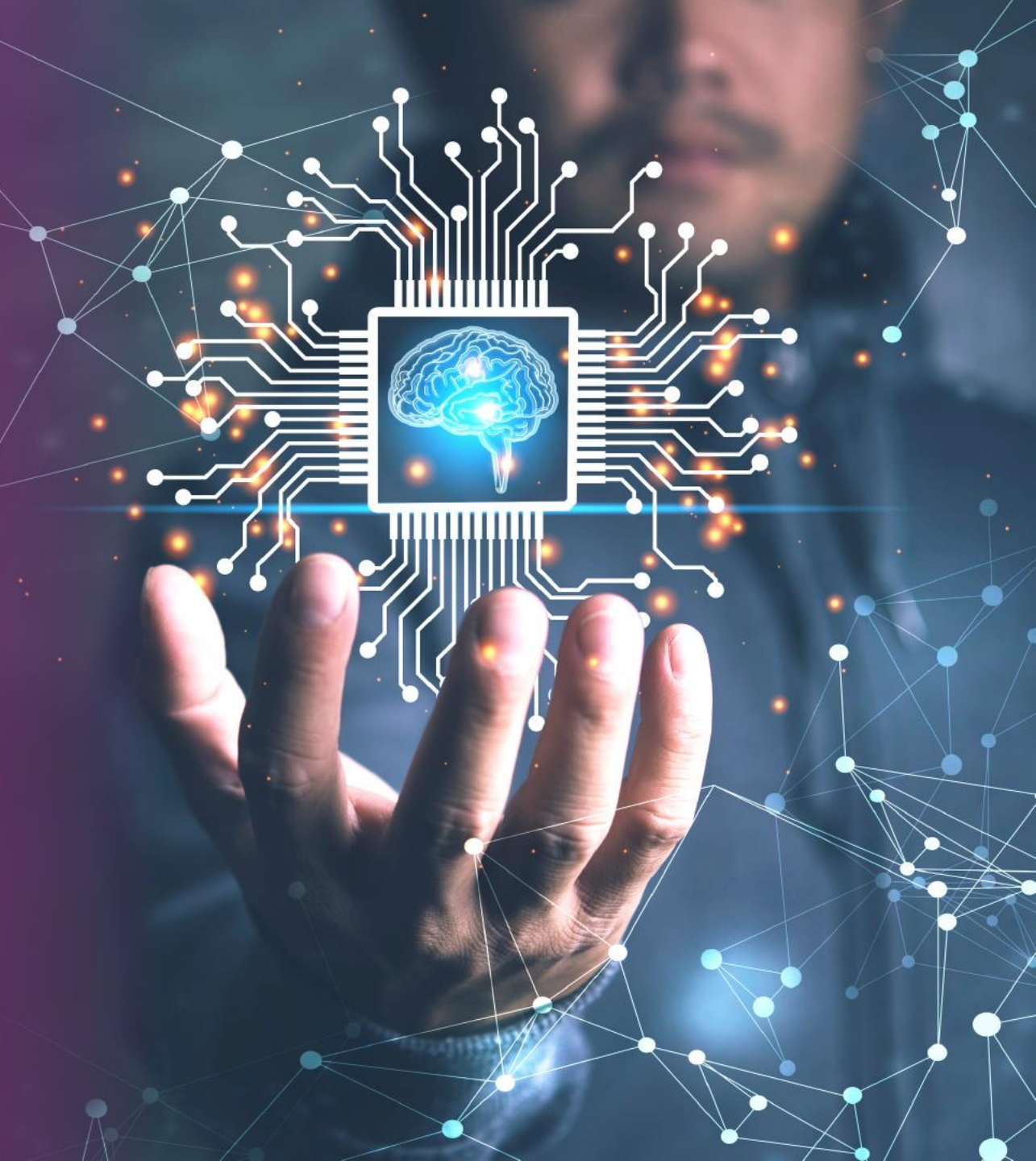
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

