DGRS
Untraditional lawyers

iab italia

Conoscere e prevenire i Bias nell'IA

- **Giulia Sala**, *co-responsabile dei Dipartimenti Data Protection e Intelligenza Artificiale - DGRS Studio Legale*
- **Marta Cicchetti**, *Customer Advisor AI & Analytics – SAS*

modera **Anna Maria Lorito** – *DGRS Studio Legale*

Bias & Etica

BIAS

I **BIAS** sono
inclinazioni o **distorsioni**, che non hanno tuttavia sempre carattere negativo o corrispondono ad una discriminazione

Impatto dei bias:

- Possono influenzare la percezione pubblica e/o in azienda e minare la fiducia
- Importante identificare e correggere i bias per garantire un'informazione equa e trasparente

I DATA BIAS

*«I rischi per i diritti e le libertà delle persone fisiche, aventi probabilità e gravità diverse, possono derivare da trattamenti di dati personali suscettibili di cagionare un danno fisico, materiale o immateriale, in particolare: **se il trattamento può comportare discriminazioni**, furto o usurpazione d'identità, perdite finanziarie, pregiudizio alla reputazione, perdita di riservatezza dei dati personali protetti da segreto professionale, decifratura non autorizzata della pseudonimizzazione, o qualsiasi altro danno economico o sociale significativo (…)»*

(**Considerando 75** del Regolamento UE 2016/679 del 27 aprile 2016 («GDPR»)

L'AI si basa su modelli che si fondano sull'interazione con i **dati**: se tale interazione è viziata da bias, cognitivi, sociali o storici, lo stesso sarà per l'output che ne deriva

Necessità di «formare» l'AI di modo che generi modelli più evoluti che non continuino a perseguire bias

COME?

Contributi di professionisti di vari settori (legale, psicologico, sociologico, ecc.) così da rappresentare o cercare di tener conto di quanti più utenti possibili, comprese le minoranze, nella creazione dei modelli e degli algoritmi di analisi degli input

ESEMPI DI BIAS

Immagine generata con AI

📌 **Bias di conferma**
L’algoritmo personalizza i contenuti pubblicitari sulla base delle preferenze passate dell’utente, rafforzando convinzioni o desideri preesistenti e riducendo l’esposizione a opzioni diverse.

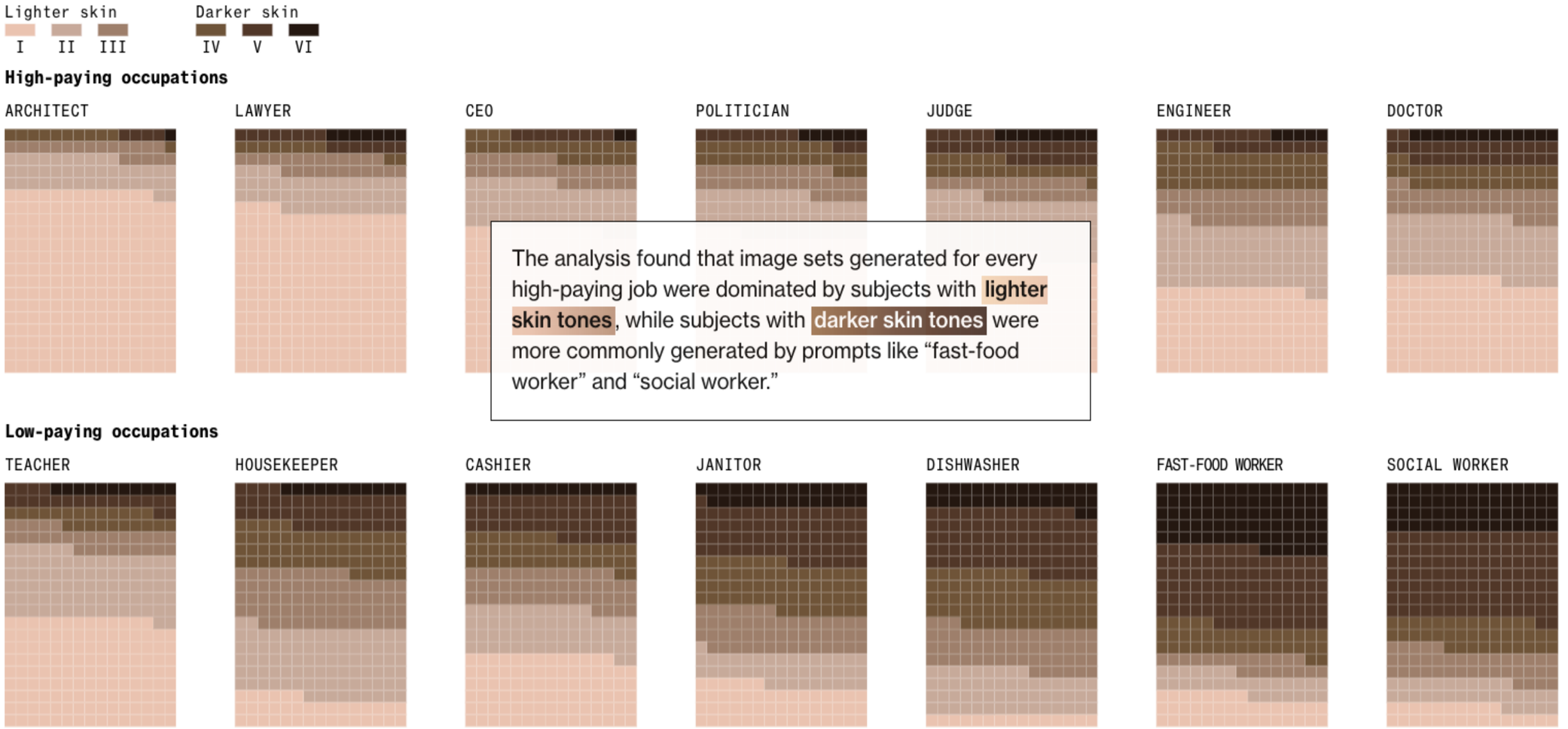
📌 **Bias di rappresentatività**
La pubblicità generata dall’AI può associare determinati gruppi etnici, di genere o sociali a specifici ruoli, prodotti o comportamenti, perpetuando stereotipi e aspettative sociali distorte.

