

# Carpe bias, quam minimum credula queries

Sabato Danzilli

Università di Catania/Universität Heidelberg, Italia/Germania, sabato.danzilli@phd.unict.it

## ABSTRACT (ITALIANO)

L'intervento analizza i rischi culturali derivanti dall'uso acritico dei Large Language Models (LLMs) nelle scienze umane, focalizzandosi sul nesso tra formulazione delle query e amplificazione di bias. Attraverso il concetto di "costituzione della domanda euristica" di Johann Gustav Droysen – atto metodologico che trasforma la curiosità storica in indagine critica attraverso la selezione intenzionale di fonti e categorie interpretative – si dimostra come le query negli LLMs non siano mai neutre, ma incorporino premesse che orientano l'output verso narrazioni dominanti. Esempi concreti (dall'analisi di testi oraziani alla ricerca su corpora parlamentari) rivelano meccanismi di conferma bias che minacciano la diversità interpretativa. Si argomenta la necessità di una formazione umanistica "aumentata", capace di coniugare rigore ermeneutico e consapevolezza algoritmica, trasformando l'IA da amplificatore di stereotipi a strumento di esplorazione critica.

**Parole chiave:** bias; domanda euristica; Verstehen; epistemologia digitale; diversità interpretativa.

## ABSTRACT (ENGLISH)

Carpe bias, quam minimum credula queries

The contribution examines the cultural risks arising from the uncritical use of Large Language Models (LLMs) in the humanities, focusing on the link between query formulation and bias amplification. Through Johann Gustav Droysen's concept of the "heuristic question constitution" – a methodological act that transforms historical curiosity into critical inquiry through the intentional selection of sources and interpretative categories – it demonstrates how LLM queries are never neutral but embed premises that steer outputs toward dominant narratives. Concrete case studies (from Horatian textual analysis to parliamentary corpora research) reveal confirmation bias mechanisms threatening interpretive diversity. The paper advocates for an "augmented" humanistic education, merging hermeneutic rigor with algorithmic awareness to transform AI from a stereotype amplifier into a tool for critical exploration.

**Keywords:** bias; heuristic question; Verstehen; digital epistemology; interpretive diversity.

## 1. INTRODUZIONE

Sebbene l'adozione di strumenti di LLM (Large Language Model) come ChatGPT sia ancora in una fase iniziale e non codificata nella prassi scientifica, il loro utilizzo informale sta rapidamente crescendo tra membri della comunità accademica, della popolazione studentesca e tra professionisti/e del settore pubblico e privato. Questi modelli sono impiegati per attività quali il *data mining* o la scrematura preliminare di vasti *dataset*, o anche per superare *gaps* linguistici, facilitando processi altrimenti complessi e dispendiosi in termini di tempo. L'intervento vuole offrire uno spunto di discussione sui mutamenti nell'approccio alla ricerca introdotti dall'interazione tra chi fa ricerca (in senso largo) nell'ambito culturale e i modelli di intelligenza artificiale (IA). Si vuole mostrare come un uso ingenuo di questi modelli rischi di amplificare narrazioni dominanti e stereotipate, emarginando prospettive minoritarie e ipotesi interpretative alternative. Questo fenomeno non solo compromette l'equità nell'accesso alla conoscenza, ma minaccia anche la diversità interpretativa, pilastro essenziale delle scienze umane.

## 2. METODOLOGIA

L'analisi presentata si basa su test condotti con ChatGPT (versione GPT-4o), utilizzando query in italiano. Il modello è stato interrogato tramite interfacce standard, senza personalizzazione dei parametri di generazione (es. temperatura, top-p), per simulare l'uso comune da parte di utenti non esperti. I dati di addestramento dei sistemi LLM utilizzati in misura maggiore nei paesi occidentali, come noto, derivano da corpora eterogenei e multilingue, con una sovrarappresentazione di fonti anglofone e occidentali che influisce sulle risposte e da cui deriva una maggiore "qualità" osservata delle risposte nella lingua di maggior addestramento, ossia l'inglese. La scelta di concentrarsi su casi di studio letterari (es. Orazio) e storici (es. verbali parlamentari) mira a evidenziare dinamiche trasversali alle diverse discipline umanistiche.

