

La conoscenza in pasto all'algoritmo: quando l'AI legge gratis

di [Antonella De Robbio](#)
4 giugno 2025

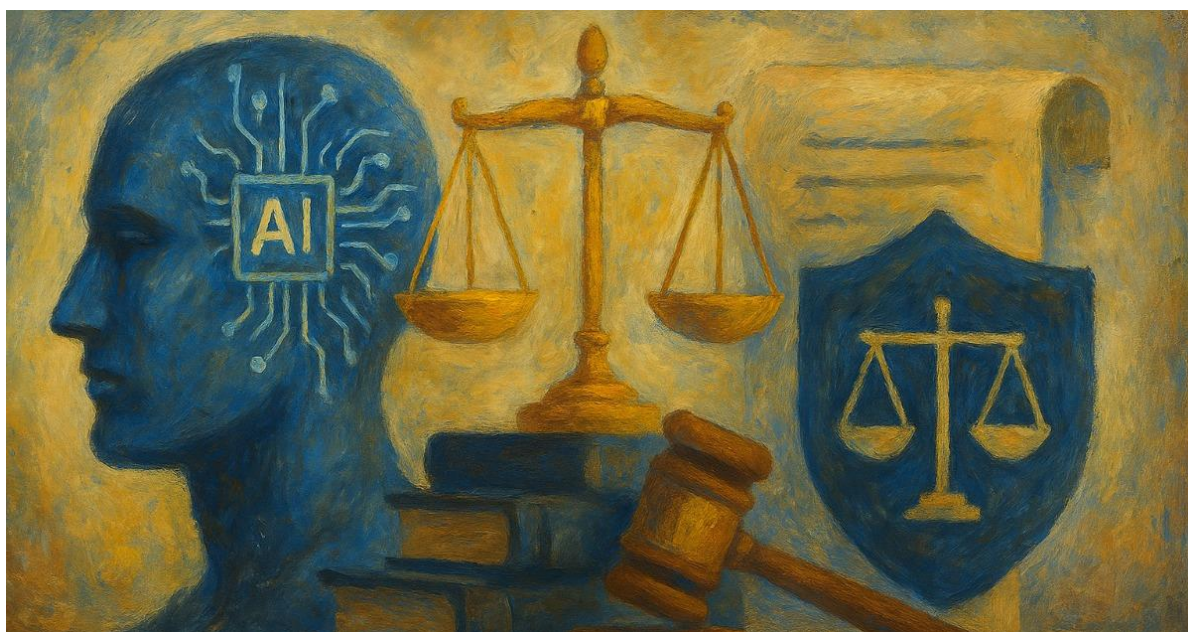


Immagine generata dall'AI

L'avvento dei grandi modelli linguistici di intelligenza artificiale (LLM), come quello alla base di chabot come ChatGPT, ha sollevato un dibattito nella comunità accademica, in particolare tra chi che ormai da 25 anni promuove l'accesso aperto (OA). L'utilizzo di vasti corpus di pubblicazioni scientifiche, compresi i contenuti ad accesso aperto, come dati di addestramento per questi modelli ha infatti suscitato preoccupazioni e richieste di regolamentazione. Sebbene sin dal [suo nascere negli anni 2000](#) – a partire dal convegno *OAI Open Access Initiative*, tenutosi nel 1999 a Santa Fè nel Nuovo Messico – l'obiettivo principale dell'OA fosse rendere i risultati della ricerca liberamente accessibili ai lettori umani, il movimento prevedeva anche la possibilità di accesso e raccolta dei metadati da parte di sistemi automatizzati tramite protocolli come OAI-PMH. Tuttavia, i promotori dell'OA probabilmente non avevano previsto l'assimilazione su vasta scala di interi corpus testuali per l'addestramento di sistemi di IA commerciali. Sebbene l'OA e i dati aperti siano stati originariamente pensati per promuovere l'innovazione, soprattutto a vantaggio di piccole imprese e ricercatori indipendenti, come affermato nelle [Raccomandazioni della Commissione Europea](#) del 2012, oggi ci troviamo di fronte a una realtà ben diversa: le grandi aziende che sviluppano modelli linguistici di grandi dimensioni, sostenute da imponenti investimenti di *venture capital*, rappresentano una nuova categoria di utilizzatori con un impatto potenzialmente destabilizzante sull'intero ecosistema della ricerca. Il loro modo di attingere ai contenuti OA solleva interrogativi seri sull'equilibrio tra apertura, equità e sostenibilità del sistema scientifico.

