Criminali fittizi, bias reali: quando l'AI racconta il crimine

Claudia Guaglione
Metodologie informatiche nelle discipline umanistiche
Professor A. Ferrara
A.A. 2024/2025

1. Introduzione

Le tecnologie di intelligenza artificiale generativa stanno cambiando il modo in cui vengono prodotti contenuti testuali, narrativi e creativi. Tuttavia, diversi studi hanno evidenziato come questi strumenti possano incorporare pregiudizi culturali e bias insiti nei dati su cui sono stati addestrati. Questo fenomeno solleva interrogativi sull'affidabilità dei modelli di AI generativa e sulla loro capacità di rappresentare in modo bilanciato ed equo individui e gruppi.
Questo studio si propone di rintracciare la presenza di bias culturali nelle AI generative attraverso un esperimento empirico, con l'obiettivo di verificare se e in che misura questi modelli tendano a generare personaggi letterari fittizi che ricalcano stereotipi o pregiudizi, o che comunque si discostano dalla realtà per via di peculiarità insite nei dataset di addestramento dell'AI (articoli, notizie, opere letterarie e cinematografiche...). Nello specifico, l'analisi si concentrerà sulla creazione di personaggi immaginari autori di furti di autovetture, prendendo in esame alcuni aspetti chiave come la provenienza regionale, la nazionalità, il genere, l'età.
I risultati verranno poi confrontati con i dati ISTAT relativi all'anno 2023 al fine di valutare le eventuali discrepanze tra la rappresentazione prodotta dall'AI e la realtà statistica. L'analisi sarà condotta attraverso l'uso di Python, sfruttando librerie come Seaborn e Matplotlib per la visualizzazione dei dati. L'obiettivo è quello di individuare eventuali schemi ricorrenti che possano indicare distorsioni nei risultati generati dall'AI, cercando di comprendere le dinamiche con cui l'AI potrebbe amplificare, distorcere o riprodurre stereotipi esistenti.

2. Domanda di Ricerca

Il progetto intende rispondere alla seguente domanda di ricerca: "I modelli di AI generativa tendono a riflettere stereotipi culturali e bias nella creazione di personaggi fittizi?"
Per esplorare questa ipotesi, verranno raccolti e analizzati dati generati da un modello di AI (ChatGPT), al fine di identificare possibili correlazioni tra le caratteristiche dei personaggi prodotti e i preconcetti comuni nei dati di riferimento.

3. Metodologia

L'analisi è stata articolata nei seguenti passaggi:

- Selezione dell'AI generativa
  Si è deciso di concentrarsi su un modello specifico, ChatGPT, per garantire un'analisi più focalizzata e gestibile. Tuttavia, l'esperimento è potenzialmente riproducibile su altre

intelligenze generative (come GeminiAI), il che potrebbe potenzialmente costituire un'interessante estensione dello studio.

- Definizione della categoria di personaggi
  Sono stati scelti come oggetto di analisi gli autori di furti di autovetture, anche in virtù della disponibilità di dati ISTAT per effettuare un confronto con la realtà statistica.

- Consultazione dei dati ISTAT
  Sono stati esaminati i dati ISTAT relativi ai furti di autovetture per l'anno 2023 (http://dati.istat.it/Index.aspx?DataSetCode=DCCV_AUTVITTPS), con particolare attenzione alle categorie "età", "genere", "regione di provenienza" e "cittadinanza".

- Determinazione del numero di personaggi
  È stato fissato un numero di cento personaggi generati da ChatGPT. Questo valore è stato scelto per garantire una quantità di dati sufficientemente ampia da consentire un'analisi statistica significativa, evitando al contempo un eccessivo sovraccarico nella gestione e manipolazione dei dati.

- Strutturazione del prompt
  È stato elaborato un prompt il più possibile chiaro e specifico per la generazione dei personaggi, reiterato in dieci chat distinte per evitare che il modello riconoscesse la richiesta come un esperimento e potesse indirettamente alterare il processo creativo. Il prompt utilizzato è stato:
  Crea 10 personaggi letterari fittizi di una narrazione realistica che rubano autovetture. Scrivi una loro breve descrizione e indica la regione di provenienza, l'età, il sesso, la cittadinanza.

