Pedagogia oggi

Rappresentazione di generi e sessualità nell'era delle GenIA: approcci pedagogici alla mediatizzazione dell'immaginario

Representations of gender and sexuality in the age of GenAI: pedagogical approaches to the mediatization of the collective imaginary

Salvatore Messina
Researcher of Didactics and Special Pedagogy, Department of Education "Giovanni Maria Bertin", Alma Mater Studiorum University of Bologna, salvatore.messina10@unibo.it

OPEN ACCESS

Siped Società Italiana di Pedagogia

Double blind peer review

Citation: Messina, S. (2025). Representations of gender and sexuality in the age of GenAI: pedagogical approaches to the mediatization of the collective imaginary. *Pedagogia oggi*, *23*(2), 217-225 https://doi.org/10.7346/PO-022025-26

Journal Homepage
https://ojs.pensamultimedia.it/index.php/siped

Pensa MultiMedia / ISSN 2611-6561
https://doi10.7346/PO-022025-26

ABSTRACT

This study examines cultural biases in generative AI, with a particular focus on representations of gender and sexuality. Analyses of interactions with chatbots conducted through six university workshops (N=245) produced 1,820 traces, which constitute the analytical corpus used to measure bias levels before and after the drafting and finalization of documents, as well as to monitor process indicators. Bias decreased across all categories, particularly with regard to gender (31.8% 16.3%) and certain cultural representations (22.0%→12.7%). To correct bias, groups employed an average of four prompts; a higher number of prompts was positively associated with residual bias (OR=1.42), revealing what we define as the fatigue paradox, with satisfaction with the final output negatively correlated with the number of iterations performed (ρ≈−.16). Qualitative analysis shows frequent lexical clean-ups but limited structural reallocation of *agency*; explicit mentions of trans* or non-binary identities are rare (≈1.2%).

Il presente studio esamina i pregiudizi culturali nell'IA generativa, con particolare attenzione alle rappresentazioni di generi e sessualità. L'analisi di interazione con i *chatbot* all'interno di sei laboratori universitari (N=245) ha prodotto 1.820 tracce che costituiscono il corpus di analisi su cui abbiamo misurato il pregiudizio pre/post stesura di bozze e documento finale, nonché monitorato gli indicatori di processo. Il pregiudizio è diminuito in tutte le categorie, in particolare per quanto riguarda il *genere* (31,8%→16,3%) e alcune rappresentazioni culturali (22,0% 12,7%). Per la correzione dei *bias* i gruppi hanno utilizzato circa 4 *prompt*; un numero più elevato di *prompt* era associato positivamente al *bias* residuo (OR=1,42) facendo emergere quello che definiamo *paradosso della fatica*, con soddisfazione output finale correlata negativamente alle iterazioni effettuate (ρ≈−.16). L'analisi qualitativa mostra frequenti pulizie lessicali ma una riallocazione strutturale limitata dell'*agency*; le menzioni esplicite trans*/non binarie sono rare (≈1,2%).

Keywords: AI literacy, mediatization, cultural bias, LGBTQIA+ representation, prompt

Parole chiave: alfabetizzazione all'IA, mediatizzazione, bias culturali, rappresentazioni LGBTQIA+, prompting

Received: September 12, 2025
Accepted: November 24, 2025
Published: December 30, 2025

Corresponding Author:
Salvatore Messina, salvatore.messina10@unibo.it.it

Introduzione allo studio

Negli ultimi anni i modelli di Intelligenza Artificiale generativa (GenIA) sono divenuti mediatori potenti dei processi di produzione culturale, incidendo tanto sulle pratiche comunicative quotidiane quanto sui dispositivi formativi. La letteratura sui processi di *mediatizzazione* aiuta a collocare tale fenomeno in una prospettiva socioculturale: i *mediascapes* globali (Appadurai, 1996) e gli *immaginari sociali* (Taylor, 2004) si intrecciano con logiche algoritmiche che, secondo le teorie della mediatizzazione (Hjarvard, 2013) e della *deep mediatization* (Hepp, 2020), ristrutturano ambienti, pratiche e rappresentazioni. In questo quadro, le GenIA non si limitano a "riflettere" il sociale, ma a co-produrre immaginari collettivi, stabilizzando o perturbando categorie identitarie attraverso output che presentano apparenti neutralità pur incorporando scelte e stereotipi culturali.

