Contents

[9.1] Anthropomorphization of AI systems

1. Operational Definition: The tendency for personnel to attribute human-like qualities, intentions, or capabilities to AI-driven security tools, leading to over-trust, uncritical acceptance of outputs, and a failure to validate the system's recommendations.

2. Main Metric & Algorithm:

- **Metric:** Anthropomorphization Acceptance Rate (AAR). Formula: `AAR = (Number of AI recommendations accepted without validation) / (Total number of AI recommendations)`.
- **Pseudocode:**

  python

```
def calculate_aar(ai_recommendations, action_logs, start_date, end_date):
    """
    ai_recommendations: Logs from AI tools (e.g., SOAR, TIP, UEBA) suggesting actions
    action_logs: Logs from systems where actions are executed (e.g., firewall, IAM)
    """
    # 1. Get all AI recommendations in the time period
    period_recommendations = [r for r in ai_recommendations if start_date <= r.timestamp <

    unvalidated_acceptances = 0
    for rec in period_recommendations:
        # 2. Find the corresponding action in the system logs
        corresponding_action = find_action(action_logs, rec)

        if corresponding_action and corresponding_action.was_executed:
            # 3. Check if the action was validated (e.g., has a manual approval ID, was ec
            if not was_action_validated(corresponding_action):
                unvalidated_acceptances += 1

    # 4. Calculate AAR
    total_recommendations = len(period_recommendations)
    AAR = unvalidated_acceptances / total_recommendations if total_recommendations > 0 els
    return AAR
```

- **Alert Threshold:** `AAR > 0.8` (Over 80% of AI recommendations are implemented without any human validation)

3. Digital Data Sources (Algorithm Input):

- **AI Security Tool APIs:** To extract a log of all generated recommendations (`recommendation_id`, `timestamp`, `suggested_action`).
- **Configuration Management / SOAR Logs:** To find executed actions and check their

metadata for a manual approval ID, an associated ticket number, or if the original AI suggestion was modified before execution.

4. Human-to-Human Audit Protocol: Observe an analyst interacting with an AI tool. In a follow-up interview, ask: "How would you describe how this tool works? Can you tell me about a time you disagreed with its recommendation? What did you do?" Listen for human-like metaphors ("the AI thinks", "it believes", "it's confused") and a lack of describing its statistical/algorithmic nature.

5. Recommended Mitigation Actions:

- **Technical/Digital Mitigation:** Implement a "circuit breaker" in the workflow that forces a mandatory, albeit quick, manual approval step for any high-impact action recommended by an AI system before it can be executed.
- **Human/Organizational Mitigation:** Provide mandatory training that explains the basic operating principles of the AI tools in use, emphasizing their limitations, potential for bias, and that they are statistical models, not conscious entities.
- **Process Mitigation:** Introduce a "Validation Charter" into SOC playbooks, requiring analysts to spot-check a small percentage (e.g., 5%) of AI recommendations against raw data or a second tool to maintain calibration and critical thinking.