### 3. PER UN RECUPERO DEL CONCETTO DI COMPrensIONE

L'IA eccelle nel generare spiegazioni causali basate su correlazioni statistiche, senza tuttavia accedere ai significati soggettivi, che restano centrali nel processo interpretativo umano. Tuttavia, essa può essere uno strumento ausiliario alle capacità umane di analizzare e interpretare dati complessi. In questa luce, il ruolo di ricercatrici e ricercatori non viene sminuito, ma enfatizzato: essi rimangono responsabili nell'attribuzione di significato.

Il concetto classico di *Verstehen* può essere utile per sottolineare la differenza tra un approccio ai dati (in senso ampio) puramente algoritmico e la comprensione umana che integra intenzionalità e contesto. Tuttavia, per evitare che il discorso diventi astratto o banalizzante, tale concetto dovrebbe essere utilizzato solo come punto di confronto metodologico: come monito per l'utente di essere consapevole dei limiti intrinseci dello strumento. Risulta esplicitativo in questo senso richiamare la "costituzione della domanda euristica" come la intendeva Johann Gustav Droysen nella *Historik*. Per Droysen, formulare una domanda non è un atto neutro, ma un processo creativo e critico che delimita il campo d'indagine, seleziona le fonti e definisce i criteri interpretativi. Nell'uso dell'IA, questa consapevolezza metodologica diventa cruciale: una query mal strutturata non solo produce risposte distorte, ma cristallizza una "pre-comprensione acritica", trasformando l'AI da strumento euristico in una gabbia ermeneutica.

Il riferimento a Droysen rimarca in questa sede l'utilizzo non di un approccio mentalistico (es. "le macchine possono pensare?"), ma funzionalistico: si analizza l'atto della query come pratica ermeneutica situata, che richiede un controllo critico sulla coerenza tra premesse e risultati. Formulare una query, anche nell'uso dell'IA, significa già caricare l'indagine di intenzionalità, selezionando e strutturando i dati secondo una visione che non è mai neutra. In questo senso, la scelta delle premesse interpretative, siano esse linguistiche o tecniche, risulta determinante nella costruzione del sapere storico e filologico. L'uso dell'IA solleva dunque domande fondamentali sul rapporto tra spiegazione e comprensione nelle scienze umane. La dipendenza da modelli computazionali sposta l'attenzione dall'intenzionalità soggettiva a un approccio dominato da correlazioni. Tuttavia, se impiegata consapevolmente, l'IA può contribuire a una lettura stratificata dei risultati, integrando le intuizioni interpretative con nuovi livelli di profondità derivanti dall'analisi statistica.

L'intervento intende sottolineare di conseguenza i rischi insiti nell'utilizzo di strumenti di intelligenza artificiale da parte di operatrici e operatori privi di una formazione adeguata. L'assenza di competenze specifiche non solo espone a errori tecnici, ma amplifica le possibilità di incorrere in distorsioni interpretative. Per studiosi e studiose con una conoscenza preliminare limitata, l'IA tende a confermare le loro premesse implicite, rafforzando interpretazioni preesistenti e impedendo una riflessione critica sui dati. Questo effetto è amplificato dalla percezione errata che le risposte fornite dall'IA siano oggettive e neutrali. Le promotrici culturali e altri operatori del settore, che spesso utilizzano strumenti come i LLMs per elaborare rapidamente grandi quantità di dati, rischiano di consolidare prospettive stereotipate o parziali, riproducendo narrazioni dominanti a scapito di visioni alternative o minoritarie. La scarsa consapevolezza epistemologica può portare a decisioni e analisi meno rigorose. Gli accademici che si affidano all'IA per supportare aspetti secondari delle loro ricerche potrebbero inconsapevolmente rafforzare bias preesistenti, compromettendo la pluralità interpretativa e riproducendo schemi interpretativi riduttivi. Questi utilizzi errati trovano terreno fertile in forme di pensiero ancora maggioritarie tra i lavoratori e le lavoratrici della conoscenza. Da una parte, una presunzione di neutralità intrinseca negli strumenti digitali, associata a una scarsa conoscenza tecnica delle loro logiche operative. Si consideri inoltre che l'unicità dell'output fornito dagli LLMs in risposta alle query ammantava psicologicamente l'output stesso di un'oggettività che esso non possiede. Dall'altra, un insieme di pregiudizi radicati in molta ricerca umanistica o filosofica, e non solo negli eredi delle tradizioni "romantiche" che percepiscono la meccanizzazione come minaccia alla creatività umana, o di posizioni che liquidano l'IA come intrinsecamente incapace di cogliere il "vero" significato dei dati. Tali atteggiamenti, seppur di natura diversa, convergono nell'ostacolare una comprensione adeguata delle trasformazioni epistemologiche e metodologiche introdotte dall'IA. Il rischio di *confirmation bias* si rivela particolarmente insidioso, in quanto capace di minare il valore epistemico della produzione scientifica, con un impatto potenzialmente grave sulle scienze ermeneutiche.