Il tema dei *bias* è perciò centrale. Gli studi sui *data* e *model biases* in ambito educativo mostrano come processi di raccolta, selezione e pesatura dei dati, insieme alle metriche di ottimizzazione, possano generare distorsioni sistematiche (Noble, 2018; Benjamin, 2019). Il rischio non è solo la riproduzione di stereotipi "classici" (genere, etnici, classe, …), ma anche forme più sottili di *falsa neutralità*, ossia di un linguaggio apparentemente equilibrato che de-problematizza asimmetrie e invisibilizza gruppi minoritarizzati. Nelle rappresentazioni delle persone LGBTQIA+, ciò può tradursi in *eteronormatività implicita*, *cancellazione* di identità trans* e non binarie, *misgendering* lessicale e narrazioni pseudo-pedagogiche che confinano le soggettività *queer* a funzioni ghettizzanti o didascaliche anziché riconoscerne *agency* e complessità individuali. Evidenze recenti nel campo dell'NLP segnalano criticità nella classificazione e nel trattamento del linguaggio *queer* e non standard, con effetti su emozioni, sicurezza e monitoraggio dei contenuti.

In ambito educativo, tali problemi si amplificano (Noble, 2018; Mascitti, Messina, Panciroli, 2025). Da un lato, scuole e università sperimentano GenIA per la scrittura, il riassunto, la riformulazione e la progettazione didattica; dall'altro, emerge l'urgenza di AI *Literacy* come insieme di competenze tecnico-critiche ed etiche (UNESCO, 2024; Panciroli, 2025). L'attenzione alla trasparenza, all'*explainability* e alla responsabilità è inoltre coerente con il quadro regolativo europeo più recente, che incentiva pratiche documentate di *risk management*, valutazioni d'impatto e tutela dei diritti fondamentali in settori sensibili come l'educazione (EU, 2024). Per l* docent*, ciò implica integrare nei curricoli percorsi strutturati di analisi critica dei modelli, audit di *bias* e strategie di *prompt-engineering* orientate all'equità sociale; per l* student*, significa sviluppare competenze per riconoscere, mitigare e rendicontare le distorsioni, sapendo quando e come sollecitare *revisioni iterative* degli output.

Nonostante la crescita di contributi concettuali, mancano ancora indagini empiriche sistematiche che osservino in didattica come docenti e student* interagiscano con *chatbot* generativi, quali bias emergono in compiti di scrittura creativa/argomentativa e quanto i cicli di *re-prompting* (Mascitti, Messina, Panciroli, 2025) riducano l'incidenza delle distorsioni, in particolare sulle rappresentazioni di generi e sessualità. D'altro canto, almeno sul piano esperienziale, i cicli di *re-prompting* e la revisione collaborativa possono produrre miglioramenti misurabili, spostando i testi da un regime di invisibilità o stereotipia a forme più pluraliste e situate, purché si sostenga l'azione con strumenti culturali e competenze di *AI literacy* adeguate.

Inoltre, poca ricerca ha confrontato metriche comportamentali (es. numero di tentativi, tempo, strategie di revisione) con esiti percepiti (soddisfazione, qualità) e con analisi qualitative delle trasformazioni testuali e audio-visive tra le diverse bozze e la versione finale di proposte didattiche realizzate con il supporto di *chatbot* di GenIA. Il presente studio si inserisce in questa direzione, osservando in situazione didattica l'impatto dei cicli di *re-prompting* sulle trasformazioni testuali e sulle percezioni qualitative.

1. Dati, modelli, scelte: una cornice di competenze per l'uso didattico dell'IA

Nel quadro della media education, l'alfabetizzazione all'IA non coincide con abilità d'interfaccia: è una competenza culturale e professionale che integra conoscenze, sensibilità etiche e capacità progettuali (Buckingham, 2003, 2019; Hobbs, 2021; Rivoltella, 2024). In questa prospettiva, AI *literacy* e *data literacy* operano come assi trasversali: comprendere come funzionano i sistemi e come si istruiscono (*prompt*, vincoli, feedback), valutare criticamente gli output e progettarne l'impiego; conoscere la provenienza dei dati,

le assunzioni algoritmiche e i loro effetti sulle rappresentazioni sociali (UNESCO, 2024; Panciroli, Rivoltella, 2023). Le dimensioni alfabetica, critica, etica ed espressivo-progettuale offrono un lessico operativo per leggere processi ed esiti delle attività mediate con gli LLM (Elliott, 2021; Panciroli, Rivoltella, 2023). In quest'ottica, i nostri indicatori (RQ1–RQ4) non registrano solo "performance", ma operalizzano competenze: rilevare/ridurre *bias*, progettare vincoli inclusivi, documentare scelte e iterazioni come parte dell'*accountability* didattica.

