Ethical Implication of AI: Assessing Trustworthy AI in Practice

Introduction

Course Coordinator:

Prof. Roberto V. Zicari, SNU

Teaching Assistant:

Sarah Hyojin, SNU

SNU Data Science, Seoul National University

September 12, 2023

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?

Definition of AI Systems

- “The OECD defines an *Artificial Intelligence (AI) System* as a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments.”

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.

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?**

New: Generative AI

- *Source Wikipedia:*

“**Generative artificial intelligence** or **generative AI** is a type of artificial intelligence (AI) system capable of generating text, images, or other media in response to prompts. Generative AI models learn the patterns and structure of their input training data by applying neural network machine learning techniques, and then generate new data that has similar characteristics.

Notable generative AI systems include ChatGPT (and its variant Bing Chat), a chatbot built by OpenAI using their GPT-3 and GPT-4 foundational large language models, and Bard, a chatbot built by Google using their LaMDA foundation model. Other generative AI models include artificial intelligence art systems such as Stable Diffusion, Midjourney, and DALL-E.”

New: Generative AI (cont.)

- *Source Wikipedia:*

"Generative AI has potential applications across a wide range of industries, including art, writing, software development, product design, healthcare, finance, gaming, marketing, and fashion. Investment in generative AI surged during the early 2020s, with large companies such as Microsoft, Google, and Baidu as well as numerous smaller firms developing generative AI models.

However, there are also concerns about the potential misuse of generative AI, such as in creating fake news or deepfakes, which can be used to deceive or manipulate people."

ChatGPT

“ In a nutshell, people are assuming that ChatGPT is “*understanding*” their question, matching it with underlying documents, and curating a response.

That’s not how it works. ”

Bill Franks Director of the Center for Data Science and Analytics at Kennesaw State University.

Source: https://www.odbms.org/2023/06/on-generative-ai-qa-with-bill-franks/

ChatGPT

“The key to what is happening is right in the name … “generative” AI.

ChatGPT is literally generating a response based on probabilities. It is not matching documents or even trying to be truthful. It simply takes your prompt, reduces it down to a series of numbers, and then predicts (one by one) what the best words would be to answer your question. “

Bill Franks Director of the Center for Data Science and Analytics at Kennesaw State University.
Source: https://www.odbms.org/2023/06/on-generative-ai-qa-with-bill-franks/

ChatGPT

“ Due to its extensive training data, ChatGPT actually gets a lot right. It is pretty amazing how well it does, but it is literally making each answer up based on the patterns in its training data. While we call the things it gets wrong, hallucinations, in reality every answer is a hallucination.

If there is a large body of information on the topic you’re asking about and that information is consistent, then ChatGPT will do a pretty good job. The more obscure your topic or the more the available training documents disagree, the more likely to get bad answers”

Bill Franks Director of the Center for Data Science and Analytics at Kennesaw State University.
Source: https://www.odbms.org/2023/06/on-generative-ai-qa-with-bill-franks/

What is Ethics?

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

Assignment: Watch YouTube Video:

