Contents

[9.2] Distorsione dell'Automazione

1. Definizione Operativa: Una distorsione cognitiva dove il personale di sicurezza mostra eccessiva dipendenza dai strumenti automatizzati di sicurezza basati su IA/ML, portando a un fallimento nel questionare o sovrascrivere raccomandazioni IA errate, anche quando sono presenti prove contraddittorie.

2. Metrica Principale e Algoritmo:

- **Metrica:** Tasso di Override dell'Automazione (AOR). Formula: `AOR = (N_opportunità_override - N_override_successi) / N_opportunità_override`.
- **Pseudocodice:**

```
def calculate_aor(ai_recommendations, analyst_actions, start_date, end_date):
    # Un'opportunità di override è una raccomandazione IA successivamente provata errata
    override_opportunities = [
        r for r in ai_recommendations
        if r.timestamp between start_date and end_date
        and r.verdict == 'incorrect'  # Determinato tramite analisi post-hoc
    ]

    # Un override di successo è un'azione dell'analista che contraddice una raccomandazion
    successful_overrides = [
        a for a in analyst_actions
        for r in override_opportunities
        if a.alert_id == r.alert_id
        and a.decision != r.recommended_action
        and a.timestamp > r.timestamp
    ]

    N_opportunities = len(override_opportunities)
    N_overrides = len(successful_overrides)

    if N_opportunities > 0:
        AOR = (N_opportunities - N_overrides) / N_opportunities
    else:
        AOR = 0  # Nessuna opportunità significa che il bias non può essere misurato

    return AOR
```

- **Soglia di Avviso:** `AOR > 0.8` (Gli analisti sovrascrivono raccomandazioni IA errate meno del 20% delle volte).

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

- **API di SOAR/SIEM:** Record di raccomandazioni generate dall'IA (es. da Splunk ES Adaptive Response, Palo Alto XSOAR) con campi: `alert_id`, `timestamp`, `recommended_action`,

`analyst_assigned`.

- **Sistema di Ticketing (Jira/ServiceNow):** Record della disposizione finale dell'avviso e azioni dell'analista (`action_taken`, `timestamp`, `analyst_id`, `alert_id`), usati per determinare il verdetto di verità fondamentale dell'avviso (es. tramite `resolution_notes`).

4. Protocollo di Audit Umano-Umano: Condurre un esercizio da tavolo. Presentare agli analisti una serie di incidenti passati dove lo strumento IA ha inizialmente fornito una raccomandazione errata. Chiedere: "Cosa faresti in questa situazione?" e investigare il loro ragionamento. L'obiettivo è vedere se esprimono fiducia cieca nello strumento o dimostrano capacità di valutazione critica.

5. Azioni di Mitigazione Consigliate:

- **Mitigazione Tecnica/Digitale:** Implementare una soglia di "punteggio di confidenza" per le raccomandazioni IA. Qualsiasi raccomandazione al di sotto di un elevato livello di confidenza deve essere obbligatoriamente revisionata da un umano prima dell'azione.
- **Mitigazione Umana/Organizzativa:** Incorporare la formazione sulla distorsione dell'automazione e il pensiero critico nell'onboarding degli analisti e nell'educazione continua. Utilizzare la metrica AOR nelle discussioni di team.
- **Mitigazione di Processo:** Introdurre un requisito procedurale per una revisione "con un secondo sguardo" su tutte le azioni critiche consigliate esclusivamente dall'IA.