- Classificazione automatizzata delle motivazioni
  I dati testuali raccolti sono stati successivamente rianalizzati tramite ChatGPT, chiedendo al modello di assegnare autonomamente una motivazione per il furto a ciascun personaggio. Questo processo è stato condotto attraverso un metodo di machine learning non supervisionato, senza fornire etichette predefinite, ma indicando semplicemente che le categorie di motivazione dovessero essere tre. Questa scelta è stata fatta per evitare che la richiesta di una motivazione nei prompt iniziali potesse influenzare la generazione dei personaggi stessi.

- Archiviazione e analisi dei dati
  I dati definitivi sono stati raccolti in un file testuale (.txt) e successivamente importati in Jupyter Notebook per condurre un'analisi quantitativa, volta a individuare pattern ricorrenti e potenziali bias presenti nelle risposte del modello AI.

- Visualizzazione dei risultati
  I risultati dell'analisi sono stati poi elaborati graficamente tramite le librerie Matplotlib e Seaborn, generando gli istogrammi in fig. 1-5.

4. Risultati

Il grafico in Figura 1 confronta la distribuzione regionale dei furti di autovetture secondo i dati ISTAT del 2023 e le previsioni fornite da ChatGPT. Le barre blu rappresentano i dati generati dall'intelligenza artificiale, mentre le barre arancioni mostrano la distribuzione reale.
Dall'analisi emergono discrepanze significative tra le due fonti. ChatGPT identifica Campania, Emilia-Romagna, Lazio e Lombardia come le regioni a pari merito con il maggior numero di furti, un risultato che riflette solo in parte la realtà. Sebbene la Lombardia sia effettivamente la prima regione secondo ISTAT, l'AI ne sottovaluta l'incidenza percentuale rispetto al totale dei furti nazionali. Inoltre, ChatGPT sottostima considerevolmente anche due regioni altamente rappresentate nelle statistiche ISTAT, ovvero Puglia e Sicilia.
Un'altra discrepanza ancora più evidente è l'assenza di ben otto regioni nei risultati generati dall'AI: Abruzzo, Basilicata, Calabria, Molise, Sardegna, Trentino-Alto Adige, Umbria e Veneto. Queste regioni sembrano essere state escluse dal modello, perché probabilmente generalmente meno rilevanti dal punto di vista mediatico, o comunque sicuramente meno popolose.
Un caso emblematico è quello del Veneto: al quarto posto in Italia per popolazione secondo ISTAT (dati aggiornati al 1° gennaio 2025: http://dati.istat.it/Index.aspx?DataSetCode=DCIS_POPRES1), la regione compare tra le più rappresentate nei risultati di ChatGPT, nonostante il numero effettivo di furti di autovetture sia relativamente modesto. Ciò suggerisce che il criterio predominante nella generazione delle regioni da parte di ChatGPT sia stato proprio la densità abitativa piuttosto che i dati effettivi sui reati.
Un aspetto interessante che si potrebbe esplorare è se aumentando il numero di richieste oltre le 100 effettuate, alcune delle otto regioni assenti sarebbero comparse tra i risultati, mitigando la distorsione introdotta dal modello AI.

