Contents

[9.7] Accettazione di Allucinazioni dell'IA

1. Definizione Operativa: La tendenza di accettare e agire sulla base di output confidenti ma errati o interamente fabbricati generati da sistemi IA (es. LLM che riassumono log o generano codice), specialmente quando l'output si allinea con le credenze preesistenti dell'utente.

2. Metrica Principale e Algoritmo:

- **Metrica:** Tasso di Incidente da Allucinazione (HIR). Formula: `HIR = N_incidenti_causati_da_allucinaz / N_azioni_generate_da_IA_totali`.
- **Pseudocodice:**

```
def calculate_hir(incident_reports, start_date, end_date):
    # Questo richiede revisione post-incidente per identificare la causa radice
    incidents_caused_by_ai = [
        i for i in incident_reports
        if i.root_cause == 'AI Hallucination'  # Tagging manuale richiesto
        and i.date between start_date and end_date
    ]

    # Questo è un approssimativo; le azioni totali coinvolte nell'IA sarebbero migliori
    total_incidents = get_total_incidents(start_date, end_date)

    if total_incidents > 0:
        HIR = len(incidents_caused_by_ai) / total_incidents
    else:
        HIR = 0

    return HIR
```

- **Soglia di Avviso:** `HIR > 0` (Qualsiasi incidente causato da un'allucinazione dell'IA deve attivare una revisione immediata e un avviso).

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

- **Piattaforma di Risposta agli Incidenti (Jira, ServiceNow):** Report di incidenti con un campo `root_cause` che può essere taggato.
- **Log di SOAR/SIEM:** Per stimare il numero totale di azioni intraprese sulla base dell'output dell'IA.

4. Protocollo di Audit Umano-Umano: Implementare un passo obbligatorio di "Verifica dell'Output dell'IA" nel processo di post-mortem dell'incidente per qualsiasi incidente dove un riassunto, codice o comando generato dall'IA è stato coinvolto. La domanda è: "L'output dell'IA era accurato e verificabile?"

5. Azioni di Mitigazione Consigliate:

- **Mitigazione Tecnica/Digitale:** Implementare modelli di guardrail che verifichino gli output dell'IA per plausibilità, fabbricazioni note, o rischi di sicurezza prima che siano presentati all'utente.
- **Mitigazione Umana/Organizzativa:** Formare gli utenti sulla possibilità di allucinazioni dell'IA. Instillare un principio di "fiducia ma verifica" per tutto il contenuto generato dall'IA, specialmente codice o comandi.
- **Mitigazione di Processo:** Stabilire una politica rigorosa che il codice o i comandi generati dall'IA devono essere revisionati e approvati da un secondo umano prima dell'esecuzione in qualsiasi ambiente di produzione o sensibile.