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

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

How to assess Trustworthy AI systems in practice

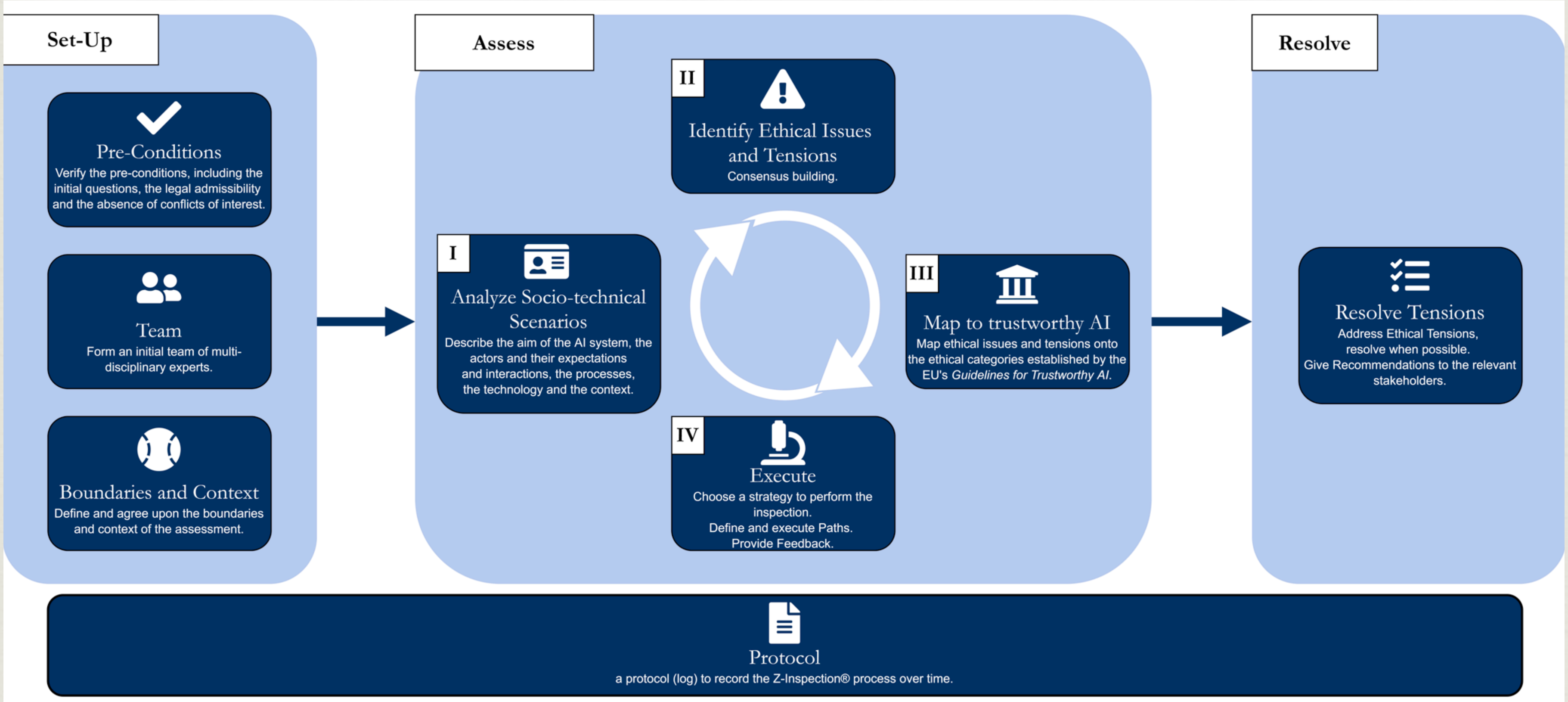
- Moreover, we also need to consider how the adoption of these new algorithms and the lack of knowledge and control over their inner workings may impact those in charge of making decisions.

- 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 5 years of applied research of the Z-Inspection® initiative, a network of high level 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.**

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.

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 oversight.**

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*

Syllabus

4.2 . Additional Frameworks

– The Claim, Arguments and Evidence Framework (CAE)

4.3 Tools

– The ALTAI Web tool

Course Resources

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

https://z-inspection.org/ethical-implication-of-ai-assessing-trustworthy-ai-in-practice-snu-fall-2023/

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 Wednesday,**
Room 302; building 942

Duration of a class: 1hour and 15 minutes.

The course is for **3 Credits**.

Course **starts on September 12, 2023** and **ends on December 6, 2023**

Course Web site

https://z-inspection.org/ethical-implication-of-ai-assessing-trustworthy-ai-in-practice-snu-fall-2023/

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.**

- – **September 12: Get together in presence.**

from 11-12.15 (Korean time).

SNU Graduate School of Data Science, Bldg. 942,

Room 302;

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 Claim, Arguments and Evidence Framework (CAE),**
- **The ALTAI Web tool.**

Assignments

Each week (see web site):

Attend lectures of the week;
Read papers of the week;

Choose a **AI product/solution→ BY SEPTEMBER 26**

- – **September 26: Q&As, Teams buildings, Select AI Systems.**

from 11 till 12.15 (Korean time)

in presence. SNU Graduate School of Data Science, Bldg. 942, **Room** 302;

Reports

- **Mid Term Report** Assess the AI system for "trustworthiness"

  → **Mid Term Report: due October 25**

- **Final Report**: Assess the AI system for "trustworthiness"

  →**FINAL Report: due December 2nd**

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

- **1**. Define the **Boundaries** and **Context** of the assessment**.**
- **2**. Create and Analyze **Socio-technical scenarios;**
- **3. Identify Claims**, provide **Arguments and Evidence**
- **4. Identify Ethical Issues and Tensions;**

1. Define the Boundaries and Context of the assessment.

- **1**. Define the **Boundaries**

  and the **Context**

  of the assessment**.**

2. Create and Analyze Socio-technical scenarios

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.

Create and Analyze Socio-technical scenarios

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?

- **How decisions are made** when using 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?

3. Claims, Arguments and Evidence

- **Identify Claims,**

provide **Arguments (pros and cons)**

and support Arguments with Strong Evidence
(they may be tensions between various evidence!)

Claims that do not have a strong evidence are assumptions and need to be analyzed (e.g. for possible harms)

4. Identify Ethical Issues and Tensions

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

Source [*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”.

Mid Term Report

→ **Mid Term Report: due October 25**

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

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).

- **→FINAL Report: due December 2nd**

Final Report

- Continue Verify Claims, Support Arguments and Evidence;
- Refine the Ethical Issues and Tensions (when possible resolve them);

Final Report

- 5. Map “Issues” to EU framework for trustworthy AI;
- Use the ALTAI web tool
- 6. 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

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

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.

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.

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.

Resources

http://z-inspection.org