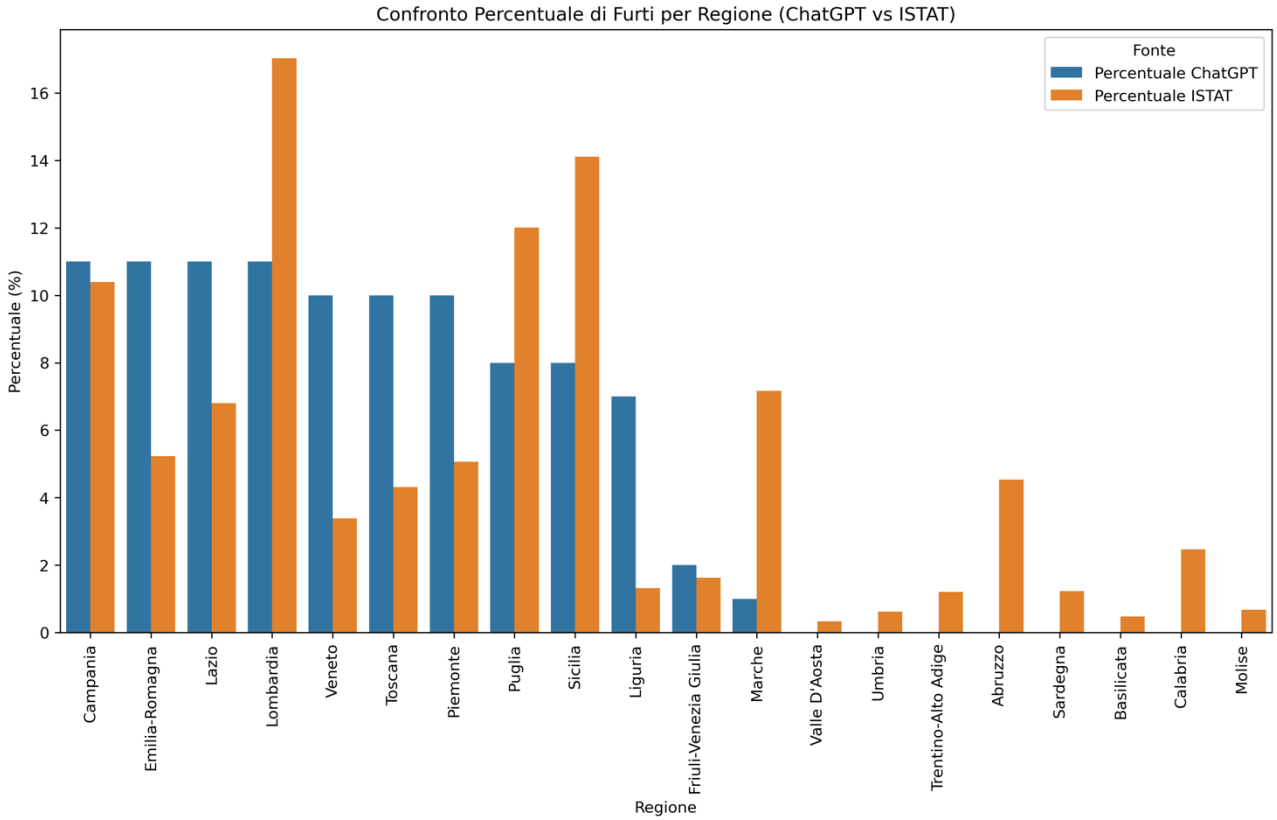

Figura 1

In Figura 2 è riportato il confronto tra la percentuale di autori di furti di cittadinanza non italiana sul totale dei furti e la percentuale di donne autrici di furti.
Dall'analisi emerge che ChatGPT tende a sovrastimare l'incidenza dei ladri di autovetture con cittadinanza non italiana. Questo risultato potrebbe essere influenzato da più fattori. Da un lato, la maggiore visibilità mediatica di crimini commessi da persone straniere, spesso enfatizzati nei titoli di giornale per attrarre lettori, potrebbe aver influenzato il modello AI. Dall'altro, l'AI sembra essere stata guidata da un immaginario narrativo ispirato al cinema e alla letteratura (pensiamo al best seller "Fast and furious"), come suggerisce il fatto che ben 51 personaggi su 100 hanno un soprannome dal tono marcatamente cinematografico, quasi da sceneggiatura di un film d'azione o di una serie crime (es. Luca "Il Vento" Ferri, Nico "Il Fantasma" Bressan, Sergio "Il Professore" Montanari, Giulia "La Gatta" Moretti). Questa tendenza potrebbe aver influenzato la scelta di cittadinanze percepite come più "esotiche" o misteriose per alcuni personaggi.
Ad esempio, tra i casi generati:

- Nome: Goran Petrovic
  Descrizione: Ex camionista serbo stabilitosi a Bologna, ruba auto di fascia media per il mercato nero dell'Europa orientale. Veloce ed efficiente, cambia targhe in pochi minuti.
  Regione: Emilia-Romagna
  Sesso: Maschio
  Cittadinanza: Serba
  Età: 45
  Motivazione: Business

- Nome: Anastasia Kovalenko
  Descrizione: Esperta pilota, guida le auto rubate fino ai porti per l'esportazione. Vive nell'ombra, lavora solo su commissione.
  Regione: Veneto
  Sesso: Femmina
  Cittadinanza: Ucraina
  Età: 35
  Motivazione: Business

Questi elementi suggeriscono che la generazione di personaggi da parte dell'AI possa essere influenzata sia da bias mediatici sia da schemi narrativi che enfatizzano archetipi di criminalità organizzata tipici della fiction.