L'uso dell'IA solleva dunque domande cruciali sul rapporto tra spiegazione e comprensione nelle scienze umane. La dipendenza da modelli computazionali sposta l'attenzione dall'intenzionalità soggettiva a un approccio dominato da correlazioni. Tuttavia, se impiegata consapevolmente, l'IA può contribuire a una lettura stratificata anche dei testi oggetto di indagine ermeneutica, integrando le intuizioni interpretative con l'analisi statistica. Emergono qui questioni di bias strutturali, ben documentate in letteratura (cfr. studi

di Buolamwini e Gebru sul gender bias, e di Caliskan et al. sui pregiudizi culturali negli algoritmi), che richiedono un'attenta valutazione delle fonti e del contesto di addestramento dei modelli.

L'analisi di grandi moli di dati, caricati sotto forma di allegati – ad esempio tramite testi storici in PDF – può evidenziare due ordini di problemi. Da un lato, errori palesi, come l'attribuzione errata di citazioni; dall'altro, rischi più sottili, quali l'amplificazione di pregiudizi latenti derivanti dalla formulazione stessa della query. L'IA può assecondare implicitamente la preconcoscenza dello studioso, falsando così l'analisi e contribuendo al *confirmation bias*. Non occorre essere tuttavia "apocalittici": come nel caso dell'IA anche la trasmissione manoscritta o le scelte editoriali non sono mai state completamente asettiche; appare evidente: sebbene le biblioteche e i database non "producessero" direttamente risposte, la loro organizzazione e selezione dei testi già rifletteva un orientamento culturale e metodologico. A maggior ragione, tuttavia l'uso dell'IA solleva questioni analoghe di controllo e responsabilità, poiché l'algoritmo stesso codifica convenzioni e preferenze determinate da chi lo progetta o addestra.

Di seguito illustreremo attraverso alcuni esempi i rischi connessi a un uso ingenuo e poco critico delle query nei modelli di intelligenza artificiale. Sottolineiamo che il soggetto ideale di questo "esperimento" non è il tecnico esperto o già formato, ma figure come studiose, lavoratori del settore culturale o membri dell'accademia che necessitano di estrarre rapidamente informazioni da una vasta mole di dati per aspetti secondari o periferici delle loro ricerche.

Immaginiamo uno di questi operatori. Analizzando come egli formuli la sua query, possiamo osservare che l'output generato dall'IA non solo varierà significativamente, ma potrebbe risultare contraddittorio o addirittura fuorviante: un ricercatore che utilizzi un corpus storico digitalizzato di documenti parlamentari, come i verbali delle sedute del Parlamento italiano tra il 1861 e il 1946, per analizzare l'evoluzione del linguaggio politico legato alle riforme sociali. Se la query prevede termini moderni come "welfare" o "diritti civili", è possibile che il modello restituirà risultati poco rappresentativi o falsati, poiché tali termini potrebbero non essere presenti o avere significati completamente diversi nei contesti storici analizzati. Una formulazione superficiale della query potrebbe quindi omettere dibattiti centrali che utilizzavano un linguaggio diverso ma pertinente, una riformulata con linguaggio storico (es. "assistenza pubblica") consentirà esiti più robusti. Questo semplice aggiustamento, però, non basta: la selezione dei documenti e la pesatura statistica delle risposte dipendono da dinamiche opache (es. rappresentatività del corpus digitalizzato), che richiedono un'ulteriore contestualizzazione.

