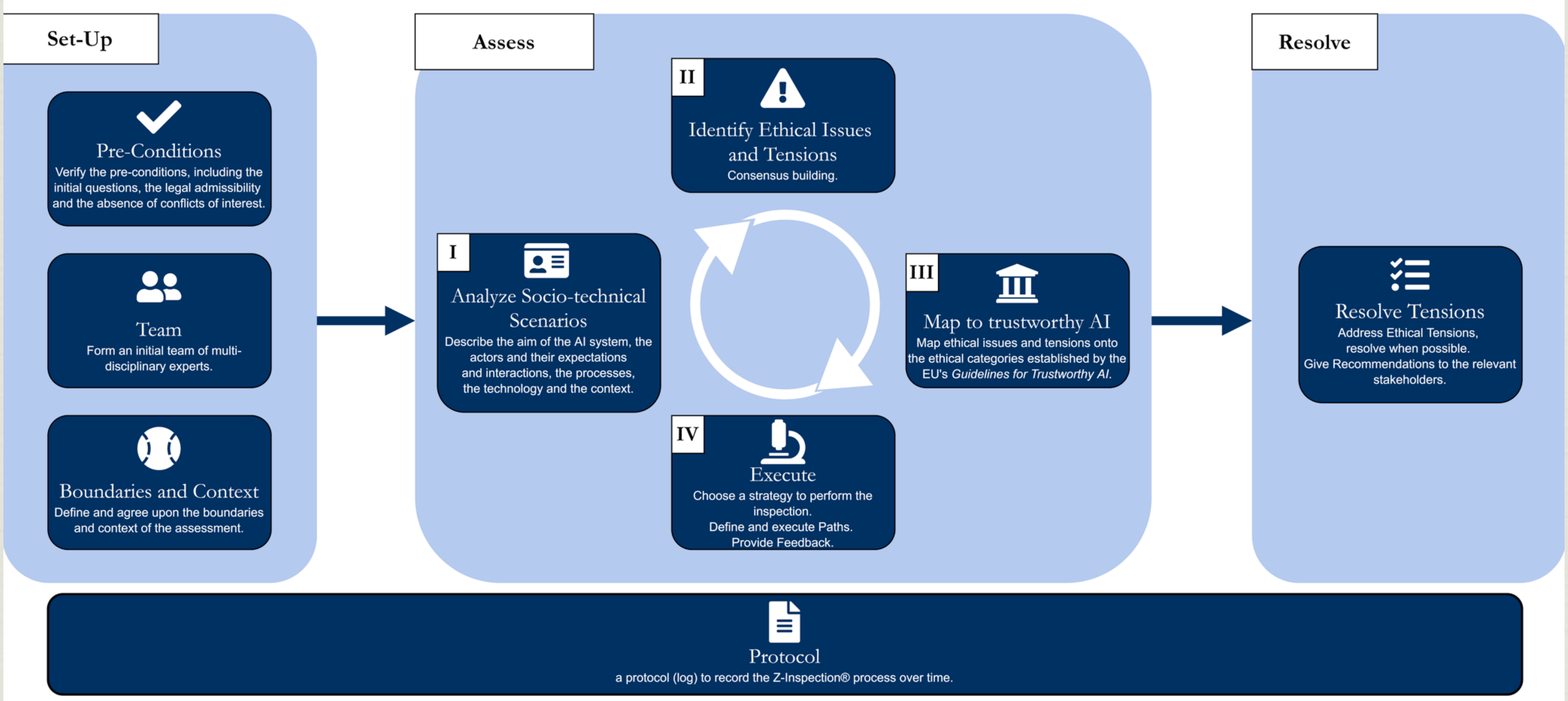
Syllabus

4. **Assessing Trustworthy AI in Practice.**

4. 1 **The Z-Inspection® process in detail**

- – *Set Up Phase*
- – *Assess Phase*
- – *Resolution Phase*

Z-inspection® Process in a Nutshell



Set-Up

Pre-Conditions
Verify the pre-conditions, including the initial questions, the legal admissibility and the absence of conflicts of interest.