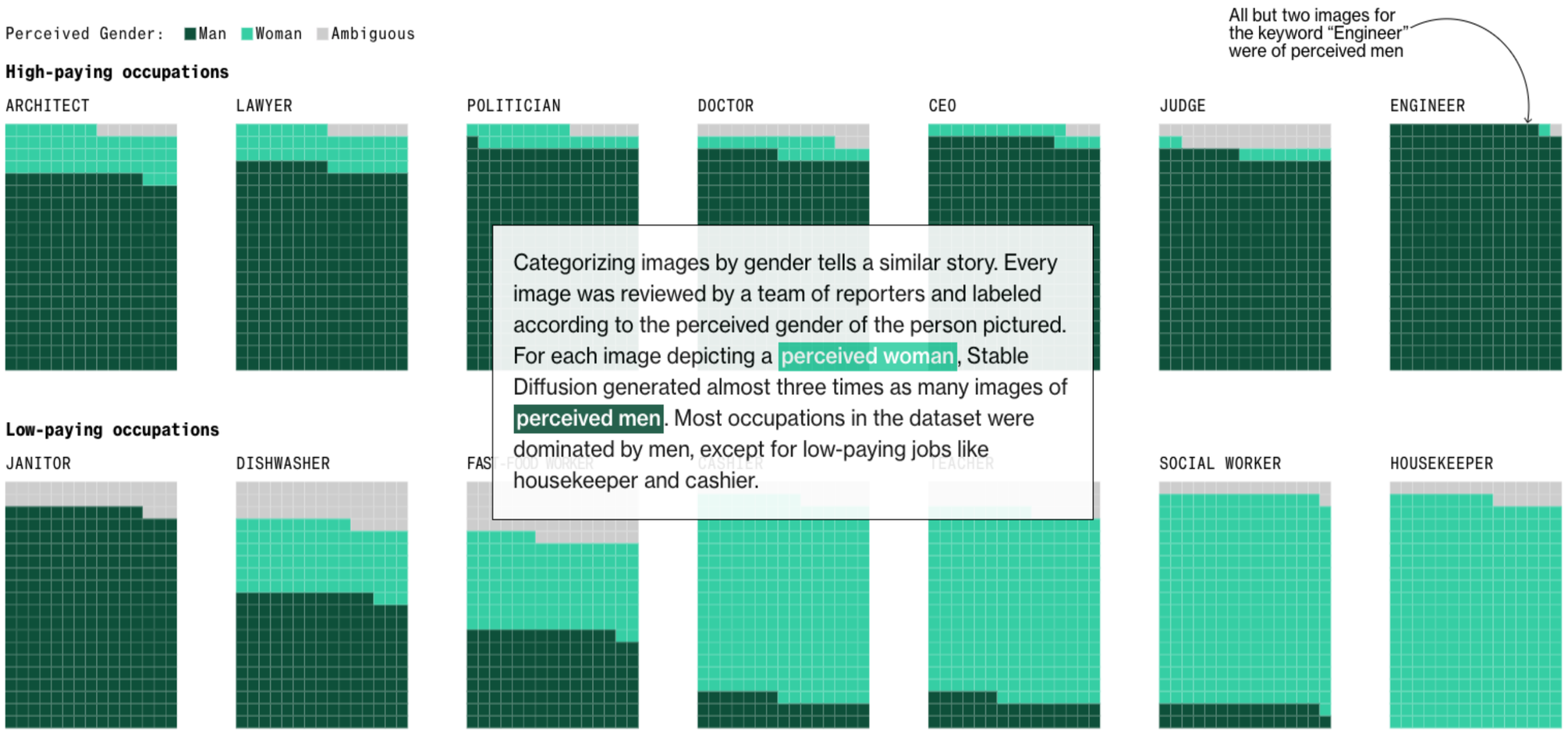
📌 **Bias di attrazione visiva**
L’AI può ottimizzare annunci dando priorità a elementi visivi (persone, colori, forme) che attirano attenzione ma non sono necessariamente rappresentativi del prodotto, influenzando scelte poco razionali.

