RECAP

Key points to remember

The key learning points for each of the five pillars of ethical AI are included below:

- **Fairness**
    - In AI, fairness is the equitable treatment of individuals or groups of individuals.
    - Fairness is achieved when unwanted bias is mitigated.
    - Protected attributes separate populations into groups.
    - Groups that traditionally receive more favorable outcomes are called privileged groups.
    - Groups that traditionally receive less or no favorable outcomes are called unprivileged groups.
    - There isn’t a defined set of protected attributes.
    - Bias is a systematic error that, intentionally or not, might generate unfair decisions.
- **Robustness**
    - A robust AI system can effectively handle exceptional conditions, like abnormalities in input or malicious attacks, without causing unintentional

harm.

    - Adversarial attacks are intentionally carried out on AI systems to accomplish a malicious end goal by exploiting AI system vulnerabilities.
    - Two types of adversarial attacks are poisoning and evasion.

- **Explainability**
    - AI systems are explainable when everyday people, who do not have any special training in AI, can understand how and why the system came to a particular prediction or recommendation.
    - Interpretability is the degree to which an observer can understand the cause of a decision.
    - Explainability looks at how the AI system arrived at the result.
- **Transparency**
    - Transparency is disclosing information related to the data used for building AI systems, design decisions made throughout the process, model creation, model evaluation, and model deployment.
    - Governance ensures the process followed during the creation and deployment follows the internal policies.
- **Privacy**
    - Personal and sensitive personal information can be used to train models, as long as privacy techniques are applied to the data to preserve the privacy of individuals whose data is included.
    - Many privacy techniques that can be applied to fortify AI against potential breaches of personal or sensitive data. Two that occur during model training are model anonymization and differential privacy. One that occurs after model training is data minimization.