EXAM 4 PRACTICE (FINAL) - WEEKS 10-16

Practice Exam

CIS 350: AI for Cybersecurity - Smart Defense for the Digital Business

Colorado State University - Computer Information Systems

About this document

This is a PRACTICE exam for the final (Exam 4). The questions differ from the real exam but cover the same Weeks 10-16 material in the same format (100 points): multiple choice, true/false, short answer, and applied scenarios. Try it closed-book and timed, then check the separate answer key. Exam 4 is taken online in Canvas with LockDown Browser, open-notes.

Total: 100 points. Part A multiple choice (20 x 2 = 40), Part B true/false (10 x 1 = 10), Part C short answer (5 x 4 = 20), Part D applied scenarios (3 x 10 = 30). No heavy math, no coding.

Part A - Multiple Choice (20 x 2 = 40 pts)

1. Which is an example of a privacy attack on a model?

A. Feature scaling
B. A train/test split
C. Model inversion (reconstructing training data)
D. Cross-validation

2. Adversarial examples succeed mainly because models:

A. Match learned patterns and are brittle near boundaries
B. Understand meaning perfectly
C. Run offline
D. Have too few parameters

3. Data poisoning is best countered by:

A. Ignoring the data
B. A faster GPU
C. Reusing passwords
D. Vetting the training data and tracking provenance

4. A jailbreak attempts to:

A. Speed up the model
B. Get the model past its safety rules
C. Back up the prompt
D. Encrypt the output

5. Prompt leakage exposes:

A. The hidden system prompt
B. The GPU
C. The password hash
D. The firewall rules

6. The root cause of prompt injection is:

A. Too few parameters
B. Offline mode
C. A slow GPU
D. No wall between trusted instructions and untrusted data

7. Equalized odds (a fairness definition) requires equal:

A. File sizes
B. True-positive AND false-positive rates across groups
C. Numbers of features
D. Training time

8. A model card documents a model's:

A. Market price only
B. The server's IP address
C. Purpose, performance, fairness, and limits
D. The training source code

9. A bias 'feedback loop' means:

A. Faster training
B. More GPUs
C. Fewer features
D. Biased outputs become tomorrow's training data

10. Homomorphic encryption lets you:

A. Compute on encrypted data without decrypting it
B. Skip training
C. Delete data
D. Speed up the GPU

11. 'Anonymize by removing names' is weak because:

A. It is illegal
B. It is slow
C. Quasi-identifiers can re-identify people
D. It encrypts the data

12. A LARGER epsilon in differential privacy means:

A. More privacy
B. Less privacy (and usually more accuracy)
C. No effect
D. A slower GPU

13. Transparency (a trust pillar) is about:

A. How the system overall works and its limits
B. Why one decision was made
C. The GPU
D. The password

14. Keeping a human in the loop provides:

A. Faster GPUs
B. More training data
C. Oversight and the ability to override
D. Cheaper storage

15. The FIRST phase of the incident-response lifecycle is:

A. Recover
B. Prepare
C. Ignore
D. Delete

16. 'The algorithm decided' is a failure of which pillar?

A. Speed
B. Size
C. Color
D. Accountability

17. Few-shot learning is enabled largely by:

A. Millions of labeled examples
B. A larger learning rate
C. A strong pretrained base (transfer learning)
D. More training epochs

18. Federated learning is often paired with ___ for privacy:

A. Differential privacy
B. A larger batch size
C. Signature-based antivirus
D. A faster network connection

19. AI-generated polymorphic malware mainly defeats:

A. Encryption at rest
B. Two-factor authentication
C. Network firewalls
D. Signature-only detection

20. The durable defensive goal in the arms race is to:

A. Win permanently
B. Raise attacker cost, detect fast, and adapt
C. Ignore attacks
D. Stop deploying anything

Part B - True / False (10 x 1 = 10 pts)

21. Data poisoning happens at training time. (T / F)
22. Overall accuracy can hide per-group unfairness. (T / F)
23. Adversarial changes must be large and obvious. (T / F)
24. In differential privacy, a smaller epsilon means more privacy. (T / F)
25. Removing the sensitive attribute guarantees a fair model. (T / F)
26. Federated learning uploads all raw data to a central server. (T / F)
27. Robustness should be monitored over time, not just tested once. (T / F)
28. There is one perfect defense that solves all AI security problems. (T / F)
29. Transfer learning can cut the number of labels you need. (T / F)
30. Explainability is only about how fast a model runs. (T / F)

Part C - Short Answer (5 x 4 = 20 pts)

31. What is an adversarial example and why does it work? When does data poisoning happen instead (timing)?

32. What is prompt injection, and name two layered defenses for an LLM app.

33. Why can a model be accurate overall but unfair? Name one fairness definition.

34. What does epsilon control in differential privacy (which way is more private)? Name one other privacy technique.

35. Name two trustworthy-AI pillars (explain one), then one advanced technique and the problem it addresses.

Part D - Applied Scenarios (3 x 10 = 30 pts)

36. A loan-approval model (10 pts)

A bank's ML model approves or denies loans. An audit finds it denies one group far more, and applicants want to know why.

(a) Which per-group metric reveals the unfairness, and one way to reduce it? (3 pts)
(b) Which trustworthy-AI pillar does 'tell me why I was denied' require? (2 pts)
(c) Name one privacy risk of the applicant data and a defense. (2 pts)
(d) Name one governance step before redeploying. (3 pts)

37. An LLM email assistant under attack (10 pts)

A company's LLM assistant reads and drafts email and can look up account info. Attackers try to make it leak data or send unauthorized messages.

(a) Name the attack and one variant (direct or indirect). (2 pts)
(b) Give one guardrail and one least-privilege decision. (4 pts)
(c) It behaves oddly after a model update - which trust pillar is at risk and why? (2 pts)
(d) Name the first two IR phases you would follow. (2 pts)

38. A malware detector adapting to attackers (10 pts)

A security team runs an ML malware detector and wants it to keep working as attackers adapt.

(a) Attackers craft inputs to evade it - name the attack and one defense. (3 pts)
(b) They try to poison its training feed - how does that hurt it, and one defense? (3 pts)
(c) Labels are scarce and a new attack appears - name one advanced technique that helps. (2 pts)
(d) Why is there 'no perfect defense,' and what should the team do? (2 pts)

End of practice exam. Check your answers with the separate answer key.