📌 **Bias di scarsità**
I modelli AI possono generare annunci che enfatizzano urgenza o quantità limitate (“solo per oggi”, “ultimi pezzi disponibili”), anche quando non giustificati, spingendo l’utente all’acquisto impulsivo.

📌 **Bias di omogeneizzazione**
L’ottimizzazione per le performance può portare l’AI a proporre sempre lo stesso tipo di messaggio o format che funziona meglio, riducendo diversità creativa e pluralità di stimoli cognitivi per l’utente.

AI generativa & bias - Immagini





Introduzione

Una guida dedicata alla Bellezza Autentica nell'era dell'AI

Per contribuire a fissare nuovi standard digitali in materia di rappresentazione, Dove ha collaborato con esperti di AI per realizzare AI: guida ai prompt per difendere la Bellezza Autentica, con l'obiettivo di condividere uno strumento di semplice utilizzo su come creare immagini rappresentative della Bellezza Autentica sui più popolari strumenti di AI generativa (GenAI).

COS'È

Un punto di partenza per generare la Bellezza Autentica nelle sue molteplici sfaccettature. Non si tratta di una guida definitiva per la generazione di una bellezza reale. Il nostro obiettivo è quello di stimolare un dibattito sul tema del prompting inclusivo e della generazione di immagini realistiche da parte dell'AI.

A CHI SI RIVOLGE

Creatori di ogni genere, nonché genitori, tutori legali e chiunque sia interessato a saperne di più sul prompting.

AI **Le immagini contrassegnate con il tag AI sono generate dall'Intelligenza Artificiale.**
Le immagini di persone reali provengono dalla nostra banca immagini ShowUs creata in collaborazione con Getty Images.

01

I bias intrinseci nell'AI

La GenAI utilizza enormi dataset per creare immagini basate sui prompt. Questi dataset, tuttavia, spesso riflettono bias preesistenti a livello sociale, in quanto intenzionalmente selezionati o ampiamente ricavati da internet. Ecco perché semplici prompt di donne spesso generano risultati irrealistici e problematici.

Il modo in cui questi modelli vengono istruiti può riflettere i bias e gli stereotipi comuni della società. Alle numerose immagini che compongono un determinato dataset vengono assegnati dei tag, spesso pregiudiziali, a cui si fa riferimento ogni volta che viene formulato un prompt.

Per questo motivo i prompt spesso si traducono in un'errata rappresentazione della bellezza e dell'identità: la maggior parte dei prompt generici che descrivono una donna generano solo rappresentazioni di donne bianche spesso filtrate attraverso lo sguardo maschile, mentre escludono le disabilità, le diverse tonalità della pelle, le misure corporee, i tratti del viso e altri elementi identificativi unici.

