XAI – Explainable ai

1. The EU parliament definition of Explainability:

* the importance of the explainability of Al systems’ outputs, processes and values, making them understandable to non-technical audiences and providing them with meaningful information, which is necessary to evaluate fairness and gain trust

1. What is Explainable AI (XAI)

* AI that produces details or reasons to make its functioning clear or easy to understand
* refers to an explanatory agent revealing underlying causes to its or another agent's decision making

1. Goals of XAI

* Improve human interpretability of Al systems
* Foster "informed trust" / appropriate levels of trust
* Support human autonomy in Al-supported decision making
* Determine the justifiability of the decision made by a machine
* Help in deciding accountability and liability (leading to good policy-making)
* Avoid discrimination and the reproduction of societal bias

1. Challenges of XAI

* The Black box problem: Good AI models have around 100Mil parameters. We cannot know for sure which parameters affect the explanation (possible hidden bias)
* What is interpretable for whom: Explanation styles must be adapted to differend needs and knowledge levels, as explanations are different for everyone