Teacher Answer Key

Case Study

1. The AI loan-scoring system produces its output through billions of mathematical operations across many variables interacting in complex, nonlinear ways. No human can trace through those operations to identify exactly which factors lowered Yolanda's score or by how much. The loan officer could see the final score but had no access to a traceable decision path that would allow them to explain the outcome in specific terms.

2. When black box AI makes consequential decisions, it may violate legal requirements for explanation (such as the Fair Credit Reporting Act's requirement for specific denial reasons), it becomes impossible for affected individuals to challenge potentially discriminatory outcomes, and it prevents regulators from auditing whether the system is complying with anti-discrimination law. The black box nature also makes it easier for historical patterns of discrimination embedded in training data to perpetuate undetected.

Multiple Choice

1. B

2. D

3. C

Matching

1-C, 2-A, 3-B, 4-E, 5-D

Fill-in-the-Blank

1) discriminatory, 2) transparency, 3) decision

Short Answer (Sample Responses)

Q1: When AI makes a music recommendation, the worst outcome is a song you dislike, which you can simply skip. When AI makes a hiring, lending, or medical decision, the stakes include someone's livelihood, financial security, or physical health. These decisions must also comply with legal requirements for fairness and explanation. An unexplainable error in a high-stakes AI decision can cause lasting harm to a real person who has no ability to understand or challenge it.

Q2: Mechanistic interpretability research attempts to understand what specific neurons and circuits inside large models are actually computing, with the goal of developing genuine understanding of how models produce their outputs. This is

important for safety because if we cannot understand what an AI is doing internally, we cannot reliably predict when it will fail, detect when it is behaving in unintended ways, or build systems to prevent harmful outputs.

Q3: Answers will vary. A strong argument against the trade-off: the hospital algorithm case shows that accepting unexplainable decisions in high-stakes domains can cause systematic discrimination that harms vulnerable groups, so the accuracy gains from black box models do not justify the legal and ethical costs. A strong argument for some trade-off: in medical diagnosis, a highly accurate black box AI that saves more lives than an explainable but less accurate alternative may still be preferable if combined with strong human oversight. Students should cite specific examples from the reading.