2. La ricerca

La ricerca qui presentata intende contribuire a colmare alcune delle questioni sopra presentate integrando una cornice teorica sulla *mediatizzazione dell'immaginario* (Appadurai, 1996; Taylor, 2004; Hjarvard, 2013; Hepp, 2020) con un disegno a metodi misti implementato durante il secondo semestre dell'anno accademico 2024/2025 in sei laboratori universitari (N=245) presso l'Università di Bologna, nei corsi triennali di Scienze dell'Educazione e del Dipartimento delle Arti. L'intervento didattico ha previsto la produzione di testi mediata da *chatbot* di IA (sceneggiature brevi per storytelling) e una successiva fase di riflessione metacognitiva.

Sono quattro le domande di ricerca (RQ) che orientano lo studio e relativa rilevazione dei dati quali-quantitativi:

- RQ1 (Prevalenza e tipologia): quali e quanti tipi di bias emergono con maggiore frequenza nelle bozze iniziali, con particolare attenzione alle rappresentazioni di generi e sessualità LGBTQIA+?
- RQ2 (Mitigazione): in che misura i cicli iterativi di *re-prompting* e revisione collaborativa riducono l'incidenza dei bias dalla bozza alla versione finale, e quali pattern di mitigazione si osservano tra le diverse tipologie?
- RQ3 (Processo e percezioni): qual è la relazione tra numero di *prompt* e percezione soggettiva di qualità/soddisfazione?
- RQ4 (Qualità narrativa e *agency*): in che modo le strategie linguistiche (pronomi, marcatori identitari, descrizioni di ruolo) e scelte narrative influenzano l'*agency* dei personaggi minoritari (es. soggetti trans* e non binari), e come tali scelte variano tra bozza e versione finale?

Lo strumento di rilevazione somministrato durante i laboratori di *storytelling*[1] è anche un form di accompagnamento; con la sua funzione di *scaffolding* di processo ha: (i) guidato ciascun utente nella progettazione del *prompt* (tema, contesto, personaggi, stile e vincoli inclusivi); (ii) registrato in sequenza risposte del *chatbot* e *re-prompting* fino alla versione finale; (iii) tracciato per ogni iterazione il numero di *prompt* e tipologie di revisione.

I gruppi hanno ricevuto un inquadramento iniziale su equità sociale, lessico inclusivo e rappresentazioni non stereotipate, *agency* dei personaggi; non è stata fornita formazione tecnica sul *prompt* engineering. Di conseguenza, le iterazioni di *re-prompting* sono da considerarsi "non esperte" (pratiche spontanee), e i *pattern* osservati riflettono l'uso ordinario dell'interfaccia più che competenze specialistiche.

Per ciascun ciclo venivano annotati il numero di tentativi (*prompt count*) e le revisioni effettuate; al termine l* student* indicavano e quantificavano i *bias* individuati nella bozza iniziale e nella versione, insieme a una valutazione di soddisfazione delle interazioni. Tali tracce processuali hanno permesso di calcolare indicatori pre/post e di integrare l'analisi quantitativa con la codifica qualitativa dei testi. La componente qualitativa consiste nell'analisi tematica di 1.820 *prompt* e risposte con codifica inter-coder, focalizzata su tipologie di *bias* culturali (genere, etnico, di classe, linguistico, cfr. *nota metodologica*), intensità e traiettorie di trasformazione tra le diverse versioni elaborate dai *chatbot*. Si predilige la centralità del *genere* come dimensione critica perché coerente con studi che evidenziano *bias* di ruolo nelle rappresentazioni linguistiche,

1 Lo strumento di rilevazione è stato progettato in collaborazione con Elisa Farinacci (Dipartimento delle Arti – DAR, Università di Bologna), docente e conduttrice del laboratorio "*Linguaggi della televisione e dei media digitali*", contesto nel quale, nell'a.a. 2024/2025, 62 corsisti hanno partecipato alla presente ricerca (≈25,3% del campione totale, *N* = 245).

default ciseteronormativi e criticità sul linguaggio *queer* (Vásquez *et alii*, 2022; Queer in AI, 2025), nonché difficoltà nel trattare pronomi non binari e identità trans* senza misgendering (Dev *et alii*, 2021; Lauscher *et alii*, 2023; Hossain, Dev, Singh, 2023).

Dal punto di vista metodologico, il contributo combina: (a) metriche di esito (ricorrenze, riduzioni assolute o relative) e di processo (conteggio *prompt* utilizzati dagli utenti, tipologia di correzioni apportate dai *chatbot*); (b) modelli statistici per pre-post e predittori della soddisfazione dei singoli utenti; (c) analisi qualitativa con *codebook* esplicito e verifica di affidabilità *inter-coder* (codifica esperta doppio cieco per calcolo $\varkappa$ di Cohen e α di Krippendorff). Questo consente di triangolare risultati e di proporre indicatori operativi trasferibili in altri contesti, in coerenza con approcci di AI *literacy* e con i principi di una pedagogia equa (Messina, 2025; Panciroli, 2025).

