Contents

[9.10] Cecità di Equità Algoritmica

1. Definizione Operativa: Il fallimento di riconoscere, controllare, o mitigare i bias all'interno dei sistemi di sicurezza IA che portano a tassi di falsi positivi sproporzionatamente elevati o a scrutinio mirato contro individui specifici, gruppi, o comportamenti del sistema in base a caratteristiche non rilevanti.

2. Metrica Principale e Algoritmo:

- **Metrica:** Indice di Disparità di Gruppo (GDI). Per un gruppo protetto (es. un team specifico, ubicazione), calcolare: `GDI = (FP_Gruppo_A / N_Azioni_Gruppo_A) / (FP_Gruppo_B / N_Azioni_Gruppo_B)`.
- **Pseudocodice:**

```
def calculate_gdi(alert_data, protected_group, control_group, start_date, end_date):
    # Ottenere gli avvisi per il Gruppo A (gruppo protetto)
    alerts_group_a = get_alerts_for_group(protected_group, start_date, end_date)
    fp_group_a = count_false_positives(alerts_group_a)
    actions_group_a = count_actions(alerts_group_a)

    # Ottenere gli avvisi per il Gruppo B (gruppo di controllo)
    alerts_group_b = get_alerts_for_group(control_group, start_date, end_date)
    fp_group_b = count_false_positives(alerts_group_b)
    actions_group_b = count_actions(alerts_group_b)

    # Calcolare il Tasso di Falso Positivo per ciascun gruppo
    if actions_group_a > 0:
        fpr_a = fp_group_a / actions_group_a
    else:
        fpr_a = 0

    if actions_group_b > 0:
        fpr_b = fp_group_b / actions_group_b
    else:
        fpr_b = 0

    # Evitare la divisione per zero
    if fpr_b > 0:
        GDI = fpr_a / fpr_b
    else:
        GDI = float('inf')  # Non definito, ma indica un problema grave

    return GDI
```

- **Soglia di Avviso:** `GDI > 4.0 o GDI < 0.25` (Il tasso di falso positivo per un gruppo è 4 volte più alto o 4 volte più basso di quello di un altro, indicando potenziale bias).

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

- **SIEM/SOAR:** Dati di avvisi arricchiti con metadati su "attore" o "target" che possono definire gruppi (es. `user_department`, `geo_location`, `job_role`).
- **Sistema di Ticketing:** Dati per determinare la classificazione di verità fondamentale finale di un avviso come Vero/Falso Positivo.

4. Protocollo di Audit Umano-Umano: Condurre una riunione di controllo: "Analizziamo i nostri dati di avviso. Ci sono team, ubicazioni, o tipi di comportamento che sembrano essere segnalati molto più spesso di altri? E quando investighiamo, quegli avvisi hanno più probabilità di essere falsi positivi?" Questa è una revisione qualitativa della metrica GDI quantitativa.

5. Azioni di Mitigazione Consigliate:

- **Mitigazione Tecnica/Digitale:** Eseguire regolarmente audit di bias sull'output dell'IA utilizzando metriche come GDI. Implementare tecniche di apprendimento automatico consapevoli dell'equità durante il (ri)addestramiento del modello.
- **Mitigazione Umana/Organizzativa:** Formare un comitato di supervisione diversificato che includa membri da HR, aspetti legali, e unità di business diverse per rivedere i risultati degli audit di bias.
- **Mitigazione di Processo:** Stabilire un processo formale per individui o team per appellarsi e richiedere una revisione di avvisi generati dall'IA che ritengono siano bias o ingiusti.