**Exercise 10.1 – Conceptual questions**

**❖ Answer the following questions:**

1. **What is Explainable AI? Find a different definition than the one on the slides.**

Solution:-

An XAI or Transparent AI or Interpretable AI is an AI whose actions can be easily understood and analysed by humans. XAI can be used to implement a social right to explanation.

Hence, XAI is envisaged to provide the following benefits:

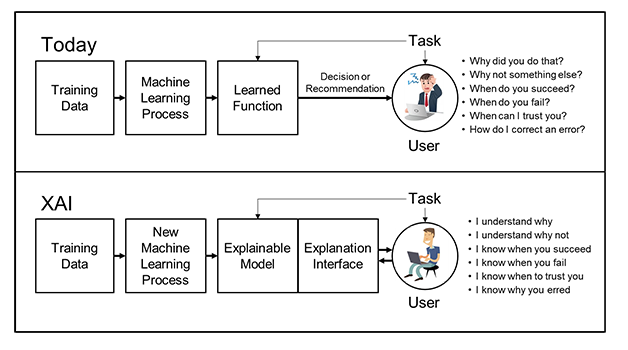
• Transparency and Compliance: It provides an auditable record including all factors and associations related with a given prediction. This enables a business to meet compliance requirements and eliminates concern that the organisation is hiding or doesn’t know how a machine is affecting an outcome of a critical decision

• Ensures that there is an auditable and provable way to defend algorithmic decisions as being fair and ethical.

Reference : H. Hagras, "Toward Human-Understandable, Explainable AI," in Computer, vol. 51, no. 9, pp. 28-36, September 2018, doi: 10.1109/MC.2018.3620965.

1. **How can you explain the so called “Explainable AI Revolution”?**

**Solution:-**



New machine-learning systems will have the ability to explain their rationale, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future. The strategy for achieving that goal is to develop new or modified machine-learning techniques that will produce more explainable models.

1. **What are the main European suggestion for developing a Trustworthy AI system?**

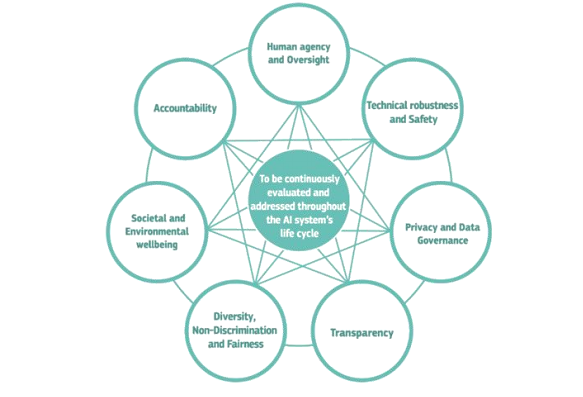
**Solutions:-**

Ethical principles that must be adhered to:

➢ Respect for human autonomy: in accordance with the principle of human self-determination, AI cannot overpower man or make a human being subordinate

➢ Prevention of harm: an AI system must not harm the human being, having to protect his dignity, freedom, privacy and security in society and at work;

➢ Fairness: an AI system must avoid discrimination and prejudice, removing bias and errors, in order to ensure equal treatment of human beings;

➢ Explicability: the processes need to be transparent, the capabilities and purposes of AI systems openly communicated, and decisions, to the extent possible, explainable to those directly and indirectly affected, despite different levels of experience and knowledge****

**\_ Note to self:- Refer to slide 18,19,20 …Each slides refers to different levels of suggestion…**

1. **What are the differences between Model-specific and Model-agnostic solutions?**

Solutions:-

Model-specific solutions: limited to specific model classes, they are able to provide further insights on a model prediction by exploiting the specificities of the model class of interest:

• e.g. compute importance scores in Deep Neural Networks (DeepLIFT [1]).

Model-agnostic solutions: they can be used on any ML model and are applied after the model has been trained. These strategies usually consider the model as a black-box and they work by analyzing only the feature input and the model output:

• e.g. LIME , SHAP , Anchors

1. **Explain in your own words how LIME works.**

Solutions:- Lime takes a local linear approximation of the model’s behaviour to provide an explanation. The model tries to sample around the instances requested for explanation and weights them according to the proximity to the instance in question. These instances are then used to learn a localised linear model which provides approximation of the model and can be used to provide an explanation.

1. **What is Algorithmic Fairness and how are the most important steps for developing a fair ML model?**

Solutions:

Algorithmic fairness refers to the capability of an AI system to guarantee the fairness of its decisions

1. For a system to be considered fair, its output must be evaluated based on some fairness criteria. (Fairness unawareness, Independence, Separation, Sufficiency.)
2. Employ Bias mitigation strategy which can arise out of an ML implementation.
   1. Pre, Post and In processing Bias Mitigation strategies.