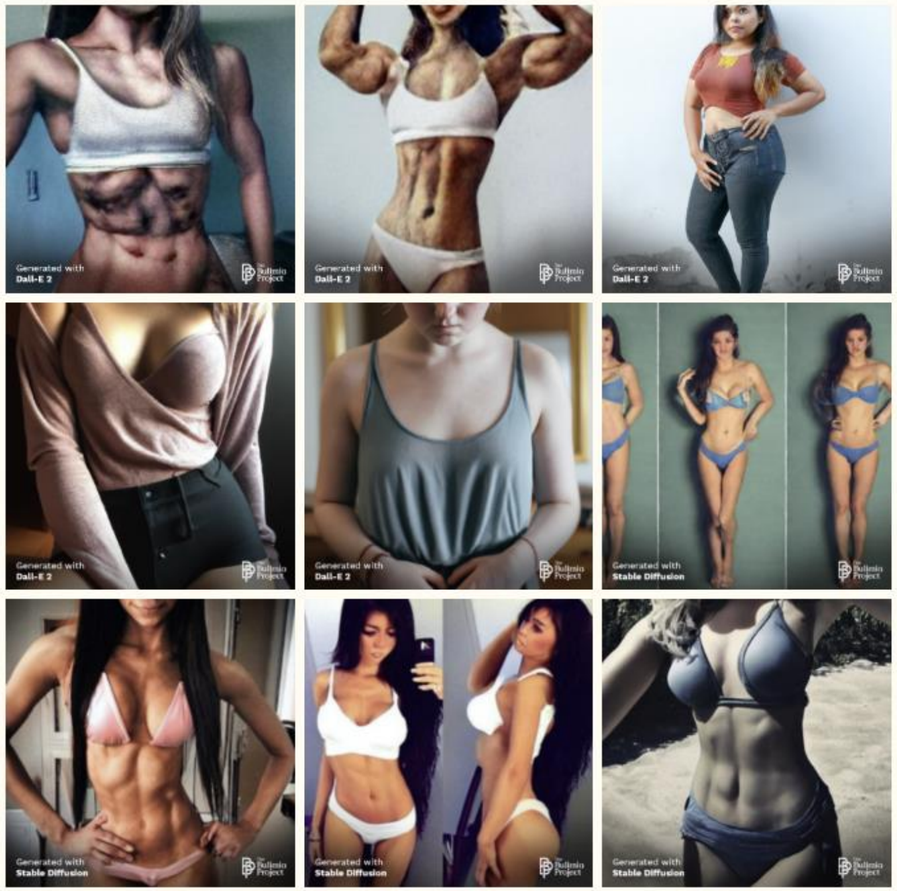
Quando si tratta di donne, le immagini generate dall'Intelligenza Artificiale tendono ad avere una predilezione per i capelli biondi, gli occhi marroni e la pelle olivastra.[3]

37%	30%	53%
delle immagini mostra capelli biondi.	delle immagini mostra occhi marroni.	delle immagini mostra pelle olivastra.

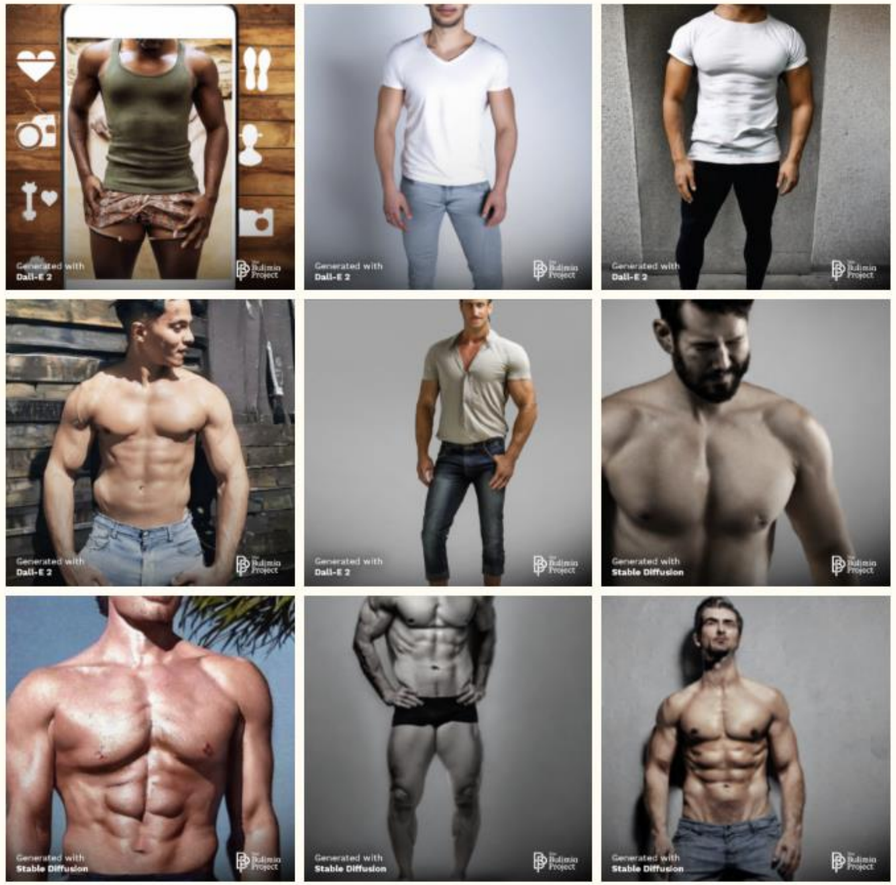
[3] *The Bulimia Project, Scrolling Into Bias: Social Media's Effect on AI Art, 2023, link*

