Exam Example II

What does "model-agnostic" mean in the context of explainability? (1pt)

A) The explanation method is specific to a particular model type.

B) The explanation method can be applied to any machine learning model.

C) The explanation method ignores the input data.

D) The explanation method is only applicable to neural networks.

Which of the following statements is FALSE regarding mechanistic interpretability: (1pt)

A) Mechanistic interpretability aims to reverse-engineer neural networks into human-understandable components like circuits and algorithms.

B) Techniques in mechanistic interpretability often involve identifying specific neurons or patterns responsible for particular behaviors.

C) All layers in a deep neural network encode the same types of interpretable features through mechanistic analysis.

D) Mechanistic interpretability has shown promising results in understanding how models internally represent features like syntax or modular arithmetic.

What are the main limitations of standard explanation methods (like saliency maps)? How do concept-based explainability methods try to solve them? (3 pt)

What is an adversarial attack? Why are adversarial attacks important? (2.5pt)

Briefly describe the core idea behind SHAP (SHapley Additive exPlanations), and how it relates to Shapley values from cooperative game theory. Mention which kind of explanation representation it provides to the user. Outline one advantage of using SHAP for model interpretability. (2.5 pt)

A healthcare provider uses an AI system to diagnose diseases from medical images (e.g., X-rays or MRIs). As part of ensuring the system is trustworthy and useful in clinical settings, it is important that the AI provides understandable explanations to doctors.
Describe at least two explainability methods that could be used to help doctors understand the decisions of the system. Discuss the types of explanation representation (e.g., feature

importance, local rules, visualization example-based) that would be most appropriate for this scenario and why. Choose one explainability method you would recommend for this use case. Clearly justify your choice based on the needs of healthcare professionals. (6 pt)