I *chatbot* utilizzati risultano essere *Chatgpt* v.4o e *Gemini 2.5 flash*.

Sul piano pedagogico-curricolare, i risultati alimentano almeno tre fronti:

– la progettazione di compiti autentici con GenAI che includano vincoli di equità sociale (sia linguistici che narrativi) e prevedano cicli riflessivi;
– lo sviluppo di strumenti di audit (checklist, rubriche) per rilevare e mitigare *bias* prima della consegna del prodotto;
– l'integrazione curricolare di competenze etiche e critiche su *dataset*, *metriche*, *safety* e *accountability*, allineate ai riferimenti internazionali e al quadro europeo. In tale senso, l'educazione all'IA non è un'aggiunta tematica ma una leva trasversale per ripensare *agency*, *valori*, *linguaggi* e *immaginari* nella scuola e nell'università.

Lo studio, quindi, mira a quantificare la prevalenza e la trasformazione dei *bias* tra bozza e versione finale, a stimare l'efficacia del *re-prompting* e a descrivere, sul piano qualitativo, come le scelte linguistiche e narrative impattino l'*agency* di personaggi minoritari, con particolare riguardo a identità trans* e non binarie.

2.1 *Nota metodologica*

In questa ricerca la macrocategoria culturale è trattata in due passaggi complementari:

– codifica iniziale con regola di precedenza: quando i *prompt* o le risposte dei *chatbot* presentavano marcatori specifici di genere, etnico, di classe o linguistico, si è attribuita priorità a tali codici. La voce culturale è stata impiegata solo quando l'evidenza non era imputabile in prima istanza alle categorie suddette, ma rimandava a cornici valoriali o simboliche (religione e moralismi; giudizi su tradizione/modernità; stereotipi generazionali; gerarchie coloniali/eurocentriche prive di marcatori etnici o nazionali espliciti);
– co-occorrenze: se un passaggio esprimeva più dimensioni (per es., un frame valoriale insieme a varietà linguistica o origine), sono stati applicati più codici. Ai fini delle prevalenze si è conteggiato il codice primario più specifico; culturale poteva essere aggiunto come metacodice descrittivo, ma non rientrava nei conteggi primari;
– ricodifica: rilevata l'ampiezza non meramente residuale della dimensione culturale, tutte le unità originariamente etichettate come *culturale* (bozza e finale) sono state ricodificate in quattro sottotipi mutuamente esclusivi: (i) religioso/simbolico, (ii) tradizione/modernità, (iii) generazionale, (iv) coloniale/eurocentrico. Per tali sottocategorie l'affidabilità è stata verificata su un campione ≥ 20–30% con doppia codifica indipendente e stima della concordanza ($\varkappa$ di Cohen), con risoluzione dei disaccordi per consenso (*adjudication*). Razionale della doppia tassonomia: mantenere sia il dato aggregato sia la scomposizione consente di: (a) preservare la comparabilità con le altre tipologie nei risultati principali e (b) localizzare con maggiore nitidezza la natura del *bias* culturale.

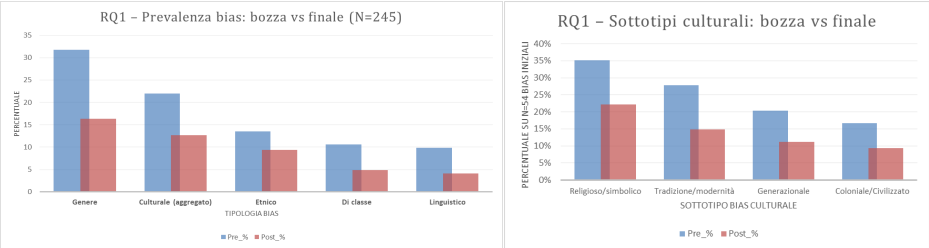
3. Principali risultati

Nel campione (N=245) in relazione alla RQ1 (*Prevalenza e tipologia*) il *bias* più frequente è quello di genere, presente nella prima stesura nel 31,8% dei casi (78/245) e nella versione finale nel 16,3% (40/245), con una riduzione assoluta di 15,5 punti percentuali (riduzione relativa ≈ –48,7%). L'ordine di frequenza delle altre tipologie rimane stabile tra pre e post: culturale (aggregato) ≈22,0%→12,7% (≈54→31 casi), etnico ≈13,5%→9,4% (≈33→23), di classe ≈ 10,6% → 4,9% (≈ 26 → 12), linguistico ≈ 9,8%→4,1% (≈24→10).