Ma un'insidia nell'uso ingenuo degli strumenti si percepisce a un livello ancora più basilare.

Abbiamo caricato su ChatGPT-4o il 10 gennaio 2025 una versione online facilmente reperibile delle *Opere complete* di Quinto Orazio Flacco.

Immaginiamo che una studentessa, o anche una promotrice culturale, cerchi per questa via di reperire facilmente informazioni riguardo la notissima *Ode I, 11*.

Immaginiamo che interroghi la macchina con questa domanda:

*In che ode Orazio afferma che bisogna cogliere l'attimo?*

Abbiamo ottenuto questa risposta:

*Il famoso invito di Orazio a "cogliere l'attimo" si trova nell'Ode I, 11, rivolta a Leuconoe. In questa ode, Orazio esprime il concetto di carpe diem: "carpe diem, quam minimum credula postero", che si traduce come "cogli il giorno, confidando il meno possibile nel domani". Questo invito rappresenta il fulcro della filosofia esistenziale oraziana, che spinge a vivere il presente senza preoccuparsi eccessivamente del futuro incerto.*

Si tratta di una risposta molto banale e che andrebbe approfondita, ma passabile per fini "enciclopedici".

Immaginiamo tuttavia che lo stesso operatore un po' più dubbioso, abbia invece domandato:

*Orazio afferma che bisogna cogliere l'attimo?*

In maniera interessante, a una domanda più neutrale, la risposta è stata più informativa:

*Sì, Orazio afferma di cogliere l'attimo, ma con una sfumatura e una prospettiva specifica. Nel famoso verso "carpe diem, quam minimum credula postero" tratto dall'ode 1.11, Orazio esorta a "cogliere il giorno, confidando il meno possibile nel domani". Non si tratta di un invito all'edonismo sfrenato, ma piuttosto a vivere consapevolmente il presente, godendo di ciò che offre senza eccessi, in modo da evitare l'angoscia per il futuro incerto. Questo tema riflette la sua adesione a un'etica di moderazione e misura, ispirata anche alla filosofia epicurea e stoica, che cerca di bilanciare il piacere e la ragione.*

Vediamo però cosa succedere se l'operatore pone una query già carica di un'interpretazione più forte:

*Dove Orazio afferma che bisogna godere ogni giorno?*

E la risposta della macchina accompagna l'interpretazione a monte dell'utente:

Orazio celebra l'importanza di godere ogni giorno nel *carpe diem* (afferrare il giorno) in *Odi I, 11*, rivolta a Leucònoe. In questo componimento, il poeta invita a non interrogarsi sul futuro o sulle sorti dettate dagli astri, ma a vivere intensamente il presente:

*"Sappi cogliere allora il giorno come viene, senza illusioni: meno che puoi fidando nel domani."*

Cosa insegnano questi esempi? Le domande formulate agli algoritmi IA non sono mai neutrali. Esse incorporano premesse semantiche e strutturali che influenzano profondamente le risposte. Nel contesto delle scienze umane, dove il significato è spesso polisemico, una query ambigua o implicante relazioni causali non dimostrate può condurre a risposte distorte, rafforzando narrative preconette.

Il cuore della questione risiede nella tendenza dei modelli di intelligenza artificiale a confermare premesse o interpretazioni già presenti nelle query poste dagli utenti. Quando un utente pone una query carica di premesse implicite (ad esempio, chiedendo "Dove Orazio afferma che bisogna godere ogni giorno?"), il modello tende a generare una risposta che asseconda tali premesse, invece di metterle in discussione o evidenziare altre possibili letture. Questo effetto amplifica il rischio che l'utente confermi una narrazione preconetta, compromettendo l'apertura ermeneutica che dovrebbe essere alla base della ricerca nelle *humanities*.