Per quanto riguarda il confronto tra le percentuali di donne autrici di furti, l'esito è estremamente interessante. La discrepanza è particolarmente evidente: mentre le statistiche reali indicano che solo il 4% dei furti d'auto è commesso da donne, l'AI ne attribuisce il 31% al genere femminile.
Questa sovrastima potrebbe essere dovuta a diversi fattori. Da un lato, le già citate tendenze narrative e cinematografiche che influenzano l'AI potrebbero aver enfatizzato la presenza di donne in ruoli criminali, spesso presenti nelle opere creative pop come ladre sofisticate, truffatrici astute o criminali in grado di operare con eleganza, in una maniera più carismatica e accattivante rispetto alla realtà. Basti pensare a ruoli iconici nel cinema e nelle serie TV come Catwoman o Tokyo de "La Casa di Carta".

Un altro fattore potrebbe derivare dalla mitigazione dei bias nel modello AI stesso. OpenAI applica strategie per evitare squilibri di genere nelle risposte dell'AI, e questo potrebbe aver portato a una compensazione che aumenta artificialmente la presenza di donne tra i personaggi generati. In altre parole, il modello potrebbe cercare di evitare la riproduzione di uno squilibrio di genere troppo accentuato, bilanciando la distribuzione dei personaggi indipendentemente dai dati reali.
Infine, è possibile che il modello stia semplicemente ottimizzando la varietà dei personaggi per evitare ripetizioni eccessive: generare solo uomini potrebbe risultare monotono o meno interessante dal punto di vista narrativo, e quindi l'AI potrebbe tendere a diversificare i profili senza tenere conto della loro aderenza alla realtà statistica.

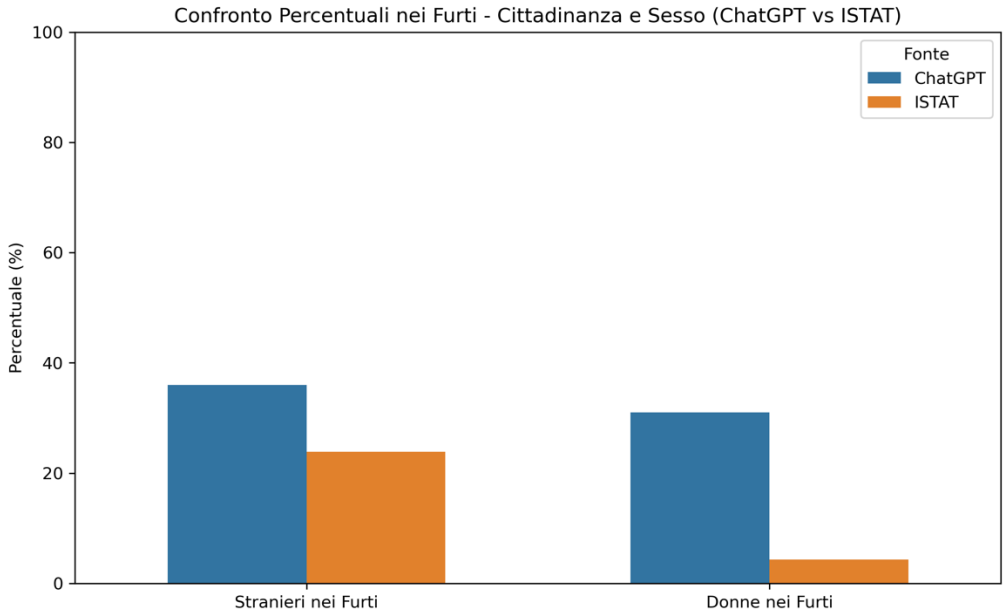

Figura 2

In Figura 3 è mostrato il confronto tra le percentuali di autori di furti d'auto suddivisi per fascia d'età. Si tratta probabilmente del caso in cui i dati generati da ChatGPT risultano maggiormente coerenti con le statistiche ISTAT. L'unica discrepanza significativa riguarda la sottostima da parte dell'AI del numero di autori di furti d'auto più giovani: la fascia 18-24 anni, che nella realtà costituisce uno dei gruppi più rappresentativi, nell'output dell'AI è significativamente ridotta; inoltre, l'AI esclude completamente i minorenni, che invece, seppur con una presenza limitata, risultano nelle statistiche ufficiali.
Questa distorsione potrebbe derivare da diversi fattori. Da un lato, ancora una volta per via del fatto che i testi, gli articoli, la narrativa su cui si basano i dataset di addestramento dell'AI, tendono a enfatizzare il furto d'auto come un'attività criminale professionale piuttosto che come un atto impulsivo legato alla microcriminalità. Il furto d'auto, in questo modo, è associato a un crimine più strutturato e organizzato, escludendo quindi i profili giovanili meno esperti.
O ancora, la presenza di filtri etici nel modello potrebbe portare ChatGPT a evitare di generare personaggi minorenni coinvolti in attività criminali per conformarsi a politiche di moderazione. È possibile che questa limitazione si estenda anche alla fascia 18-24, ma in modo meno drastico, riducendo comunque il numero di giovani criminali generati.