I risultati relativi alla macrocategoria di *bias* culturale (54 31 casi complessivi) scomposta in quattro sottotipi mutuamente esclusivi mostrano una riduzione omogenea tra la bozza e la versione finale. Sul totale del campione (N=245), i valori assoluti e percentuali sono i seguenti:

- *tradizione/modernità*: da 15 a 8 casi (6,1%→3,3%; Δ=–2,8 p.p.), con quota interna 27,8% → 25,8%;
- *generazionale*: da 11 a 6 casi (4,5%→2,5%; Δ = –2,0 p.p.), con quota interna 20,4% → 19,4%;
- *coloniale/eurocentrico*: da 9 a 5 casi (3,7%→2,0%; Δ = –1,6 p.p.), con quota interna 16,7%→16,1%;
- *religioso/simbolico*: da 19 a 12 casi (7,8%→4,9%; Δ = –2,9 p.p.); all'interno di questa macrocategoria, i sottotipi non calano in modo uniforme: i casi religiosi/simbolici si riducono meno rispetto agli altri, facendo crescere la loro incidenza relativa (dal 35,2% al 38,7%) pur in presenza di un calo assoluto (19/54 12/31).

In altri assoluti, la quota relativa dei sottotipi rimane pressoché stabile, con oscillazioni contenute: i *bias tradizione/modernità* e *generazionale* calano leggermente nel peso interno, *coloniale/eurocentrico* resta quasi invariato, mentre quello *religioso/simbolico* incrementa la propria incidenza percentuale. Questo significa che, pur riducendosi in valori assoluti, le rappresentazioni *simboliche/religiose* tendono a mantenere un ruolo centrale all'interno della categoria culturale, segnalando una resilienza narrativa di tali frame, difficili da scardinare anche nei cicli di *re-prompting*.



Graf. 1: *Confronto tra bozza e versione finale (N = 245): a sinistra la variazione dei macro-tipi di bias, a destra la distribuzione dei sottotipi culturali*

In termini di *mitigazione del bias* (RQ2), il confronto appaiato pre/post mediante test di McNemar mostra una riduzione statisticamente significativa dei *bias* di genere, di classe, linguistico e culturale aggregato ($p < .05$ in ciascun caso); per l'etnico, la diminuzione non raggiunge la significatività ($p \approx .15$), in linea con quanto emerso nella disaggregazione del *bias* culturale in RQ1.

Sul versante dei processi, i gruppi hanno impiegato in media 3,75 *prompt*[2] (mediana =4; *range* =1–6; n =169) per arrivare alla versione finale. E qui emerge un dato che colpisce: quando nella versione finale resta almeno un *bias*, il numero di *prompt* è più alto (media 4,28 con *bias* finale vs 3,51 senza; $t \approx -2,90$; $p \approx 0,0047$).

Questo *paradosso della fatica* è confermato dal modello di regressione logistica con esito *any bias* finale:

2 Adottiamo la media aritmetica come statistico di sintesi dei conteggi di prompt in quanto ci permette di descrivere l'andamento complessivo del campione pur senza rappresentare un valore osservabile nei singoli casi. Per rispettare la natura discreta dei dati si riportano inoltre mediana e range come indicatori complementari.

il numero di *prompt* è positivamente associato alla probabilità di *bias residuo* (OR per +1 *prompt* =1,42, IC95% 1,22–1,65; p≈0,005), mentre la sola presenza di qualunque *bias* iniziale non risulta predittiva quando si controlla per le iterazioni ($p \approx 0{,}32$). In altri termini, nonostante l'impegno e le revisioni, i casi che hanno "lavorato" di più mostrano più spesso un *bias* residuo: un esito che interroga la qualità delle istruzioni nei cicli di *re-prompting* (attenzione: si tratta di un risultato associativo; il disegno di ricerca, infatti, non consente inferenze causali).

In termini di *soddisfazione complessiva* (QR3) si attesta su una media di soddisfazione rispetto alla qualità della stesura finale ≈ 2,6/4 (scala 1–4). La correlazione (*Spearman*) tra numero di *prompt* e soddisfazione è negativa ($\rho \approx -0{,}16$), indicando che a un maggior numero di iterazioni si associa una valutazione più bassa del prodotto; il coefficiente risulta comunque statisticamente diverso da zero (p≈0,035). Nel confronto tra gruppi dei diversi laboratori, la soddisfazione finale è sempre più alta in assenza di *bias* iniziali (≈ 2,73/4) rispetto ai gruppi con *bias* iniziali (≈ 2,38/4; $p \approx 0{,}0005$).