Alcuni ricercatori nell'ambito dell'IA, soprattutto quelli orientati alla scienza aperta, che vedono nell'addestramento con contenuti scientifici una continuità con pratiche di *text and data mining* già diffuse in ambito accademico, ritengono che l'impiego da parte di modelli di IA avanzati di vasti corpus di conoscenze accademiche, inclusi i contenuti OA, rappresenti un'attività di ricerca legittima e necessaria per far progredire la scienza e la tecnologia a beneficio dell'umanità. Nel 2023 [LAION](#) (Large-scale AI Open Network), un'organizzazione che promuove l'intelligenza artificiale open source e l'accesso ai modelli di IA per la comunità globale, aveva inviato una [lettera aperta al Parlamento Europeo](#) chiedendo di limitare clausole che potessero danneggiare la ricerca e lo sviluppo *open source* nel campo dell'intelligenza artificiale. Tra i firmatari della lettera aperta al Parlamento Europeo vi erano importanti istituzioni di ricerca e accademiche come il Max Planck Institute for Intelligent Systems in Germania, uno dei centri di ricerca più prestigiosi nel campo dell'IA, e l'Université de Montréal, un polo di eccellenza per l'intelligenza artificiale in Canada, oltre a GitHub, una delle piattaforme di sviluppo software più utilizzate al mondo. Restrizioni eccessive sull'utilizzo di questi contenuti per "nutrire" l'IA possono rallentarne lo sviluppo, e di conseguenza frenare anche il progresso in aree cruciali come la medicina e la scoperta di nuovi farmaci, privandoci dei potenziali benefici rivoluzionari dell'intelligenza artificiale in questi campi. I firmatari propongono che la legislazione imponga regolamenti più equilibrati, adeguati al rischio effettivo, in modo da non ostacolare l'innovazione e evitare che ai modelli OS vengano applicati requisiti sproporzionati. Chiedono infine la creazione di infrastrutture pubbliche di calcolo, per consentire alla comunità accademica di esplorare e sviluppare modelli *open source* in un ambiente controllato e sicuro. **In sostanza, la lettera auspica una regolamentazione che sostenga un ecosistema di ricerca aperto, competitivo e trasparente, senza compromettere il progresso nell'IA open source.**

Lavori di ricerca accademica, spesso frutto di anni di sforzi intellettuali, messi in accesso aperto per condividere liberamente le conoscenze, vengono oggi impiegati per addestrare questi sistemi di IA commerciali. Molti ricercatori temono che il loro prezioso lavoro possa essere assimilato senza un adeguato riconoscimento del loro contributo alla conoscenza e senza ricadute open alla comunità. L'assimilazione di questi contenuti da parte dell'intelligenza artificiale potrebbe inoltre minare le consolidate pratiche di citazione e attribuzione, elementi chiave per garantire l'integrità della ricerca accademica.

L'addestramento dei sistemi di intelligenza artificiale non si limita a semplici operazioni di indicizzazione o recupero dei contenuti ma va ben oltre, consentendo a questi modelli di apprendere e interiorizzare le caratteristiche dei testi, al punto da poter generare nuovi contenuti con stili e strutture simili a quelli di partenza. Le controversie legali in corso, che coinvolgono aziende come OpenAI ed editori o autori, potrebbero ridefinire i confini del concetto di uso trasformativo, soprattutto laddove la natura commerciale di molti sistemi di IA indebolisca le tutele offerte dal *fair use*. Se questo ha un impatto evidente sui contenuti accademici a pagamento, la situazione si fa molto più complessa quando si prende in considerazione l'editoria ad accesso aperto con le sue molteplici declinazioni, ormai sempre più centrale nel panorama accademico. È importante chiarire che il presente dibattito riguarda specificamente l'utilizzo di pubblicazioni accademiche e scientifiche, prodotte da ricercatori le cui attività di ricerca sono finanziate e retribuite dalle rispettive istituzioni accademiche o enti di ricerca. Non si sta facendo riferimento ad altri tipi di contenuti come opere letterarie, saggistica, contenuti artistici o scritti di autori che traggono la loro sussistenza economica dai diritti sulle loro opere.

