Intelligent Systems Engineering

Bias and Fairness in Generative Models: Ethical Challenges and Engineering Solutions

Nicola Montanari

A.Y 2024/2025

Obiettivo del progetto

L'obiettivo di questa ricerca è stato quello di analizzare e sintetizzare, tramite una revisione sistematica, le conoscenze attuali sul bias e l'equità nei modelli di Intelligenza Artificiale Generativa, identificando le sfide etiche e le soluzioni ingegneristiche proposte per promuovere uno sviluppo responsabile.

Introduzione

Bias:

Errori sistematici e ripetibili che si manifestano negli output del modello, spesso riflettendo e amplificando pregiudizi sociali, stereotipi o disuguaglianze presenti nei dati di addestramento.

Equità:

Obiettivo di garantire che i modelli generativi producano risultati giusti, imparziali e rappresentativi per tutti i gruppi demografici, evitando discriminazioni e perpetuazione di stereotipi.

Metodologia

Database utilizzato: Scopus

Periodo di Ricerca: 1 Gennaio 2024 - 1 Novembre 2025

Procedimento: Stringa di ricerca specifica, screening titoli/abstracts & revisioni complete

Domande di Ricerca (RQs)

Come si manifestano e sono caratterizzati i problemi di bias ed equità nei modelli generativi?

Quali problemi etici derivano da problemi di bias e equità nei modelli generativi?

Quali soluzioni ingegneristiche e strategie di mitigazione sono state proposte per affrontare bias ed equità nei modelli generativi?

Manifestazioni del Bias (RQ1)

Tipi di Bias:

- **Representational Bias:** Rappresentazione inaccurata/incompleta o omissione di gruppi demografici.
- **Output Bias:** Riflette stereotipi interni (es. bias di genere).
- **Stereotype Amplification:** Amplificazione degli stereotipi esistenti.
- **Disparate System Performance:** Degradazione della qualità del linguaggio per specifici gruppi sociali o variazioni linguistiche.

Fonti del Bias:

- **Training Data:** Dati massivi non curati, pregiudizi sociali/storici.
- **Model Design:** Scelte algoritmiche che privilegiano gruppi di maggioranza.
- **Human Annotations & Feedback:** Pregiudizi degli annotatori umani.
- **Evaluation & Deployment:** Dataset benchmark non rappresentativi.

Questioni Etiche derivanti da Bias ed Equità (RQ2)

Conseguenze dirette: Risultati Discriminatori e Danni

- **Danni Rappresentativi:** Perpetuazione di stereotipi dannosi, misrappresentazione o cancellazione di gruppi sociali. (Es. linguaggio tossico, astrazioni negative).
- **Danni Allocativi:** Distribuzione iniqua di risorse o opportunità. (Es. discriminazione diretta in sanità o finanza).

Conseguenze generali: Erosione della Fiducia e Sfide di Responsabilità

- **Erosione della Fiducia Pubblica:** Minare la fiducia nella tecnologia IA.
- **Sfide di Responsabilità:** La natura "black box" rende difficile attribuire la responsabilità del bias.
- **Trade-off Equità-Performance:** Spesso migliorare l'equità degrada altre metriche di performance.

Soluzioni Ingegneristiche (RQ3)

Le soluzioni sono categorizzate in base al punto di intervento nel ciclo di sviluppo dell'IA.

Pre-processing:

Questa strategia interviene modificando i dati o i prompt prima che il modello venga addestrato o effettui l'inferenza. (Es: Data Augmentation)

In-training:

Questo approccio integra modifiche o vincoli direttamente nella procedura di addestramento del modello. (Es: Reinforcement Learning from Human Feedback)

Intra-processing:

Queste tecniche regolano il comportamento del modello durante la fase di inferenza senza la necessità di un addestramento aggiuntivo. (Es: Modified Token Distribution)

Post-processing

Questo metodo opera sugli output generati dal modello dopo che sono stati prodotti, identificandoli e modificandoli. (Es: Keyword Replacement)

Fine