Explainable artificial intelligence (XAI extensions)

Explainable AI entails a collection of methods and processes that allow humans to understand machine learning algorithms. Explainable AI essentially tries to explain or describe the decisions made, its biases and weights.

XAI can help humans overcome the black box problem of Artificial intelligence, where we don't ultimately know why a decision was taken. This method therefore can help with transparency and trust in the system, factors like this can be crucial in certain applications such as in law enforcement or in the judicial system.

Besides that explainable AI can offer an advantage in bug testing, by looking at explainable AI engineers can determine if a system is working as intended in the first place. For example an image could be classified correctly but not for the right reasons, for example an AI tasked with interpreting characters in an image checked for copyright tag in the corner of an image.

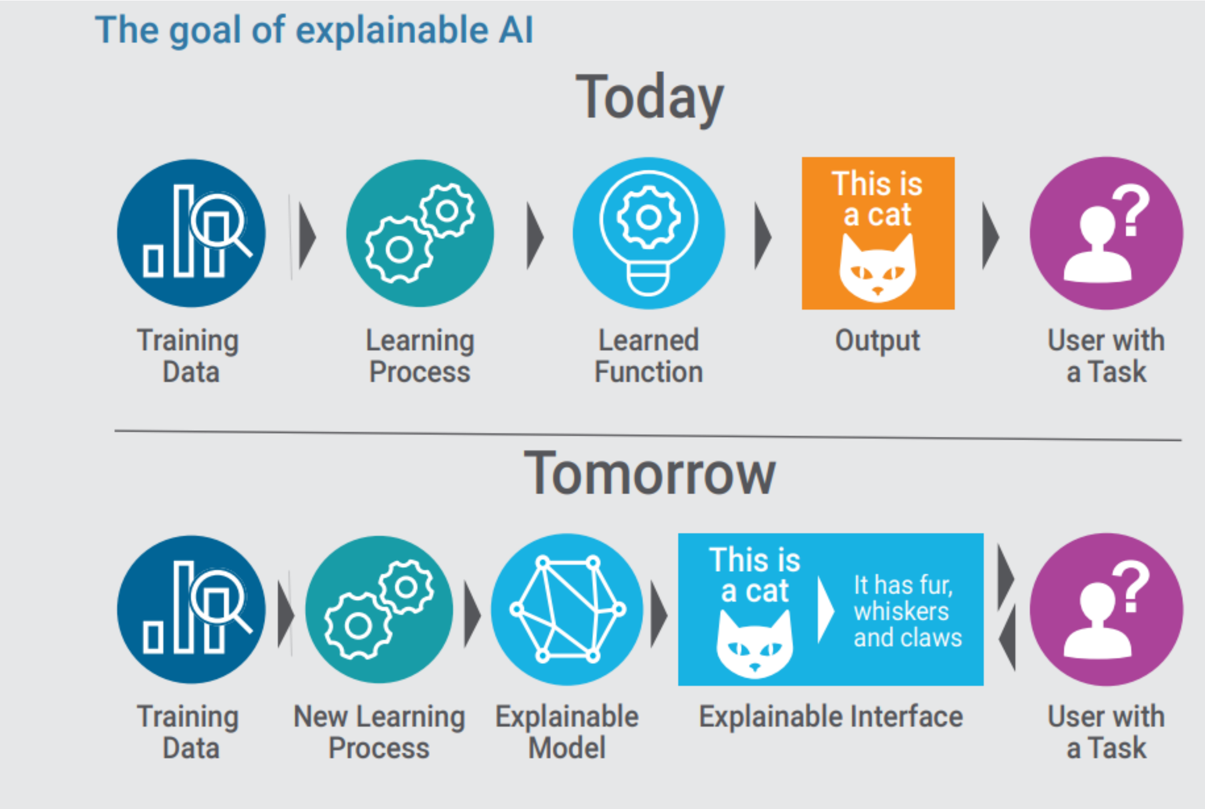

Fig 1 (Visual explainable AI)