Human Agency and Oversight

The first of the seven requirements on the [Ethics Guidelines for Trustworthy AI] refers to **Human Agency and Oversight.**

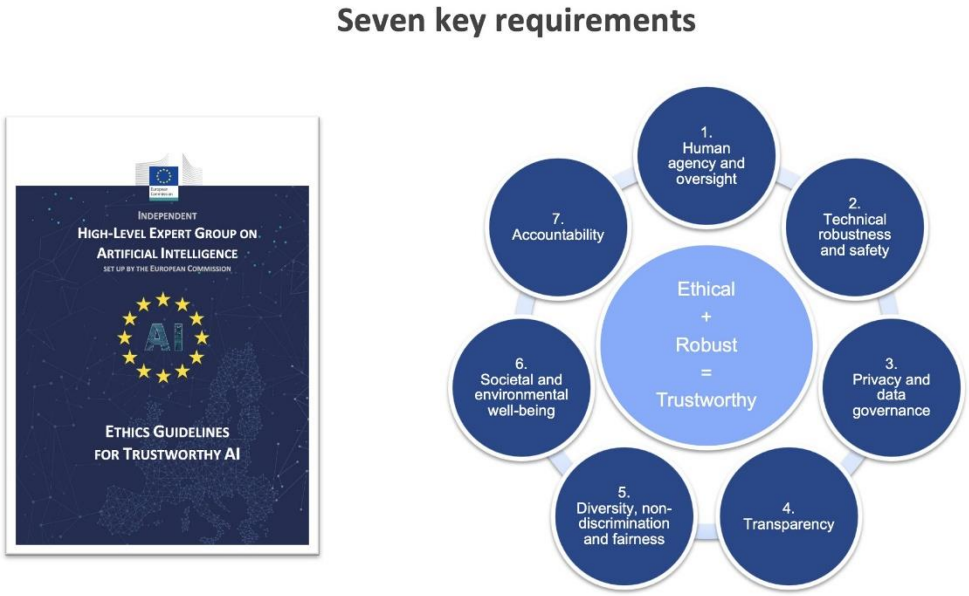

Figure 1: Seven Key Requirements [Source]

Human Agency ensures that AI systems support **human autonomy**, fostering informed and unmanipulated interactions, while also **securing the rights** to rectify AI-driven decisions. On the other hand, Human Oversight refers to the supervision of AI systems, including guiding and monitoring its learning and execution to prevent harm and ensure that the AI system that AI acts in alignment with **human values and ethical standards** (especially in high-risk AI cases, as emphasized in the EU AI Act).

In particular, AI systems should empower human beings by enhancing their capabilities, while allowing them to make informed decisions, respecting their autonomy, and fostering their fundamental rights. At the same time, proper oversight mechanisms need to be established, which can be achieved through human-in-the-loop (**HITL**), human-on-the-loop (**HOTL**), and human-in-command approaches (**HIC**).

When assessing the Human Agency and Oversight requirement in AI systems it is essential to consider multiple factors to ensure that the AI system is designed and operated in ways that uphold human autonomy and facilitate effective human oversight, ultimately **fostering trust and safety in AI technologies**. These factors include:

1. Fundamental Rights Impact Assessment

Following the provisions of the AI Act, an assessment of the impact on fundamental rights that the use of such system may produce must be performed, especially for high-risk AI systems. A Fundamental Rights Impact Assessment (FRIA) evaluates whether and how risks to fundamental rights or central ethical principles can be reduced or justified as necessary in democratic societies, through carefully risk mitigation activities including feedback from stakeholders likely to be affected by the system.

2. Transparency and Explainability

Transparency in AI systems means that the processes and decisions made by the AI are understandable to humans. An AI system should provide explanations that are accessible and comprehensible to non-experts, allowing users to understand how decisions are made and to question or challenge those decisions if necessary. Tools and techniques such as explainable AI (XAI) can be employed to provide insights into the AI's decision-making processes.

3. Level of human control for or involvement in particular AI systems

Three different mechanisms are mentioned in the guidelines: (i) Human-in-the-loop (**HITL**) mechanisms ensure that **humans can intervene and modify** the outcome of an AI system. These mechanisms can range from simple stop buttons to more complex systems where human oversight is integrated into the AI's workflow; (ii) Human-on-the-loop (**HOTL**) mechanisms ensure that humans can effectively **monitor and stop an AI system.** This is usually done through enriched user interfaces that include elaborate visual analytics to inform the user about the status of the system; and (iii) Human-in-Command (**HIC**) mechanisms that ensure that a human has the **ultimate authority and responsibility** over an AI system. This is

usually the case for **healthcare applications**, in which the AI system supports the decision making of a doctor, who is responsible for the actual decision making.

