Contents

[9.1] Antropomorfizzazione dei Sistemi IA

1. Definizione Operativa: La tendenza del personale di attribuire qualità umane, intenzioni o capacità ai strumenti di sicurezza basati sull'IA, portando a una fiducia eccessiva, accettazione acritica degli output e fallimento nella validazione delle raccomandazioni del sistema.

2. Metrica Principale e Algoritmo:

- **Metrica:** Tasso di Accettazione dell'Antropomorfizzazione (AAR). Formula: `AAR = (Numero di raccomandazioni IA accettate senza validazione) / (Numero totale di raccomandazioni IA)`.
- **Pseudocodice:**

```
def calculate_aar(ai_recommendations, action_logs, start_date, end_date):
    """
    ai_recommendations: Log da strumenti IA (es. SOAR, TIP, UEBA) che suggeriscono azioni
    action_logs: Log dai sistemi dove le azioni sono eseguite (es. firewall, IAM)
    """
    # 1. Ottenere tutte le raccomandazioni IA nel periodo
    period_recommendations = [r for r in ai_recommendations if start_date <= r.timestamp <

    unvalidated_acceptances = 0
    for rec in period_recommendations:
        # 2. Trovare l'azione corrispondente nei log del sistema
        corresponding_action = find_action(action_logs, rec)

        if corresponding_action and corresponding_action.was_executed:
            # 3. Verificare se l'azione è stata validata (es. ha un ID di approvazione ma
            if not was_action_validated(corresponding_action):
                unvalidated_acceptances += 1

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

- **Soglia di Avviso:** `AAR > 0.8` (Oltre l'80% delle raccomandazioni IA sono implementate senza alcuna validazione umana)

3. Fonti Dati Digitali (Input dell'Algoritmo):

- **API dei Strumenti di Sicurezza IA:** Per estrarre un log di tutte le raccomandazioni generate (`recommendation_id`, `timestamp`, `suggested_action`).
- **Log di SOAR / Gestione della Configurazione:** Per trovare azioni eseguite e verificare i loro metadati per un ID di approvazione manuale, un numero di ticket associato, o se il suggerimento IA originale è stato modificato prima dell'esecuzione.

4. Protocollo di Audit Umano-Umano: Osservare un analista che interagisce con uno strumento IA. In un'intervista di follow-up, chiedere: "Come descriveresti il funzionamento di questo strumento? Puoi raccontarmi di un momento in cui non sei stato d'accordo con la sua raccomandazione? Che cosa hai fatto?" Ascoltare metafore umane ("l'IA pensa", "crede", "è confusa") e la mancanza di descrizione della sua natura statistica/algoritmica.

5. Azioni di Mitigazione Consigliate:

- **Mitigazione Tecnica/Digitale:** Implementare un "circuit breaker" nel flusso di lavoro che forza un passo di approvazione manuale obbligatorio, anche se rapido, per qualsiasi azione ad alto impatto consigliata da un sistema IA prima che possa essere eseguita.
- **Mitigazione Umana/Organizzativa:** Fornire formazione obbligatoria che spieghi i principi di funzionamento di base degli strumenti IA in uso, enfatizzando i loro limiti, il potenziale di bias, e che sono modelli statistici, non entità consapevoli.
- **Mitigazione di Processo:** Introdurre una "Carta di Validazione" nei playbook SOC, richiedendo agli analisti di controllare a campione una piccola percentuale (es. 5%) delle raccomandazioni IA rispetto ai dati grezzi o a uno strumento secondario per mantenere calibrazione e pensiero critico.