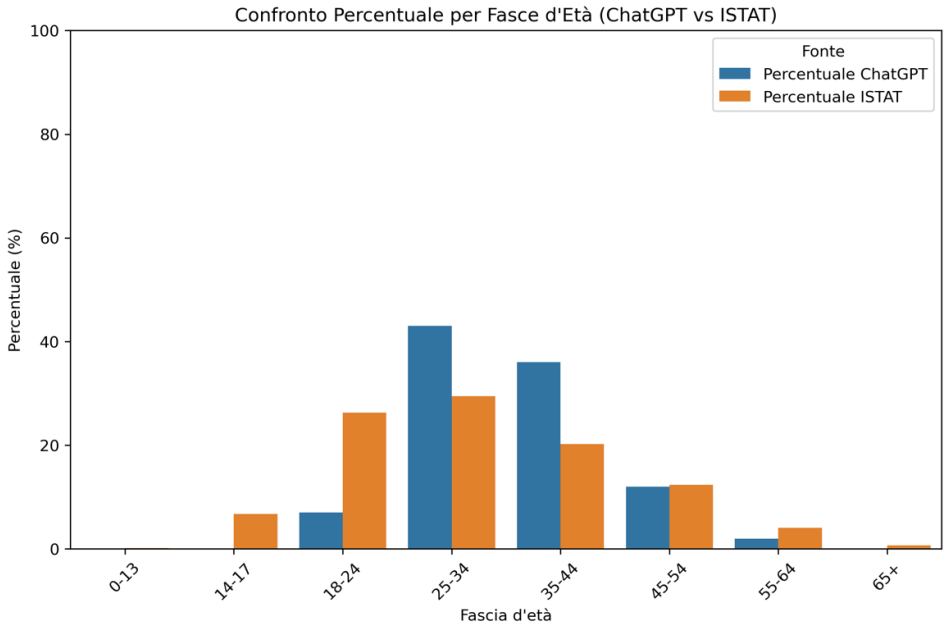

Figura 3

In Figura 4 è mostrata la percentuale di furti che riguardano auto non utilitarie. Ben il 27% dei personaggi generati da ChatGPT è coinvolto nel furto di veicoli d'epoca, di lusso o da collezione, una percentuale sovrastimata rispetto alla realtà (basti leggere il Dossier di Lojack Italia per il 2024, che rielabora i dati del Ministero dell'Interno e mostra come a essere più rubate siano principalmente auto di massa, utilitarie o al massimo SUV: https://www.lojack.it/wp-content/uploads/2024/04/CS_Furti_Dossier_LoJack_2024_final.pdf?). Questo dato può essere interpretato da un lato come un'ulteriore dimostrazione dell'influenza che l'immaginario cinematografico ha sulla generazione di contenuti da parte dell'AI, dall'altro è probabile che l'AI abbia generato questa percentuale perché i furti di auto di lusso, da collezione, o d'epoca fanno più scalpore a livello mediatico, e quindi potrebbero essere sovra-rappresentati nel modello rispetto alla loro effettiva frequenza. Tuttavia, oltre alle componenti cinematografica e mediatica, è possibile che, come nel caso della generazione del sesso del personaggio, anche qui l'AI abbia compensato artificialmente la varietà delle storie, introducendo furti più particolari per evitare ripetizioni nei profili criminali generati. Inoltre, il furto di auto di lusso può essere percepito come un crimine più "narrativamente interessante", il che potrebbe aver spinto il modello a produrre più casi di questo tipo rispetto a furti più ordinari e meno memorabili.