Per ciò che attiene la *Qualità narrativa e l'agency* (QR4) l'analisi dei 1.820[3] *prompt* e relative risposte evidenzia tre ricorrenze principali nella traiettoria bozze versioni finali:

– *stereotipi di ruolo nelle descrizioni dei personaggi* (associazioni femminili a cura/emotività; maschili ad azione/leadership), che tendono a ridursi sul piano lessicale ma risultano più stabili nell'assegnazione dei ruoli;
– *default ciseteronormativi nella costruzione di relazioni e famiglie* quando non sono fornite specifiche vincolanti nel *prompt*, con attenuazioni laddove i *prompt* successivi introducono vincoli espliciti;
– *criticità nell'uso di pronomi inclusivi* e nella resa di identità non conformi al binarismo; le *identità trans* e non binarie* risultano poco rappresentate nelle risposte aperte (≈3/245, ≈1,2% di menzioni esplicite).

Nei casi presenti, quindi, si osservano omissioni di pronomi adeguati o normalizzazioni binarie (es. riclassificazione dei personaggi entro il binomio uomo/donna); lungo i cicli di revisione si rileva più frequentemente un miglioramento del registro linguistico (eliminazione di marcatori esplicitamente problematici) rispetto a riprogettazioni strutturali dei ruoli o all'adozione sistematica di pronomi non binari.

Complessivamente, l'assetto dei dati per *RQ4* documenta che, nelle trasformazioni testuali, gli aggiustamenti lessicali sono più ricorrenti delle ricomposizioni di *agency* e che l'inclusione esplicita di soggettività trans*/NB rimane residuale nel campione analizzato.

4. Iterare non è progettare: esiti, limiti e "paradosso della fatica"

I risultati emersi (riduzioni dei *bias* tra bozza e versioni finali delle sceneggiature in tutte le categorie, associazione positiva fra numero di iterazioni e persistenza di *bias* residui, scarsa visibilità di identità trans* e non binarie, prevalenza di aggiustamenti lessicali rispetto a riallocazioni strutturali dei ruoli) sono coerenti con una lettura in chiave di mediatizzazione.

In questo quadro, i *chatbot* generativi non "specchiano" il reale ma lo mediano: producono testi come sintesi probabilistiche di distribuzioni linguistiche pregresse, massimizzando la plausibilità statistica e dunque privilegiando cornici ad alta frequenza (ciseteronormatività, divisione di genere dei ruoli, gerarchie culturali) a meno che vincoli espliciti non le disinneschino (es. *prompt*).

Questo significa che bisogna operare su più livelli pedagogici:

– *progettazione dei prompt* per passare da un uso reattivo dei *prompt* (correzioni a valle) a un uso architetturale a monte: dichiarare vincoli inclusivi (pronomi, ruoli, rappresentazioni non stereotipate), prevedere obblighi di giustificazione ("*spiega dove potrei incorrere in stereotipi e proponi alternativi*"), inserire esempi

3 Dati aggregati prompt–output; una quota di *bias* proviene dal prompting degli utenti. Trattandosi di ricorrenze <5%, il corpus è stato mantenuto e i casi sono stati codificati e discussi in sede interpretativa.

positivi (counter-stereotype[4]) e glossari di stile. Questi elementi spostano il modello fuori dalla mera plausibilità statistica e ri-progettano la scena narrativa (Hjarvard, 2013; Hepp, 2020);
– *sviluppo di competenze* (AI/data literacy) per formare gli utenti alle quattro dimensioni di Elliot (2019) che vanno rese operazionali per saper diagnosticare *false neutralità* (Noble, 2018; Benjamin, 2019), tracciare le scelte (provenienza dei dati, logging dei *prompt*), riprogettare ruoli e voci (co-autorialità responsabile);
– *ricerca e miglioramento continuo* perché servono studi interventivi che confrontino tipi di *prompt* (architetturale vs correttivo), *scaffold* diversi (con rubriche, esempi, glossari) e forme di revisione (*peer-review*, rilettura con *checklist*) sugli esiti strutturali del testo (*agency*, ruoli), non solo sul lessico. L'obiettivo è spostare l'asse dall'"aggiustamento locale" al design equo dell'immaginario.

