1.5 Leveraging AI in Cyber Defense

Name: _____ Class: _____ Date: _____ **Total: 20 marks**

KEY WORDS anomaly • threshold • false positive • false negative • configuration • alert

Objective

Build the skills to answer exam questions on **AI in cyber defense** : finding attacks in huge amounts of data, choosing a **threshold** , and why people must review AI decisions.

You must be able to:

- explain how AI helps defenders handle the scale of network and login data
- work out the **false positives** and **false negatives** for a given threshold
- explain why people must review AI suggestions, and how an adversary may try to fool an AI detector

1 Key knowledge and worked examples

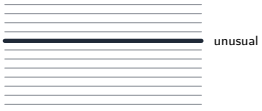
■ AI on the defender's side

idea	rule
scale	networks make millions of events a day; AI looks at all of them and flags the few that look unusual
speed	AI works all the time and reacts in seconds
safer settings	AI can suggest configuration changes, but a person reviews them, because the AI may not know what the business needs
score and threshold	the AI gives each event a score for how likely it is an attack; events above the **threshold** are blocked or flagged
false positive	normal activity flagged as an attack: wasted time, blocked users, and real alerts lost in the noise
false negative	an attack that is not flagged: the adversary gets through
moving the threshold	a lower threshold catches more attacks but flags more normal events; a higher one does the opposite
limits	AI learns from past data, so a new kind of attack may not match; an adversary may act slowly to look normal

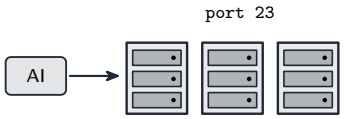
> **Answer shapes:** for a threshold, sort the events into flagged and not flagged, then compare each with what it really was.

2 Practice

2.1 A network creates 2 million log events a day. An analyst can check one event every 10 seconds, for 8 hours a day. Work out how many events one analyst can check in a day, and explain why AI is needed. [2]



2.2 An AI tool suggests closing port 23 on every server. Explain why a person should review this before it is applied. [2]



2.3 An AI alarm flags 500 logins a day. Checks show that 495 of them were normal staff. Name this kind of error, and explain one problem it causes. [2]

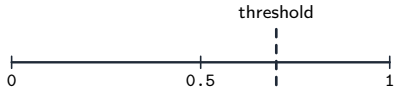


3 Exam-style questions

3.1 An AI model gives each login a score from 0 to 1 for how likely it is an attack. The team blocks every login whose score is above a threshold. Later, the team finds out what each login really was.

login	score	really
L1	0.95	attack
L2	0.80	normal
L3	0.72	attack
L4	0.65	attack
L5	0.40	normal
L6	0.30	normal

(a) The threshold is 0.7. State the logins that are blocked, and the number of false positives and false negatives. [3]



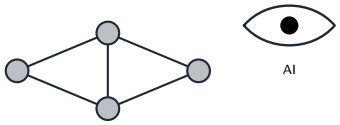
(b) Explain the effect of lowering the threshold to 0.6. [2]

(c) State which kind of error is worse for a bank's login system, with a reason. [1]

exercises.lol exercises.com.cn

3.2 A company uses AI to watch its network and to suggest changes to its firewall rules.

(a) Give two advantages of AI over human analysts for finding attacks. [2]



(b) Explain why the AI might miss a new kind of attack. [2]

(c) Explain why people should approve the AI's rule changes. [2]

(d) Describe how an adversary could try to fool the AI detector. [2]

Solutions

Marking: **M1** a correct method, **A1** the correct answer, **B1** an independent mark.

Each answer shows the steps, then the **result**.

2.1

$8 \times 3600 \div 10 = 2880$ events; this is far fewer than 2 million, so AI must check every event and pass only the unusual ones to people

2.2

the AI may not know that a system still needs that port; applying it without checking could stop a service that people need

2.3

false positives; analysts waste time and stop trusting the alarm, so the 5 real attacks may be missed

3.1

(a) L1, L2 and L3 are blocked; one false positive, L2; one false negative, L4

(b) L4 is now blocked, so there are no false negatives; a lower threshold also flags more normal logins in general, so there can be more false positives

(c) a false negative, because an adversary gets into a customer's account

3.2

(a) any two: it is faster; it can check every event; it works all day and night without getting tired

(b) the AI learned from past attacks; a new attack may not look like anything in that data

(c) the AI may block traffic the business needs; people understand the business and are responsible for the decision

(d) the adversary acts slowly, or makes the attack look like normal traffic; so the scores stay below the threshold and nothing is flagged