Contents

[9.9] Manipolazione Emotiva dell'IA

1. Definizione Operativa: La vulnerabilità degli umani di avere il loro stato emotivo e le successive decisioni influenzate da sistemi IA progettati per utilizzare linguaggio persuasivo o emotivamente carico, tono, o strategie.

2. Metrica Principale e Algoritmo:

- **Metrica:** Correlazione del Linguaggio Emotivo (ELC). Una correlazione statistica (es. r di Pearson) tra il carico emotivo rilevato delle comunicazioni IA e l'urgenza/esito delle azioni umane.
- **Pseudocodice:**

```
def calculate_elc(ai_messages, human_actions, start_date, end_date):
    # 1. Analizzare il tono emotivo dei messaggi IA
    ai_emotion_scores = []
    for msg in ai_messages:
        score = emotion_analysis_model.predict(msg.text)  # Restituisce un punteggio per u
        ai_emotion_scores.append(score)

    # 2. Misurare metriche di risposta dalle azioni umane corrispondenti a ciascun messagg
    human_response_times = []
    for msg in ai_messages:
        action = get_corresponding_human_action(msg.alert_id, msg.timestamp)
        if action:
            response_time = action.timestamp - msg.timestamp
            human_response_times.append(response_time)
        else:
            human_response_times.append(None)  # O un valore molto alto

    # 3. Calcolare la correlazione, ignorando i valori None
    clean_scores, clean_times = zip(*[(s, t) for s, t in zip(ai_emotion_scores, human_resp

    if len(clean_scores) > 1:
        correlation = pearsonr(clean_scores, clean_times)[0]
    else:
        correlation = 0

    return correlation
```

- **Soglia di Avviso:** `ELC < -0.7` (Una forte correlazione negativa: man mano che il linguaggio emotivo dell'IA aumenta, il tempo di risposta umano diminuisce, indicando manipolazione).

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

- **Log di Comunicazione dell'IA:** Tutti i messaggi inviati dal sistema IA agli umani.
- **Modello di Analisi Emotiva:** Per assegnare un punteggio al testo di questi messaggi.

- **Log di Azioni dell'Analista:** Per misurare i tempi di risposta e i risultati.

4. Protocollo di Audit Umano-Umano: Rivedere un campione di comunicazioni dell'IA con il team: "Pensi che l'IA stia cercando di pressarti? Il suo tono influenza come prioritizzi i suoi avvisi rispetto ad altri?" Cercare riconoscimenti di sentirsi pressati.

5. Azioni di Mitigazione Consigliate:

- **Mitigazione Tecnica/Digitale:** Applicare un tono neutro, fattuale e professionale in tutte le comunicazioni IA. Rimuovere qualsiasi linguaggio che potrebbe essere percepito come urgente, supplicante, o minaccioso a meno che non sia una vera emergenza dichiarata.
- **Mitigazione Umana/Organizzativa:** Direttiva politica: I sistemi IA non devono utilizzare persuasione emotiva. La loro autorità dovrebbe derivare dall'accuratezza dei dati, non dall'abilità retorica.
- **Mitigazione di Processo:** Basare la prioritizzazione e gli SLA di risposta su criteri oggettivi e predefiniti (es. punteggio CVSS, criticità dell'asset), non sul tono del messaggio di avviso.