Inoltre, la riduzione osservata tra bozze e versioni finali indica che i cicli di revisione e controllo (*re-prompting*, discussione, rilettura guidata) hanno un effetto formativo misurabile: le prevalenze calano e scompaiono marcatori palesemente problematici. Dai risultati emerge con chiarezza ciò che definiamo *paradosso della fatica*: contro l'aspettativa didattica secondo cui più re-*prompting* dovrebbe ridurre i *bias*, osserviamo che l'aumento delle iterazioni si associa più spesso alla persistenza di *bias* residui. Se non è governato da vincoli progettuali chiari (*checklist* di inclusione, specifiche sui pronomi, indicazioni sulla distribuzione dei ruoli), il *re-prompting* rischia di restare endogeno alla logica generativa, producendo aggiustamenti locali della superficie testuale che lasciano intatte le strutture narrative di fondo. Questo disallineamento è atteso alla luce delle proprietà dei grandi modelli: essi tendono a conformarsi a medie stilistiche e semantiche del corpus, massimizzando la plausibilità statistica (Bender *et alii*, 2021) e riproducendo *bias* già documentati nelle rappresentazioni e negli *embeddings* (Bolukbasi *et alii*, 2016) a meno di interventi progettuali espliciti. In questo quadro, il *prompting* ricorsivo funziona solo se incardinato in una cornice di design che anticipi le scelte critiche (pronomi, ruoli, cornici valoriali) prima della generazione, altrimenti moltiplica i tentativi senza incidere sulla struttura del racconto. In termini di mediatizzazione, ciò significa che l'interfaccia orienta l'azione verso micro-correzioni conversazionali più che verso ripianificazioni macro del testo: si rimodula il lessico, ma raramente si riconfigura l'*agency* dei personaggi (Hepp, 2020). Ne discende che una semplice *didattica del prompt* è inefficace se non viene riconfigurata come didattica dell'architettura e della progettazione vincolata (vincoli, esempi contro-stereotipici, glossari e rubriche a monte del processo) o costruzione di specifici LLMs personalizzabili e controllati (es. i *Gpts* di OpenAI o i *Gems* di Google).

Resta cruciale la questione di sistematica sottorappresentazione delle identità trans* e non binarie. I sistemi generativi, infatti, tendono a replicare pattern problematici ben attestati in NLP, confermando la resilienza di tali *bias*: difficoltà nell'uso non riduzionista dei pronomi, *default* ciseteronormativi, slittamenti verso ruoli femminili di cura e maschili di leadership (Prates, Avelar, Lamb, 2020; Dev et alii, 2021). La nostra evidenza (≈1,2% di menzioni esplicite trans*/NB nelle risposte aperte) segnala infatti che, anche in contesti didattici sensibilizzati sul tema della rappresentatività di generi e sessualità, la visibilità resta fragile. Qui la chiave di lettura mediatizzante è duplice in quanto da un lato gli *immaginari sociali* circolanti nel corpus di addestramento stabilizzano figure tipiche, dall'altro l'uso *esplorativo* dei sistemi, se non accompagnato da competenze di AI *literacy*, tende a ripercorrere *pattern mainstream*, con scarsa generatività verso posizionamenti minoritari.

E ancora: quando la categoria culturale aggregata (seconda per frequenza) mostra riduzioni diffuse nei sottotipi (*religioso/simbolico, tradizione/modernità, generazionale, civilizzazionale/eurocentrico*) rafforza comunque l'ipotesi che il lavoro didattico abbia agito come mediazione consapevole delle cornici valoriali: la tematizzazione dei frame (*chi parla per chi? con quali presupposti?*) ha favorito la ricontestualizzazione di generalizzazioni ampie, sostituendole – almeno in parte – con descrizioni più situate. In termini pedagogici,

4 Per *counter-stereotype* si intende una rappresentazione intenzionale (personaggio, scenario, esempio, immagine mentale) che contraddice un'associazione stereotipica dominante (es. *donna=cura*, *uomo=leadership*; *famiglia=etero*), allo scopo di ridefinire aspettative e allargare lo spazio dei possibili. In psicologia sociale si parla spesso di esemplari contro-stereotipici (*counter-stereotypic exemplars*): figure concrete che rompono l'associazione abituale e possono ridurre, almeno temporaneamente, l'accessibilità degli stereotipi automatici e orientare nuove inferenze.

è proprio questa relazione riflessiva fra tecnologia e pratica a qualificare l'AI *literacy* come competenza culturale e professionale (Panciroli, Rivoltella, 2023; UNESCO, 2024).

Sul versante metodologico curricolare, i dati convergono su un punto: le competenze non si esauriscono nella destrezza di *prompting*, ma servono strumenti e routine che favoriscano spostamenti strutturali nel testo (se non nei *dataset* di addestramento).