The assessment should check for the presence and functionality of these intervention mechanisms.

4. Accountability

Establishing clear accountability frameworks is essential to ensure Human Oversight. This involves defining who is responsible for the decisions made by AI systems and ensuring that there are mechanisms in place for reporting and addressing any issues that arise. Accountability can be assessed by examining **whether there are clear policies and procedures for handling errors, biases, and adverse outcomes**, as well as whether there are **designated roles and responsibilities for overseeing AI operations**. These frameworks also act as safeguards to prevent overconfidence in or overreliance on the AI system.

5. Training, Education and Raising Awareness

Another important aspect of assessing Human Agency and Oversight is evaluating the awareness, training, and education provided to users and operators of AI systems. Users should be aware of the **capabilities and limitations of AI**, should be **trained on how to interact with it effectively, and how to exercise control over it.** Training programs and materials should be reviewed to ensure they adequately prepare users to maintain Oversight of the AI systems they use. Users should also be aware of the potential emotional or behavioural implications of interactions with non-human agents and the potential negative effects on human autonomy.

6. Feedback Mechanisms

Implementing and assessing (external) feedback mechanisms is crucial for continuous improvement, especially when it comes to assessing impact on Fundamental Rights. Users should be able to provide **feedback** on the AI system's performance, and there should be processes in place to incorporate this feedback into system updates and improvements. This ensures that the AI evolves in ways that better support Human Agency and Oversight over time.

7. Human Agency and Oversight in MANOLO

In MANOLO, **Arcada University of Applied Sciences** leads a dedicated task, that runs in parallel and together with the implementation activities, ensuring that all AI systems or components developed under MANOLO will adhere to the above principles. Arcada, together with the other technical partners, will **develop methods and tools for human oversight in monitoring and validating AI systems within the cloud-edge continuum.**

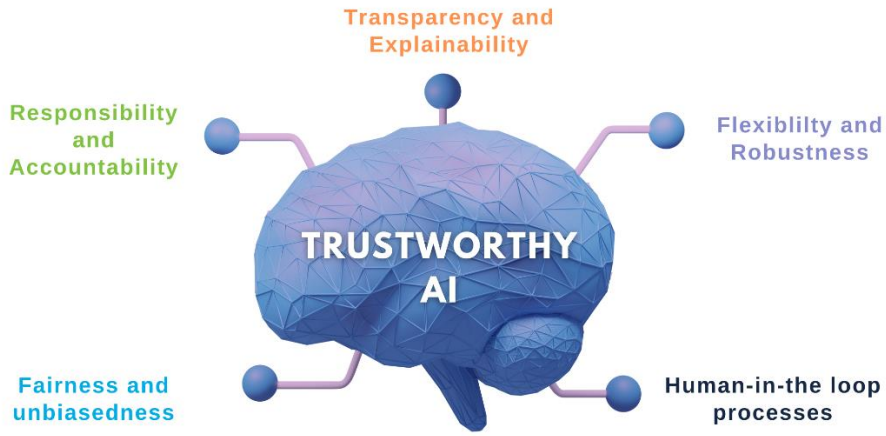

Figure 2: Trustworthy AI components [source]

Wrap Up

To summarise, the "Human Agency and Oversight" key requirement in the Ethics Guidelines for Trustworthy AI ensures AI systems enhance human autonomy, support informed interactions, and provide mechanisms to rectify AI decisions. It emphasizes **designing AI to empower users**, **maintaining their decision-making capabilities**, and **respecting fundamental rights**. Oversight is established through human-in-the-loop (HITL), human-on-the-loop (HOTL), and human-in-command (HIC) mechanisms, which allow humans **to intervene, monitor, and ultimately control AI systems.**

Assessment involves evaluating the impact on fundamental rights, transparency, level of human control or intervention, ethical and social implications, accountability frameworks, user training, and feedback mechanisms.

Specifically, within the MANOLO project, Arcada University of Applied Sciences leads efforts to ensure all AI systems or components developed adhere to these

principles. The dedicated task employs the Z-Inspection® process for Trustworthy AI assessments, a comprehensive assessment framework considering **ethical and legal factors,** to develop methods and tools for human oversight in monitoring and validating AI systems within the **cloud-edge continuum.**