Contents

[9.5] Effetti della Valle Inquietante

1. Definizione Operativa: Il senso di disagio, sfiducia o repulsione scatenato da sistemi IA che imitano il comportamento umano (es. chatbot conversazionali) in un modo che è quasi, ma non perfettamente, realistico, compromettendo la collaborazione efficace.

2. Metrica Principale e Algoritmo:

- **Metrica:** Rapporto di Sentimento Negativo (NSR) nelle Interazioni Umano-IA. Formula: `NSR = N_interazioni_negative / N_interazioni_totali`.
- **Pseudocodice:**

```
def calculate_nsr(chat_logs, ai_agent_id, start_date, end_date):
    # Ottenere tutte le interazioni con l'agente IA
    interactions = get_chat_interactions(ai_agent_id, start_date, end_date)

    # Utilizzare un modello di analisi del sentimento sui messaggi umani all'interno di qu
    negative_interactions = set()

    for interaction in interactions:
        for message in interaction.human_messages:
            sentiment = sentiment_analysis_model.predict(message.text)
            if sentiment == 'negative':
                negative_interactions.add(interaction.id)
                break  # Un messaggio negativo contrassegna l'intera interazione

    N_total = len(interactions)
    N_negative = len(negative_interactions)

    if N_total > 0:
        NSR = N_negative / N_total
    else:
        NSR = 0

    return NSR
```

- **Soglia di Avviso:** `NSR > 0.3` (Oltre il 30% delle interazioni conversazionali con l'agente IA contengono sentimento negativo rilevabile).

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

- **API della Piattaforma di Chat (Slack/Teams):** Log delle conversazioni con il chatbot/agente IA (`message`, `user`, `timestamp`, `thread_id`).
- **Un Modello di Analisi del Sentimento Pre-addestrato:** Per elaborare il testo dei messaggi umani in queste conversazioni (es. utilizzando una libreria come VADER o un'API dedicata).

4. Protocollo di Audit Umano-Umano: Condurre focus group o interviste: "Come ti senti nell'interagire con il chatbot di sicurezza? C'è qualcosa che sembra 'strano' o frustrante?" Analizzare il feedback per temi relativi a risposte innaturali, frustrazione, o inquietudine.

5. Azioni di Mitigazione Consigliate:

- **Mitigazione Tecnica/Digitale:** Riprogettare lo stile conversazionale dell'IA. Spesso, allontanarsi dal tentativo di imitare gli umani verso una personalità chiaramente artificiale ma utile e trasparente (es. "Sono un'IA progettata per…") può ridurre gli effetti della valle inquietante.
- **Mitigazione Umana/Organizzativa:** Essere trasparenti riguardo alle capacità e limitazioni dell'IA. Spiegare che è uno strumento, non un umano.
- **Mitigazione di Processo:** Fornire una via di fuga facile e ovvia verso un operatore umano se l'interazione con l'IA diventa frustrante o inefficace.