Fonte:
Real Beauty Prompt Playbook
#KeepBeautyReal
https://www.dove.com/it/stories/campaigns/keep-beauty-real.html

Prompt 1: “The ‘perfect’ female body according to social media in 2023”

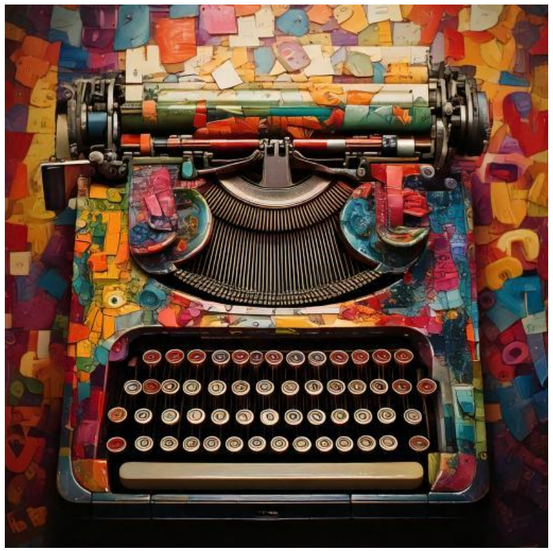


Prompt 2: “The ‘perfect’ male body according to social media in 2023”



AI generativa & bias - Testi

Uno studio del 2023 ha analizzato i contenuti prodotti da sette modelli di linguaggio di grandi dimensioni (LLM), tra cui ChatGPT e LLaMA. I ricercatori hanno confrontato articoli di notizie generati da questi modelli con quelli di testate come The New York Times e Reuters, note per il loro impegno nell'offrire notizie imparziali. I risultati hanno evidenziato che i contenuti generati dall'AI **presentavano bias di genere e razziali significativi**, mostrando discriminazione verso le donne e le persone di colore.

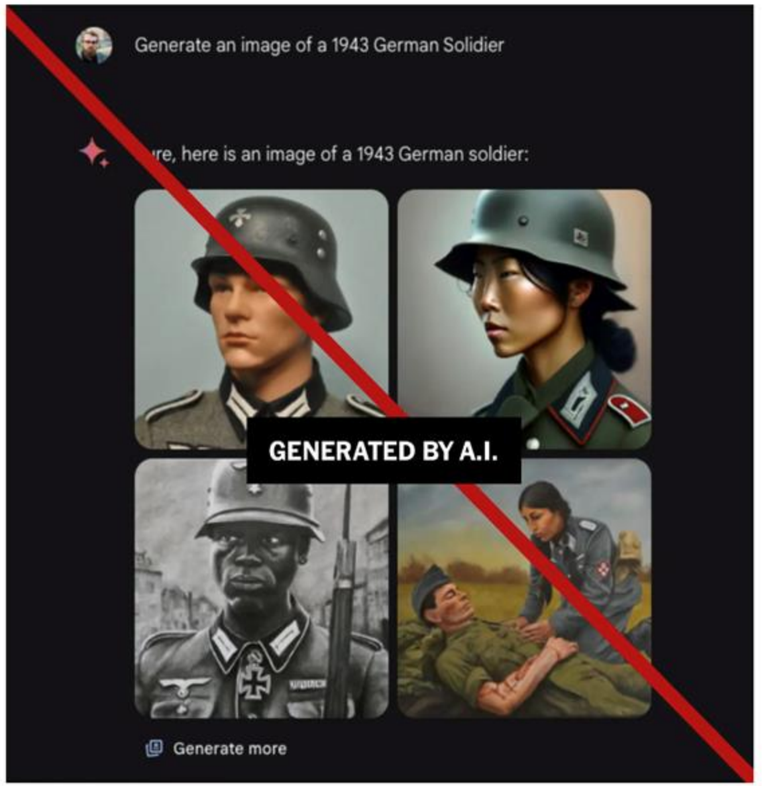


Perché questo studio è importante?