Il rischio epistemico evidente è che l'IA, confermando le premesse implicite dell'utenza, riduca la diversità interpretativa e rafforzi narrative dominanti o stereotipate. Ad esempio, nell'analisi di corpora storici, una query mal formulata potrebbe selezionare soltanto i dati che confermano una specifica ipotesi, lasciando in ombra contronarrazioni potenzialmente significative.

La progettazione delle query non dovrebbe essere vista solo come una tecnica ingegneristica, pertanto, ma come un atto ermeneutico. La precisione semantica nella formulazione non serve solo a ottenere risposte accurate, ma evita che il sistema IA diventi uno specchio dei pregiudizi dell'interlocutore. Una riflessione filosofica sulle premesse interpretative diventa centrale per affrontare i rischi epistemici. Il problema è pressante perché coinvolge il modo stesso in cui sono strutturati questi modelli, nella fase storica in cui essi diventano una tecnologia di massa.

Il motivo per cui una query più neutrale produce una risposta più informativa è che, in assenza di premesse forti o di un'interpretazione implicita nella domanda, il modello IA è "costretto" a generare una risposta che copra un ventaglio di significati più ampio, senza focalizzarsi su una lettura preconfezionata. Quando la query è carica di premesse, come nel caso di "Dove Orazio afferma che bisogna godere ogni giorno?", il modello utilizza quelle premesse come vincoli per generare una risposta. Di conseguenza, tenderà a produrre un'interpretazione che conferma l'input, privilegiando i dati che meglio si adattano a quella cornice semantica, viceversa, una query più neutrale, come "Orazio afferma che bisogna cogliere l'attimo?", lascia invece il modello libero di attingere a una gamma più ampia di contesti, permettendo una risposta più equilibrata che tiene conto di diverse sfumature interpretative.

I modelli linguistici avanzati, come ChatGPT, si basano su correlazioni statistiche apprese dai dati. Quando una query presenta un focus specifico (ad esempio, "godere ogni giorno"), il modello cerca corrispondenze strette, che limitano implicitamente la varietà delle informazioni richiamate. Le premesse cariche di significato nelle query agiscono come "guide" per il modello, ma spesso portano a confermare interpretazioni riduttive o parziali. Nel caso dell'esempio, chiedere esplicitamente "dove Orazio afferma che bisogna godere ogni giorno" trasforma il concetto di *carpe diem* in una lettura edonistica, portando il modello a costruire una risposta che rafforza questa interpretazione. Una query neutrale, invece, consente al modello di esplorare una maggiore varietà di significati, riflettendo, ad esempio, anche l'etica della moderazione epicurea o la prospettiva stoica, come nel secondo esempio fornito.

Le implicazioni per la pratica quotidiana nell'ambito delle *humanities* sono evidenti e determinano la necessità di educare l'utenza di questi strumenti a come mitigare i rischi di vedersi IA che si limitano ad accondiscendere ai loro pregiudizi pensando di fare *data mining*.

Una query apparentemente innocente, come quelle che abbiamo visto su Orazio, ha in realtà un numero molteplice di fallacie. Ne richiamiamo velocemente alcune:

**Bias di conferma:** Domande come "Dove Orazio afferma che bisogna godere ogni giorno?" guidano il modello a confermare un'interpretazione edonistica implicita, ignorando altre prospettive come l'etica della moderazione epicurea o stoica.

**Bias di selezione:** Le premesse nelle query spingono il modello a privilegiare dati che confermano interpretazioni dominanti nei testi di addestramento, escludendo contronarrazioni o letture filologicamente più accurate.

*Bias culturale e linguistico:* L'associazione di carpe diem con il godimento immediato è spesso rafforzata da interpretazioni moderne e da pregiudizi culturali nei dati di addestramento, ignorando la complessità del contesto storico e culturale.

*Bias di disponibilità:* Il modello attinge alle interpretazioni più comuni presenti nei dati di addestramento, trascurando quelle meno frequenti ma più rilevanti per una lettura critica del testo.

*Bias algoritmico:* Gli algoritmi ottimizzano risposte basate su correlazioni statistiche, non considerando il contesto critico della query. Questo limita la capacità del modello di generare risposte che sfidino le premesse implicite.