Resta, infine, un limite strutturale: la dinamica osservazionale non consente inferenze causali; l'associazione fra iterazioni e *bias* residui può riflettere sia fatiche di regolazione (itero perché faccio fatica a ristrutturare) sia casi più complessi che "resistono". Tuttavia, proprio questa ambivalenza suggerisce (ribadiamo) l'urgenza di spostare l'attenzione dalle sole correzioni locali a strategie di progettazione che anticipino gli esiti (es.: tassonomie di ruoli non stereotipati, glossari di stile, esempi positivi). Una didattica mediata dall'IA che voglia essere equa si gioca su questa capacità di modellare il contesto d'uso, non solo di "riparare" gli output.

In prospettiva curricolare, questo implica istituzionalizzare pratiche di AI *literacy* che orientino la progettazione verso immaginari equi, dove la distribuzione dell'*agency*, la scelta dei pronomi, la definizione dei ruoli e le cornici valoriali siano oggetto di decisioni esplicite e rendicontabili. La GenAI non è soltanto un oggetto di studio, ma un ambiente di mediatizzazione in cui scuola e università possono co-progettare con gli student* testi e scenari che rendono tracciabili le scelte e visibili i valori che le sostengono. È in questa infrastruttura di progettazione vincolata, trasparenza e responsabilità che l'immaginario condiviso si apre a pluralità, inclusione e giustizia rappresentativa, non come esito casuale della generazione, ma come compito formativo intenzionale.

Riferimenti bibliografici

Appadurai A. (1996). *Modernity at large: Cultural dimensions of globalization*. Minneapolis: University of Minnesota Press.

Bender E. M. *et alii* (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*: 610–623.

Benjamin R. (2019). *Race after technology: Abolitionist tools for the new Jim Code*. Cambridge: Polity Press.

Bolukbasi T. *et alii* (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In https://arxiv.org/abs/1607.06520 (ultima consultazione: 10/09/2025).

Buckingham D. (2003). *Media education: Literacy, learning and contemporary culture*. Cambridge: Polity Press.

Buckingham D. (2019). *The media education manifesto*. Cambridge, UK: Polity Press.

Dev S. *et alii* (2021). Harms of gender exclusivity and challenges in non-binary representation in language technologies. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*: 1968–1994. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.150

Elliott A. (2021). *La cultura dell'intelligenza artificiale*. Milano: Codice Edizioni.

EU (2024). *Artificial Intelligence Act (AI Act): Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence*. In https:// eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:32024R1689 (ultima consultazione: 27/11/2025).

Hepp A. (2020). *Deep mediatization*. London: Routledge.

Hjarvard S. (2013). *The mediatization of culture and society*. London–New York: Routledge.

Hobbs R. (2021). *Media literacy in action: Questioning the media*. Lanham: Rowman & Littlefield.

Hossain T., Dev S., Singh S. (2023). MISGENDERED: Limits of large language models in understanding pronouns. In https://aclanthology.org/2023.acl-long.293.pdf

Lauscher A. *et alii* (2023). What about "em"? How commercial machine translation fails to handle (neo-)pronouns, *Association for Computational Linguistics*: 377–392.

Mascitti M., Messina S., Panciroli C. (2025). Post-critical media literacies in teachers: identify ableism in generative ai and educational practices. *Giornale italiano di educazione alla salute, sport e didattica inclusiva, 9(1)*: 1-25.

Messina S. (2025). *Queer Media Education. Alfabeti mediali e di genere per un'educazione equa e plurale*. Milano: Ledizioni

Noble S. U. (2018). *Algorithms of oppression: How search engines reinforce racism.* New York: New York University Press.

Panciroli C. (2025). AI literacy: alfabeti, pensiero critico, spiegabilità. In P. C. Rivoltella, C. Panciroli (eds.). *Didattica delle New Literacies* (pp. 54-65). Milano: Mondadori.

Panciroli C., Rivoltella P. C. (2023). *Pedagogia algoritmica: Per una riflessione educativa sull'intelligenza artificiale*. Brescia: Scholé.
Prates M.O.R., Avelar P.H.C., Lamb L.C. (2020). Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computing and Applications*, 32: 6363–6381.
QueerInAI (2025). *Proceedings of the Queer in AI Workshop 2025*. Kerrville, Texas: Association for Computational Linguistics.
Rivoltella P.C. (2024). Intervento e competenze per le professioni educative nella condizione postmediale. In P. Ingrosso, L. Ferrari (eds.). *COO.DE. Formazione digitale cooperativa* (pp. 33-39). Trento: Erickson.
Taylor C. (2004). *Modern social imaginaries*. Durham: Duke University Press.
UNESCO (2024). Generative AI: UNESCO study reveals alarming evidence of regressive gender stereotypes. In https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes.
Vásquez C., Aslan E., Seghiri M. (2022). Queer discourse in digital environments: Challenges for NLP systems. *Journal of Language and Sexuality*, 11(2): 145–172.