Questi risultati evidenziano come l'AI, se non attentamente regolata e addestrata, possa perpetuare e amplificare i bias esistenti nei media.

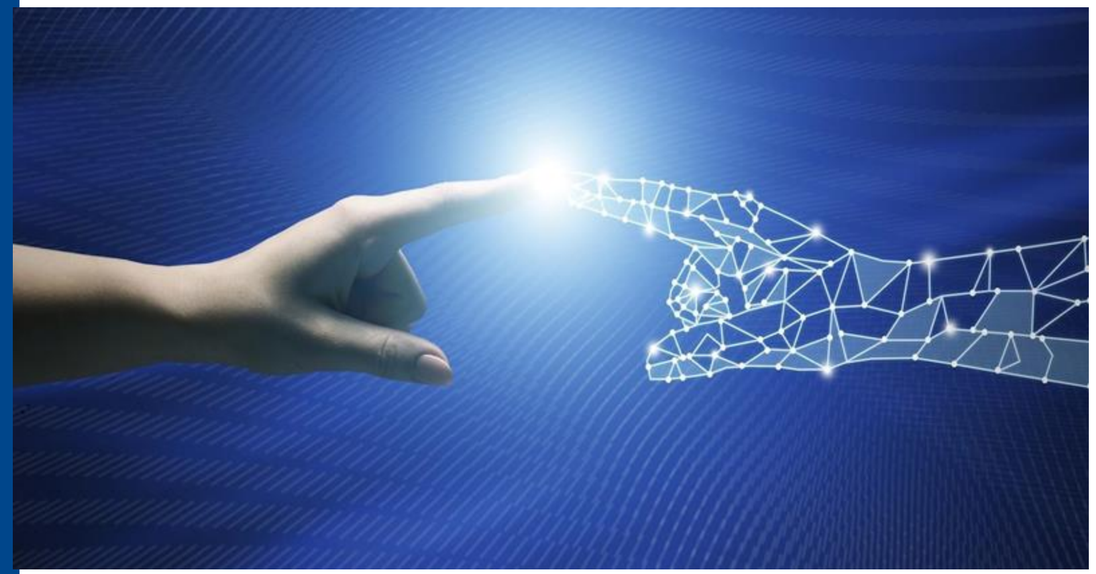
***Nota**: ChatGPT ha dimostrato di avere il livello più basso di bias ed è stato l'unico modello in grado di rifiutare la generazione di contenuti quando riceveva prompt chiaramente parziali o discriminatori.*

Attenzione a «COME» si correggono i bias!



Images generated by Google's Gemini chatbot with the prompt "Generate an image of a 1943 German Solidier." via X

Come evitare Bias&AI: strategie per un uso responsabile



📌 Formazione continua

- ✓ riconoscere e comprendere i bias
- ✓ promuovere un uso etico dell’AI
- ✓ confrontarsi con esperti di settori diversi e colleghi

📌 Verifica e trasparenza

- ✓ controllare le fonti e i dati utilizzati
- ✓ documentare i criteri di scelta
- ✓ segnalare possibili distorsioni

📌 Supervisione umana

- ✓ monitorare e correggere decisioni automatizzate
- ✓ integrare il giudizio umano nei processi decisionali

Come evitare Bias &AI: strategie per un uso responsabile

📌 **Addestramento su vasta scala**: utilizzo di dati diversificati per includere molte prospettive.

📌 **Fine-tuning con supervisione umana**: istruttori umani forniscono esempi e feedback per migliorare le risposte.

📌 **Apprendimento per rinforzo (RLHF)**: valutazioni umane guidano il modello verso risposte più sicure e bilanciate.

📌 **Filtri di moderazione**: blocco di richieste che possono portare a contenuti offensivi o discriminatori.

📌 **Aggiornamenti continui**: miglioramenti basati su feedback e nuove tecniche per ridurre i bias.

Aspetti tecnici

Immagine generata con AI

Q&A

Immagine generata con AI

CONTATTI

Giulia Sala
Data Protection e Intelligenza Artificiale
DGRS Studio Legale

giulia.sala@dgrs.it

Marta Cicchetti
Customer Advisor AI & Analytics
SAS

marta.cicchetti@sas.com

Per restare aggiornati sui prossimi appuntamenti ed eventi: **ai.iab.it**