Contents

[9.6] Fiducia nell'Opacità dell'Apprendimento Automatico

1. Definizione Operativa: La propensione di fidarsi degli output di modelli di apprendimento automatico complessi "black box" senza questione, a causa dell'incapacità di comprendre i loro meccanismi interni, che sono spesso percepiti come misticamente autorevoli.

2. Metrica Principale e Algoritmo:

- **Metrica:** Tasso di Consultazione della Spiegazione (ECR). Formula: `ECR = N_richieste_spiegazione / N_raccomandazioni_IA_presentate`.
- **Pseudocodice:**

```
def calculate_ecr(ai_recommendations, explanation_logs, start_date, end_date):
    N_recommendations = count_ai_recommendations(start_date, end_date)

    # Contare quante volte gli utenti hanno fatto clic su "Perché?" o una funzione di spie
    N_explanation_requests = count_explanation_requests(explanation_logs, start_date, end_

    if N_recommendations > 0:
        ECR = N_explanation_requests / N_recommendations
    else:
        ECR = 0

    return ECR
```

- **Soglia di Avviso:** `ECR < 0.05` (Gli utenti richiedono una spiegazione per meno del 5% delle raccomandazioni IA, indicando fiducia cieca).

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

- **Log dell'Interfaccia del Sistema IA:** Log specifici dell'applicazione che registrano i clic degli utenti su funzioni di spiegabilità (es. un pulsante "Spiega questa raccomandazione").
- **API del Sistema IA:** Log di tutte le raccomandazioni presentate agli utenti.

4. Protocollo di Audit Umano-Umano: Durante osservazioni o interviste, dopo che una raccomandazione è mostrata, chiedere direttamente all'analista: "Puoi guidarmi attraverso il motivo per cui l'IA potrebbe aver suggerito questo?" L'incapacità di articolare alcun motivo, o il ricorso a "perché l'IA l'ha detto", indica fiducia nell'opacità.

5. Azioni di Mitigazione Consigliate:

- **Mitigazione Tecnica/Digitale:** Implementare i principi di IA Spiegabile (XAI) per progettazione. Forzare il sistema a fornire una rationale breve e leggibile per umani per ogni raccomandazione *per impostazione predefinita*, non dietro un clic.
- **Mitigazione Umana/Organizzativa:** Formare gli analisti sui fondamenti di come funziona l'IA (es. "Cerca modelli simili a incidenti passati") per demistificarla.
- **Mitigazione di Processo:** Rendere la consultazione della spiegazione un passo formale e obbligatorio nel protocollo operativo standard per gestire avvisi generati dall'IA.