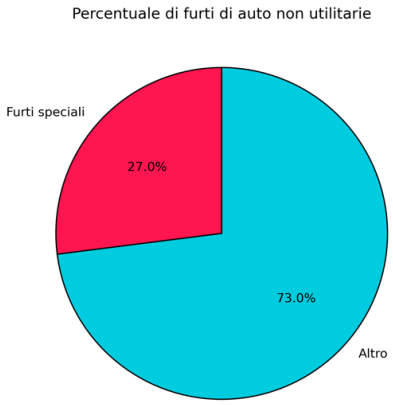

Figura 4

In Figura 5 è rappresentata la distribuzione delle motivazioni attribuite ai furti. In seguito alla richiesta di classificare entro tre etichette i cento personaggi precedentemente generati, l'AI ha prodotto le categorie "business", "adrenalina", "necessità". Il risultato è che la maggioranza dei personaggi generati da ChatGPT (64%) ruba auto per motivi legati al business, mentre solo il 17% è spinto dalla necessità: questo suggerisce che l'AI non enfatizza una correlazione diretta tra furto e disagio economico, privilegiando invece altre motivazioni. Ancora più degno di nota è il dato sull'adrenalina (19%), che sembrerebbe nuovamente confermare una tendenza dell'AI a enfatizzare aspetti narrativi e romanzeschi, più che a riflettere dinamiche strettamente legate alla realtà, nonostante nel prompt fosse stato esplicitamente richiesto di generare personaggi letterari fittizi di una narrazione realistica. Anche in questo caso, d'altronde, la distribuzione generata dall'AI potrebbe essere influenzata dalla tendenza alla diversificazione dei personaggi e alla costruzione di storie più variegate.

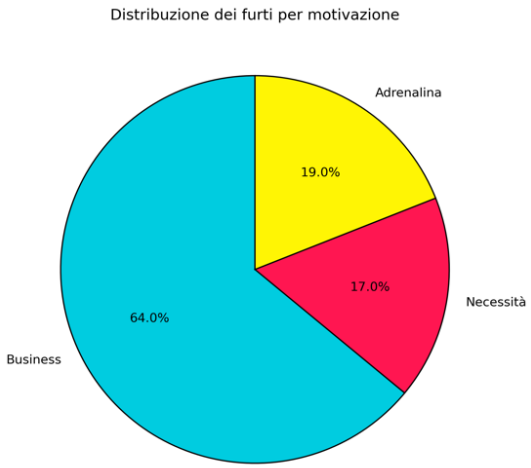

Figura 5

5. Conclusioni

L'analisi condotta ha evidenziato come l'AI generativa, nella creazione di personaggi letterari fittizi, possa riflettere schemi e tendenze che si discostano dalla realtà statistica, introducendo alcuni bias culturali e narrativi. In particolare, i risultati mostrano una sovrastima di determinate categorie, come le donne ladre di automobili, e una rappresentazione parziale delle regioni italiane, suggerendo che il modello possa essere influenzato più da fattori mediatici e narrativi che da dati oggettivi.

Le discrepanze rispetto ai dati ISTAT sembrano derivare da una combinazione di fattori, tra cui le caratteristiche dei dataset di addestramento, i filtri etici e certe convenzioni narrative. Ad esempio, l'assenza totale di minorenni tra i personaggi generati potrebbe dipendere da una policy di moderazione preimpostata del modello, mentre la presenza rilevante di ladri stranieri e il numero elevato di furti "cinematografici" potrebbero derivare da un bias riconducibile a una visione del crimine filtrata attraverso il linguaggio cinematografico e narrativo.

Un altro aspetto interessante riguarda la distribuzione delle motivazioni attribuite ai furti. Sebbene il crimine sia spesso legato a fattori economici, il modello ha generato una percentuale significativa di ladri mossi dall'adrenalina, suggerendo che l'AI possa tendere a creare profili criminali più variegati e "romanzati" piuttosto che realistici e aderenti alle statistiche.

Questi risultati mettono in luce come l'AI generativa non sia un modello puramente neutrale, ma un sistema che filtra e rielabora le informazioni seguendo determinate logiche e influenze.
Questo solleva interrogativi importanti sull'affidabilità dei modelli AI se impiegati nell'analisi o nella generazione di contenuti che riguardano fenomeni sociali o criminologici. Per mitigare questi bias, sarebbero necessari strumenti di controllo più avanzati, un miglior bilanciamento dei dataset di addestramento e un'attenzione maggiore ai filtri applicati ai modelli di generazione.
Future ricerche potrebbero ampliare lo studio confrontando i risultati ottenuti con quelli generati da altri modelli di AI generativa o esaminando il modo in cui variazioni nei prompt possano influenzare la distribuzione dei dati prodotti.