Team
Form an initial team of multi-disciplinary experts.

Boundaries and Context
Define and agree upon the boundaries and context of the assessment.

Assess

I
Analyze Socio-technical Scenarios
Describe the aim of the AI system, the actors and their expectations and interactions, the processes, the technology and the context.

II
Identify Ethical Issues and Tensions
Consensus building.

III
Map to trustworthy AI
Map ethical issues and tensions onto the ethical categories established by the EU's *Guidelines for Trustworthy AI.*

IV
Execute
Choose a strategy to perform the inspection.
Define and execute Paths.
Provide Feedback.

Resolve

Resolve Tensions
Address Ethical Tensions, resolve when possible.
Give Recommendations to the relevant stakeholders.

Protocol
a protocol (log) to record the Z-Inspection® process over time.

Syllabus

4.2 . **Additional Frameworks**

– The Claim, Arguments and Evidence Framework (CAE)

– The fundamental rights and algorithm impact assessment ('FRAIA')

4.3 Tools

– The ALTAI Web tool

Course Resources

- EU Trustworthy AI Guidelines
- EU Draft Proposed AI Law
- Socio-Technical Scenarios
- Catalog of Ethical tensions
- Claims Arguments and Evidence (CAE)
- Fundamental Human Rights
- The Z-Inspection® process

http://z-inspection.org/ethical-implication-of-ai-assessing-trustworthy-ai-in-practice/

Additional Resources

AI and Ethics: Reports/Papers classified by topics

http://z-inspection.org/educational-resources-on-ethics-and-ai/

Course Modalities and Schedule

- Remote Zoom video call, and in presence sessions
- **65%** of lecture remote, **35%** present
- Class schedule**: Every Tuesday and Thursday, from 5pm -6.15pm** Korean time (KST) (10.00am – 11.15 am CET)
- Duration of a class: 1hour and 15 minutes. The course is for **3 Credits**.
- Course starts on September 1, 2022 and ends on December 14, 2022
  NOTE: the first lesson will be on Thursday, **September 15.**

Course Web site

http://z-inspection.org/ethical-implication-of-ai-assessing-trustworthy-ai-in-practice/

Working Groups

- Students will work in small groups and will evaluate the trustworthiness of real AI systems in various domains, **e.g. healthcare, government and public administrations, justice, etc.**
- **Sept 22 and Sept 27 :** Get to know meetings at 18.30 Bldg. 942, **Room #302 (**SNU Graduate School of Data Science)

Assignments

Each week (see web site):

Attend lectures of the week;

Read papers of the week;

Choose a **AI product/solution→ BY SEPTEMBER 29**

Mid Term and Final Report: Assess the AI system for "trustworthiness"

trustworthy AI, based on the EU "Framework for Trustworthy AI", using the Z-inspection® process.

Z-inspection® is a registered trademark.

The content of this work is open access distributed under the terms and conditions of the Creative Commons (**Attribution-NonCommercial-ShareAlike CC BY-NC-SA)** license https://creativecommons.org/licenses/by-nc-sa/4.0/)

Assess AI systems for "trustworthiness"

To Assess AI systems students will learn how to apply:

- **The EU Ethics Guidelines for Trustworthy AI,**
- **The Z-inspection® process,**
- **The fundamental rights and algorithm impact assessment ('FRAIA'),**
- **The Claim, Arguments and Evidence Framework (CAE),**
- **The ALTAI Web tool.**

Mid-Term Report

- The goal of the mid-term report is to select an **AI system** (i.e. an AI-product and or an AI-based service) and start with the evaluation process.

- In teams of **three students**:
  Choose a AI data-driven product/solution and assess for "trustworthiness.

