**Notes:**

**Black Box AI and Explainable AI:**

Black box AI models are what currently rule the market. They are AI systems that use complex deep learning algorithms, so when used, the user cannot understand how the model has ended up at that answer. This is an issue, because it makes it much more difficult to understand why certain biases arise. Black box systems can exist for two reasons, either it is intentionally done by the software designers, or it is a design flaw that arises from training. Many genAI tools especially are black box systems, as they use such complex deep learning algorithms, that even the software engineers do not understand. This can lead to a phenomenon called the “Clever Hans effect” where the system arrives at the correct answer using the wrong method. This can have drastic effects, as they can pick up on patterns that have nothing to do with the result, and give false positives. This was seen during the COVID-19 pandemic as the AI model was misdiagnosing patients by looking at annotations on the x-rays. This problem therefore led to the development of explainable AI/XAI which uses two methods for trying to understand why the AI reached a conclusion. These two methods are local explanations and global explanations where local explanations are when the developers are focusing solely on a single decision, while global explanations look at the broader picture and asses all the decisions. Local explanations are used for debugging the decisions and understanding why they happen, while global explanations are used to assess biases and fairness within the AI system.