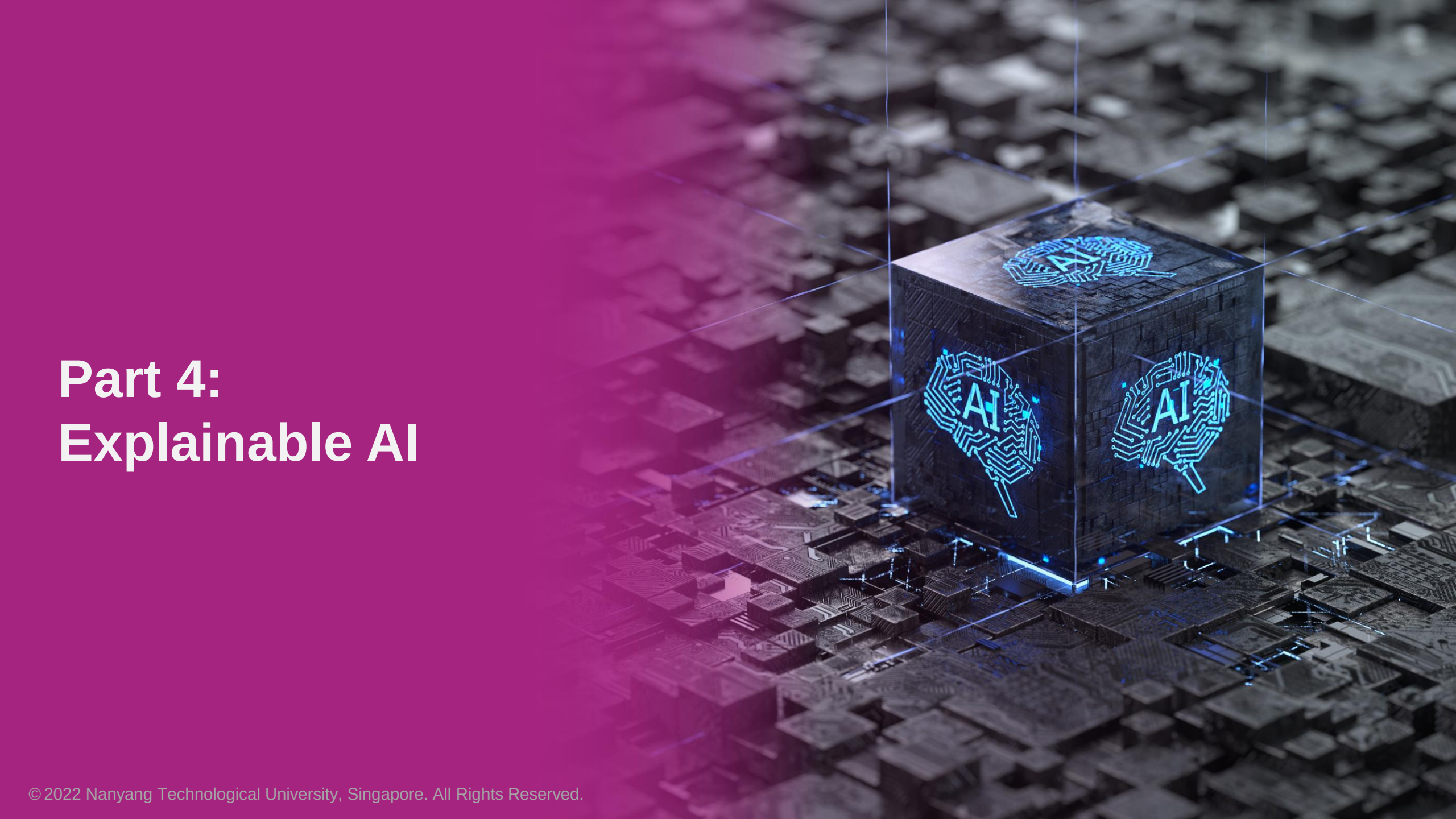
CC0007 Science and Technology for Humanity

