Written Exam - Example

Explainable and Trustworthy AI

Eliana Pastor

Written exam

- Structure

- 1-3 multi-choice questions – approx. 1-1.5 points each
- 2-3 open questions – short/medium answers – approx. 3-4 points each
- 1-2 open questions – medium/long answers – approx. 6-8 each

Multi-choice questions - Q1 – 1 point

What does local interpretability in Explainable Artificial Intelligence (XAI) refer to?

A) Understanding the overall behavior of the model

B) Understanding individual predictions or decisions made by the model

C) Understanding the data preprocessing steps in the model

D) Understanding the training process of the model

Open questions – short/medium answer – 4 points

Discuss how counterfactual explanations can be used to provide insights into model predictions. Provide an example scenario where counterfactual explanations would be useful.

Open questions – short/long answer – 7 points

Consider a trained black box healthcare AI system operating with tabular data that predicts the likelihood of disease. How would you ensure that this system provides understandable explanations to doctors? Discuss some methods and the types of explanations that would be appropriate.

Open questions – short/medium answer – 3 points

Outline the concept of surrogate models in XAI. How can surrogate models be used to interpret machine

Multi-choice questions - Q1 – 1 point

Which of the following best describes accountability in the context of Trustworthy AI?

A) The requirement that AI systems operate without any human oversight
B) The responsibility of the entities involved to ensure and answer for the outcomes produced by AI systems
C) The ability of AI systems to self-correct when errors occur
D) The focus on minimizing the cost of deploying AI systems