Mid Term Report

- Define the **Boundaries** and **Context** of the assessment**.**
- Create and Analyze **Socio-technical scenarios**;
- Apply **The fundamental rights and algorithm impact assessment ('FRAIA')**
- **Identify Claims**, provide **Arguments and Evidence**
- **Identify Ethical Issues and Tensions**;

Mid-Term Report

By collecting relevant resources, **socio-technical scenarios** should be created and analyzed by the team of students.

Create and Analyze Socio-technical scenarios

- **Aim of the system**
  Goal of the system, context, why it is used.
- **Actors** (*primary:* directly involved with the use of the AI system, and *secondary* and *tertiary* only indirectly involved with the use of the AI system)
  Who designed and implemented the system?
- Who has authorized the deployment of the system?
- Who is currently using the system?
- Who are the end users for this system?
- Who is directly influenced by decisions made by the system?
- Who is indirectly influenced by decisions made by the system?
- Who is responsible for this system?

Create and Analyze Socio-technical scenarios

- **Actors' Expectation and Motivation**
  Why would the different groups of actors want the system? What are their expectations towards the system behavior? What benefits are they expecting from using the system?
- **Actors' Concerns and Worries**
  What problems / challenges can the actors foresee? Do they have concerns regarding the use of the system? What risks are they concerned about with the system? Are there any conflicts?

Create and Analyze Socio-technical scenarios

- **Context where the AI system is used**
  What additional context information about the situation where the AI system is used? (e.g. urgency, budget constraints, for profit, academic, conflicts, environmental). What are potential future usage of the AI system?

Create and Analyze Socio-technical scenarios

- **Interaction with the AI system**
  What is the intended interaction between the system and its users? If and how the 'human in control' aspect is envisaged? Why is it like this?
- **AI Technology used**
  Technical description of the AI system. An important part of considering AI trustworthiness is that it is robust and if the technical description is not clear, this cannot be assessed.

Create and Analyze Socio-technical scenarios

- **Clinical studies /Field tests**
  Was the system's performance validated in (clinical/field tests) studies? What were the results of these studies? Are the results openly available?

- **Intellectual Property**
  What parts of the AI system are open access (if any)? What IP regulations need to be considered when assessing / disseminating the system? Does it contain confidential information that must not be published? What is and how to handle the IP of the AI and of the part of the system to be examined. Identify possible restrictions to the Inspection process, in this case, assess the

Create and Analyze Socio-technical scenarios

- **Legal framework**
  What is the legal framework for use of the system? What special regulations apply? What are the data protection issues?", "Was the data aspect compliant with the GDPR?
- **Ethics oversight and/or approval**
  Has the AI system already undergone some kind of ethical assessment or other approval? If not - why not? If so, was this internal/external, volunteer/regulated, what was covered?

Mid Term Report

- **Mid-term report due → October 28, 2022**

- It must be **min. 3-max. 5 pages long**, including references and written with the Google docs presets (i.e. Normal Text: Arial, 11pt).

Mid-Term Report Requirements

Identify Ethical Issues and Tensions

We use the term ‘tension’ as defined in **[Whittlestone et al. 2019]** :

„tensions between the pursuit of different values in technological applications rather than an abstract tension between the values themselves.“

Use a Catalog of Examples of Ethical tensions

- Accuracy vs. Fairness
- Accuracy vs. Explainability
- Privacy vs. Transparency
- Quality of services vs. Privacy
- Personalisation vs. Solidarity
- Convenience vs. Dignity
- Efficiency vs. Safety and Sustainability
- Satisfaction of Preferences vs. Equality

Classify ethical tensions

[Whittlestone et al. 2019] :

- – **true dilemma** , i.e. “a conflict between two or more duties, obligations, or values, both of which an agent would ordinarily have reason to pursue but cannot”;
- – **dilemma in practice** , i.e. “the tension exists not inherently, but due to current technological capabilities and constraints, including the time and resources available for finding a solution”;
- – **false dilemma** , i.e. “situations where there exists a third set of options beyond having to choose between two important values”.