Artificial Intelligence I (Technology Aspect)

Assoc Prof Andy Khong, NTU



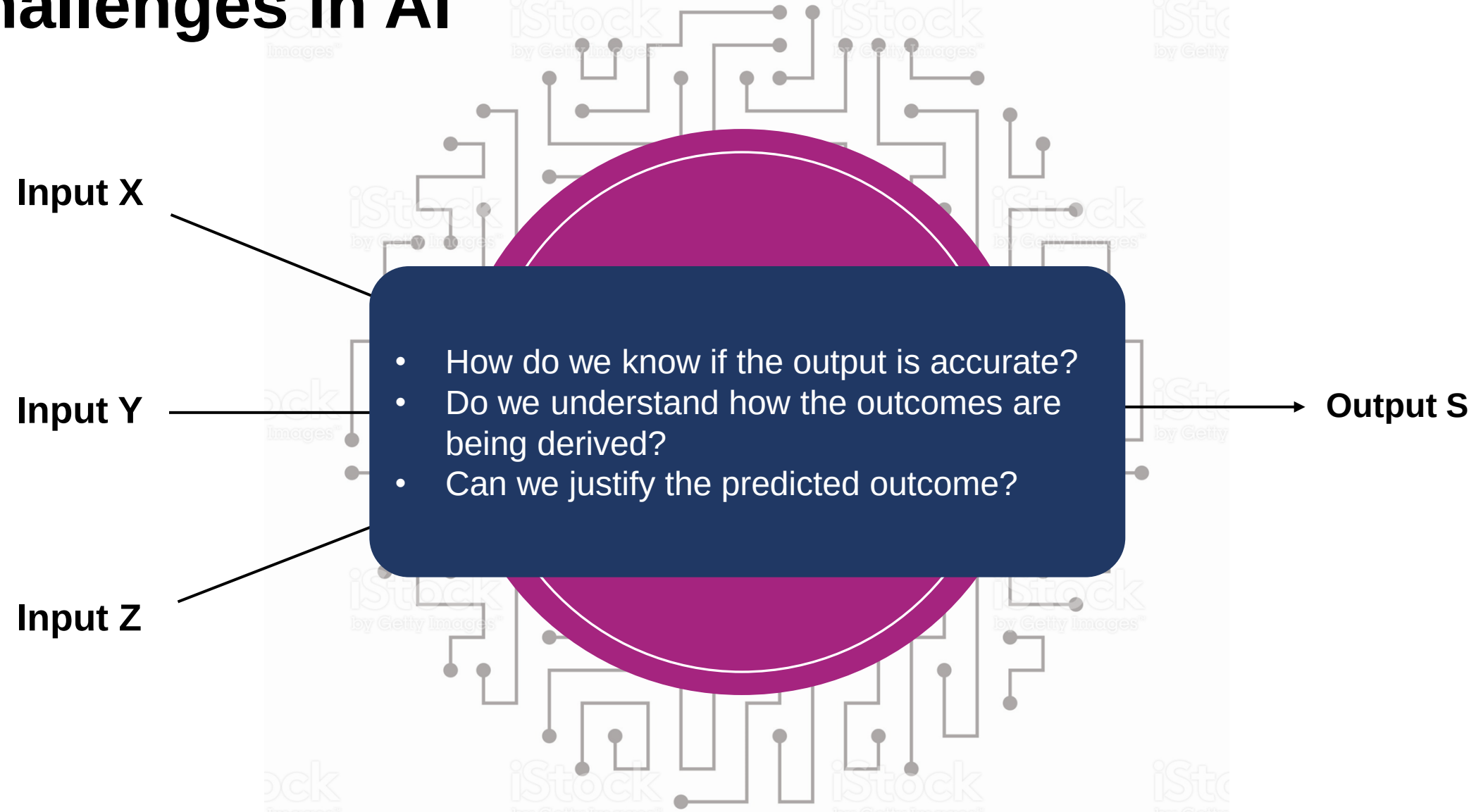
Part 4: Explainable AI



Scenario

Scenario

Challenges in AI



Challenges in AI

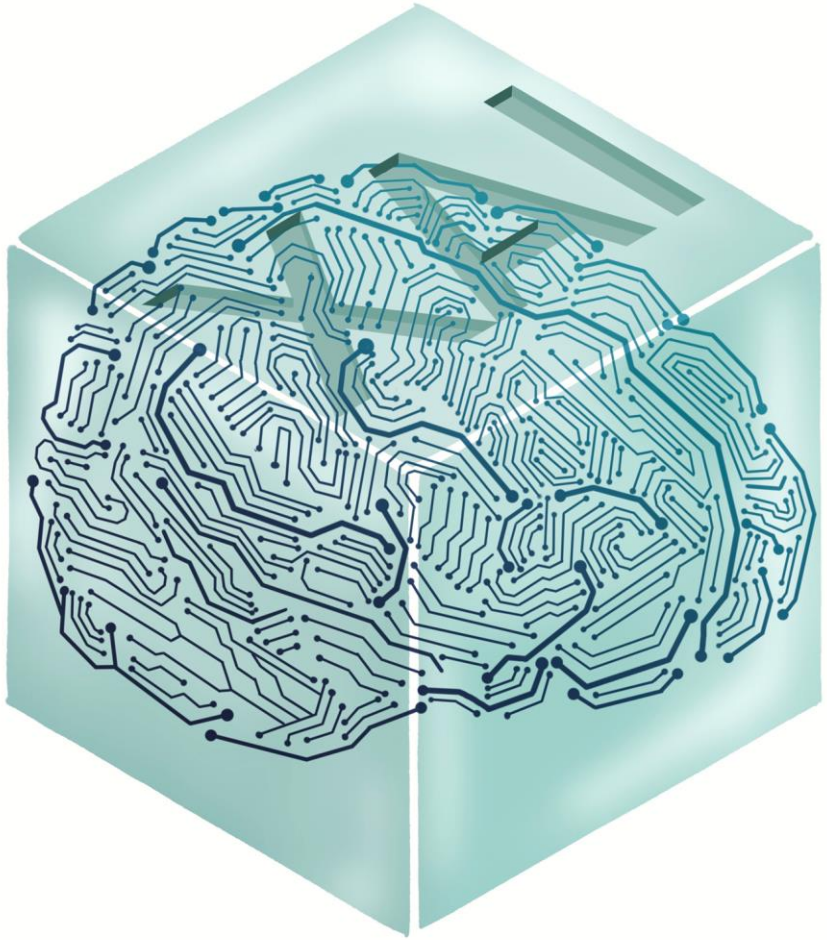
- Explainable artificial intelligence
  - Machine learning has always been used as a '**black box**'.
- Data privacy and security
  - To develop effective and accurate machine learning models, training is usually performed via **sensitive** data.
  - It is important to maintain **data privacy** and **security**.

Challenges in AI

- Fairness in AI model
  - Preventing discriminatory bias will require **bias testing** in development cycle.
- Model validation
  - Under a particular business setting, the data that the business is dealing are consistently **changing** and **evolving**.
  - AI models need to be **revalidated** on new data to maintain good prediction capability.

Explainable Artificial Intelligence

- Explainable artificial intelligence (XAI) is a set of **processes and methods** that allows humans to **comprehend** and **trust the results** created by AI models.
- It is used to **describe** a model, its expected impact and potential biases.
- It helps to characterise **model accuracy**, **fairness**, **transparency** and **outcomes**.
- The **main purpose** of explainability is to ensure that the blackbox models that are used, are not making decisions based on data points that are ambiguous.



Thank you!