L'amplificazione dei bias attraverso query cariche di premesse ha anche un impatto diretto sull'eliminazione o sottorappresentazione delle minoranze e delle ipotesi interpretative alternative. L'esperimento con le *Odi* di Orazio illustra il contrasto tra una domanda carica di premesse ("godere ogni giorno") e una aperta ("Orazio afferma di cogliere l'attimo?"). Come nella "costituzione della domanda euristica" di Droysen, la differenza non sta nella "verità" dell'IA, ma nell'interazione tra struttura della query e contesto culturale. Una query del tipo "Quali diritti civili sono stati discussi nel Parlamento italiano nel XIX secolo?" potrebbe portare il modello a enfatizzare le discussioni più comuni (es. riforme elettorali) e ignorare temi meno rappresentati, come il dibattito sull'emancipazione femminile o i diritti delle minoranze religiose. Una query che riflette una visione maggioritaria (es. "Come il Rinascimento ha promosso la libertà intellettuale?") amplifica la narrazione dominante, ignorando eventuali critiche, come quelle legate all'esclusione delle donne, e al filone femminista, o delle minoranze religiose ed etniche durante il periodo.

La scena è comune, quotidiana: uno studente universitario, alle prese con la ricerca di informazioni rapide per un esame, o una professionista che deve esplorare un argomento secondario fuori dal proprio ambito di *expertise*, si rivolge a un modello di intelligenza artificiale, usando magari lo strumento di caricamento di file per "sentirsi più sicuro". La query posta è diretta, apparentemente innocente, formulata per ottenere risposte sintetiche e immediate. Tuttavia, proprio in questo gesto, che può sembrare una prassi per ottimizzare il tempo, talvolta necessaria, si nasconde un rischio insidioso e profondo.

Dalla stesura della prima versione di questo contributo (gennaio 2025), l'ecosistema degli LLM ha subito un'accelerazione: modelli come DeepSeek e strumenti integrati in app di messaggistica, lettori PDF (es. Adobe AI Assistant) o ulteriore diffusione di piattaforme accademiche (es. Scite, Elicit) hanno reso l'IA ubiqua. Non si tratta più solo di sottoporre un prompt a un modello, ma di processi in cui l'IA "sgrossa" o addirittura ricerca dati preliminari, sintetizza letteratura o addirittura suggerisce ipotesi di lettura. Questa automazione parziale crea un salto qualitativo: l'utente medio interagisce e sottopone ad analisi output che sono essi stessi già mediati da algoritmi, spesso inoltre senza accesso al ragionamento sottostante. Modelli di *depththinking* (es. Chain-of-Thought prompting) che esplicitano il processo logico della macchina offrono un contraltare promettente, ma rimangono pienamente comprensibili solo a utenti "avvertiti" dei rischi. Gli esempi discussi (dall'analisi di Orazio alle query sui verbali parlamentari) mostrano che il rischio principale non è la "malafede" dell'IA, ma l'assenza di consapevolezza metodologica nell'utente. Come sottolinea Droysen, la storia (e per estensione le scienze umane) non è mera accumulazione di dati, ma costruzione guidata da domande critiche. Gli strumenti di IA, se usati passivamente, rischiano di ridurre la ricerca a un enciclopedismo frammentario. In Italia, dove l'educazione digitale fatica a illustrare le criticità ermeneutiche, ciò richiede interventi sistemici: non basta insegnare a "verificare le fonti", come suggeriscono gli stessi modelli di IA in fondo alle loro interfacce, ma serve formare a decostruire le premesse implicite delle query, riconoscere i bias cognitivi e valutare l'adeguatezza contestuale delle risposte. Questo approccio richiama in parte l'avvertimento di Droysen contro il positivismo: la ricerca non può ridursi a leggi universali o correlazioni statistiche, ma deve integrare categorie interpretative (politiche, culturali, morali). A tal proposito, strumenti come i framework di "AI literacy" proposti da UNESCO o il progetto europeo AI4T offrono spunti operativi, ma vanno adattati alle specificità delle discipline umanistiche.