Final Report

Final Report
Continue the work of the Mid Term and create a Final Report

Total MAX 10 pages (including the min 3 pages and max. 5 pages of the Mid term report)
including references and written with the google docs presets (i.e. Normal Text: Arial, 11pt).

Delivered by →December 2, 2022

Final Report

- Continue The fundamental rights and algorithm impact assessment (‘FRAIA’),
- Continue Verify Claims, Support Arguments and Evidence;
- Refine the Ethical Issues and Tensions (when possible resolve them);
- **Map “Issues” to EU framework for trustworthy AI;**
- **Use the ALTAI web tool**
- **Give recommendations to relevant stakeholders.**

Mid-Term/Final Report Requirements

- The Mid Term/Final report contributes to your grade and every team member will receive the same grade.

- The report must be delivered as a **google doc** (we will create one document per team).

Remarks

NEVER EVER COPY AND PASTE text from the internet and other sources.

Two options:

1. You can describe what the source is saying using your words and quoting the source.

e.g As indicated in **[Roig and Vetter 2020]** the moon is flat. **Reference [Roig and Vetter 2020]** Why the Moon is Flat. Roig Gemma, Vetter Dennis Journal of Dreams, Issue No. 1, November 2020-

Remarks

2.You can quote what the source is saying using their words in "… "

e.g **[Roig and Vetter 2020]** has made a case that "the moon is completely flat and not round".**Reference [Roig and Vetter 2020]** Why the Moon is Flat. Roig Gemma, Vetter Dennis Journal of Dreams, Issue No. 1, November 2020

- -Make sure to READ the source when you use them to make sure you understand the context! In the example above, if you quote that "the moon is flat" without understanding that this was a dream and not a scientific evidence, you are using a quote in a WRONG way!

Assignment

- Watch this pre-recorded lesson

The Ethics of Artificial Intelligence (AI)

Dr. Emmanuel Goffi)

YouTube Video:
https://www.youtube.com/watch?v=uuh7spVdf0c&t=7s

Assignment

- Read this report:

EU Trustworthy AI Guidelines

- Ethics Guidelines for Trustworthy AI. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019. Link to .PDF

https://ai.bsa.org/wp-content/uploads/2019/09/AIHLEG_EthicsGuidelinesforTrustworthyAI-ENpdf.pdf

Resources

http://z-inspection.org

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

- In 2019 the High-Level Expert Group on Artificial Intelligence (AI HLEG), set up by the European Commission, published the Ethics Guidelines for Trustworthy Artificial Intelligence.
- The third chapter of those Guidelines contained an Assessment List to help assess whether the AI system that is being developed, deployed, procured or used, adheres to the **seven requirements of Trustworthy Artificial Intelligence (AI),** as specified in our Ethics Guidelines for Trustworthy AI

ALTAI WEB SITE

- This website contains the Assessment List for Trustworthy AI (ALTAI).
- ALTAI was developed by the High-Level Expert Group on Artificial Intelligence set up by the European Commission to help assess whether the AI system that is being developed, deployed, procured or used, complies with the seven requirements of Trustworthy AI, as specified in our Ethics Guidelines for Trustworthy AI.
- **https://altai.insight-centre.org**

How to complete ALTAI

- For each question ALTAI provides **guidance in the glossary** and by **referencing to the relevant parts of the Ethics Guidelines for Trustworthy AI** and examples in text boxes alongside the questions.

Upon completing ALTAI, the following will be generated:

- **A visualisation** of the self-assessed level of adherence of the AI system and it's use with the 7 requirements for Trustworthy AI. These results are based on your organisation's own assessment and are solely meant to help you identify the areas of improvement.
- **Recommendations** based on the answers to particular questions.

Register to the ALTAI Online Tool.

- You need to create an account to use the Assessment List for Trustworthy AI Online Tool. This allows you to save and edit ALTAIs.
- https://altai.insight-centre.org/Identity/Account/Register

ALTAI web tool

- The self-assessment results and the list of recommendations are confidential and available to you only.
- ALTAI has seven sections, corresponding to the 7 requirements for Trustworthy AI. You can move between the sections by clicking on the side menu.

Colour code
Questions with **white background** are aimed at (describing) the features of the AI system.

Answers to **blue questions** will contribute to recommendations.

Text in red allows you to self-assess your organisation's compliance with the respective requirement.

Fundamental Rights

- Fundamental rights encompass rights such as human dignity and non-discrimination, as well as rights in relation to data protection and privacy, to name just some examples.

Fundamental rights impact assessment

- **Prior to self-assessing an AI system with this Assessment List, a fundamental rights impact assessment (FRIA) should be performed.**
- A FRIA could include questions such as the following – drawing on specific articles in the Charter and the European Convention on Human Rights (ECHR) its protocols and the European Social Charter.
- **https://altai.insight-centre.org/Home/FundamentalRights**

The Fundamental Rights and Algorithm Impact Assessment (FRAIA)

- **Impact Assessment Fundamental Rights and Algorithms**
- The Fundamental Rights and Algorithm Impact Assessment (FRAIA) helps to map the risks to human rights in the use of algorithms and to take measures to address this. FRAIA is the English translation of Impact Assessment Mensenrechten en Algoritmes.

https://www.government.nl/documents/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms

Fundamental rights impact assessment

- **Does the AI system potentially negatively discriminate against people on the basis of any of the following grounds** (non-exhaustively):

sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation?

Fundamental rights impact assessment

- **Does the AI system potentially negatively discriminate against people on the basis of any of the following grounds**
- Have you put in place processes to test and monitor for potential discrimination (bias) during the development, deployment and use phase of the AI system?
- Have you put in place processes to address and rectify for potential discrimination (bias) in the AI system?

Fundamental rights impact assessment

- **Does the AI system respect the rights of the child, for example with respect to child protection and taking the child's best interests into account?**
- Have you put in place processes to test and monitor for potential harm to children during the development, deployment and use phase of the AI system?
- Have you put in place processes to address and rectify for potential harm to children by the AI system?

Fundamental rights impact assessment

- Does the AI system protect the right to privacy, including personal data relating to individuals in line with GDPR?

Fundamental rights impact assessment

- Have you put in place processes to assess in detail the need for a data protection impact assessment, including an assessment of the necessity and proportionality of the processing operations in relation to their purpose, with respect to the development, deployment and use phases of the AI system?
- Have you put in place measures envisaged to address the risks, including safeguards, security measures and mechanisms to ensure the protection of personal data with respect to the development, deployment and use phases of the AI system?
- See the section on Privacy and Data Governance in this Assessment List, and available guidance from the European Data Protection Supervisor.

Fundamental rights impact assessment

- **Does the AI system respect the freedom of expression or assembly?**
- Could the AI-system potentially limit a person's
- freedom to openly express an opinion, partake in a peaceful demonstration or join a union?

Foundations. The European View Framework for Trustworthy AI

Photo RVZ

Trustworthy Artificial Intelligence

EU High-Level Expert Group on AI presented their ethics guidelines for *trustworthy* artificial intelligence:

- (1) **lawful** - respecting all applicable laws and regulations
- (2) **ethical** - respecting ethical principles and values
- (3) **robust** - both from a technical perspective while taking into account its social environment

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Framework for Trustworthy AI
Four Ethical Principles.

Four ethical principles, rooted in fundamental rights

(i) Respect for human autonomy

(ii) Prevention of harm

(iii) Fairness

(iv) Explicability

- Tensions between the principles

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Requirements of Trustworthy AI Including *Sub-requirements*

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

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Requirements of Trustworthy AI Including *Sub-requirements*

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.

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.