L'IA non è un osservatore neutrale; le risposte che produce non sono mere fotografie oggettive della realtà. Ogni query agisce come una lente che amplifica le prospettive e le narrazioni presenti nei suoi dati di addestramento, oltre a indurre un falso senso di sicurezza nello scienziato sociale, che si sente spesso gratificato da informazioni che vanno nella direzione che lui auspicava, ma solo perché il prompt è stato scritto in un modo poco avvertito. Il risultato è un ciclo che rafforza schemi già consolidati, relegando al margine le voci minoritarie, le interpretazioni alternative e le ipotesi innovative. Invece di arricchire il panorama epistemico, rischia di impoverirlo. Infine, resta aperto il problema della possibile "responsabilizzazione" degli algoritmi, ovvero di come la futura implementazione tecnica possa mitigare

bias e distorsioni. Sul punto, possono risultare utili riflessioni che prospettano una maggiore trasparenza e verificabilità nell'intero processo di sviluppo dei modelli, permettendo così di arginare parzialmente i rischi epistemici connessi al loro impiego nelle scienze umane.

Questo, va da sé, non è solo un limite tecnico, ma un rischio etico e scientifico. Se lasciato incontrollato, l'uso ingenuo dell'IA potrebbe minare la diversità interpretativa, uno dei pilastri fondamentali delle scienze umane. Diffondere prospettive stereotipate non è solo un errore di metodo; è un insuccesso del compito della ricerca accademica. La sfida, quindi, non è se utilizzare l'IA, ma come promuovere un utilizzo consapevole, evitando che un atto quotidiano si trasformi in uno strumento di consolidamento di stereotipi e di riduzione della pluralità interpretativa.

#### **4. CONCLUSIONI**

Il problema, in conclusione, non è nella "verità" dell'IA, ma nell'interazione tra struttura della query, bias del modello e contesto culturale dell'utente. Non si tratta pertanto di essere "apocalittici o integrati", ma riconoscere che l'IA necessita un cambiamento cognitivo nella definizione delle pratiche di ricerca. La posta in gioco non è etica in astratto, ma metodologica: come evitare che strumenti progettati per ottimizzare efficienza riducano la complessità interpretativa. La soluzione non sta, o non sta solo, in controlli tecnici ex post (es. filtri anti-bias), ma nella formazione di una professionalità umanistica che coniughi filosofia del linguaggio e filosofia della scienza, teoria delle fonti e critica degli algoritmi. Solo così si potrà scongiurare un'IA quale amplificatore di bias e rendere, al contrario, tale strumento utile per una ricerca più rigorosa e inclusiva.

#### **BIBLIOGRAFIA**

Bender, E. M. et al. (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, DOI:10.1145/3442188.3445922

Buolamwini, Joy (2024) Unmasking AI: My Mission to Protect What Is Human in a World of Machines.

Caliskan, A. et al. (2017) Semantics derived automatically from language corpora contain human-like biases, *Science*, Vol 356, Issue 6334, 183-186, DOI: 10.1126/science.aal4230;

Crawford, K. (2021) *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press;

Droysen, Johann Gustav, *Sommario di Istorica*. Pisa: Edizioni della Normale, 2014; New York: Random House USA;

EU AI4T Project (2024) *AI Pedagogy in Humanities: A Framework for Critical Engagement*. Brussels: European Commission;

Gebru, T. et al. (2021) Datasheets for Datasets, v8 <https://doi.org/10.48550/arXiv.1803.09010>;

Mercier, H., Sperber, D. (2017) *The Enigma of Reason*. Cambridge (Massachusetts): Harvard University Press, 2017;

O'Neil, C. (2017) *Armi di distruzione matematica*. Milano: Bompiani, 2017;

Orazio. (2013) *Tutte le opere*. Roma: Salerno, 2013;

Ricoeur, P. (2000) *La mémoire, l'histoire, l'oubli*. Paris: Seuil;

Ruha, B. (2019) *Race After Technology: Abolitionist Tools for the New Jim Code*. Medford: Polity, 2019

Umoya Noble, S. (2018) *Algorithms of Oppression. How Search Engines Reinforce Racism*. New York: NYU, 2018;

UNESCO (2023) *Guidance for Generative AI in Education and Research*. Paris: UNESCO.