Le licenze attuali, siano esse basate sul copyright tradizionale o su formule open come le CC-BY, raramente distinguono in modo esplicito tra riutilizzo umano e automatico, o tra uso diretto da parte delle persone e uso mediato da sistemi "intelligenti". Eppure questa distinzione sta diventando cruciale, sostiene [Stephanie Decker](#), docente di Strategia alla Birmingham Business School (Regno Unito). Decker, che in questo periodo è [visiting professor a Ca'Foscari](#), e proprio nelle scorse settimane ha tenuto dei seminari al [Dipartimento di Scienze Storiche, Geografiche e dell'Antichità dell'Università di Padova](#), si occupa di studi organizzativi con un approccio storico utilizzando archivi digitali e metodologie storiche e, all'interno della [British Academy of Management](#), è responsabile per l'Open Access e partecipa alle attività sulla valutazione dell'impatto dell'intelligenza artificiale nella ricerca e nella didattica. Durante la recente [consultazione sull'accesso aperto condotta nell'ambito del Research Excellence Framework del Regno Unito](#) la British Academy of Management, portavoce di molte società accademiche nel campo delle scienze sociali e umanistiche, ha espresso preoccupazione per il riutilizzo da parte dell'intelligenza artificiale dei contenuti OA, evidenziando non solo i potenziali effetti negativi, ma anche l'assenza sul tema di un confronto pubblico approfondito. Sebbene le licenze OA permettano legalmente l'uso nei processi di *machine learning*, è improbabile che i ricercatori aderiscano all'OA con l'intento di alimentare gratuitamente lo sviluppo di sistemi IA commerciali: la loro priorità resta infatti l'accessibilità e il riutilizzo da parte della comunità scientifica e del pubblico umano. Quando i modelli di intelligenza artificiale vengono addestrati su testi accademici le parole vengono scomposte in *token*, elaborate da reti neurali e trasformate in rappresentazioni statistiche che non rientrano nelle categorie tradizionali di copia, distribuzione o adattamento. Questo processo, fondato su un'aggregazione massiva, sfugge ai presupposti su cui si basano le attuali licenze. **Il fatto che da contenuti scientifici resi liberamente disponibili si possa estrarre un rilevante valore economico per fini privati sta emergendo con forza.** È significativo e in parte sconcertante, scrive [Decker in suo recente articolo su Kitchen](#), che consultazioni recenti sull'open access, come quella promossa nel 2024 dagli enti di finanziamento della ricerca del Regno Unito, ignorino del tutto la questione dell'addestramento dei modelli di IA su pubblicazioni accademiche e che non venga sollevato il tema se una più ampia adozione di licenze con restrizioni non commerciali (come la CC-BY-NC) possa o meno offrire maggiori tutele.

L'intelligenza artificiale non dovrebbe essere vista come uno sfruttamento commerciale del lavoro intellettuale, ma piuttosto come un'opportunità di ricerca trasformativa. Quando si afferma che l'intelligenza artificiale dovrebbe essere vista come [un'opportunità di "ricerca trasformativa"](#), ci si riferisce al potenziale dell'IA di abilitare nuovi modi rivoluzionari di condurre la ricerca scientifica e accademica, che potrebbero portare a scoperte e progressi dirompenti in vari campi. La ricerca trasformativa implica l'utilizzo di approcci, metodi o tecnologie all'avanguardia che possono cambiare radicalmente le prospettive, le domande o persino i paradigmi di un determinato campo di ricerca, in quanto mira a produrre conoscenze o soluzioni che vanno ben oltre i miglioramenti incrementali, aprendo nuove frontiere o strade inesplorate. **Di conseguenza addestrare modelli di IA sui contenuti accademici, OA e non, potrebbe portare a nuove scoperte, analisi e applicazioni che promuovono il progresso della conoscenza umana in modi che attualmente non possiamo nemmeno immaginare.**

Sul versante copyright editoriale le case editrici "roccaforti", che hanno acquisito i diritti di pubblicazione dagli autori accademici in cambio della possibilità di pubblicare nelle loro prestigiose riviste, stanno reclamando diritti sui contenuti finanziati con fondi pubblici che, si potrebbe argomentare, non dovrebbero nemmeno detenere. **Sarebbe auspicabile un**

quadro normativo aggiornato che consentisse l'uso di contenuti accademici, OA e non-OA, per l'addestramento dell'IA a scopi di ricerca, ma con requisiti chiari di attribuzione e citazione. I modelli di IA addestrati su pubblicazioni accademiche dovrebbero essere in grado di tracciare e attribuire esplicitamente le idee e le intuizioni ai rispettivi autori e articoli originali. Questo potrebbe richiedere l'integrazione di funzionalità di citazione avanzate nei modelli di IA, in modo che ogni output generato sia collegato alle sue fonti sottostanti. Inoltre, potremmo dover considerare nuovi indicatori di impatto accademico che tengano conto del riutilizzo da parte dell'IA, in aggiunta ai tradizionali conteggi di citazioni.

Un quadro normativo equilibrato tutelerebbe gli interessi di accademici e istituzioni di ricerca, assicurando il giusto riconoscimento al loro lavoro intellettuale. Al contempo, permetterebbe di sfruttare appieno il potenziale trasformativo dell'intelligenza artificiale per far avanzare la conoscenza e affrontare le grandi sfide globali. Pensiamo, ad esempio, allo sviluppo di terapie per malattie ancora incurabili, grazie alla capacità dell'IA di analizzare enormi volumi di dati biomedici e accelerare la scoperta di nuovi farmaci. O alla lotta contro la crisi climatica, con l'uso dell'IA nella modellizzazione del clima e nell'ottimizzazione dei sistemi energetici per una transizione verso fonti rinnovabili e sostenibili. Anche la sicurezza alimentare potrebbe beneficiare dell'IA, attraverso il miglioramento delle pratiche agricole e il monitoraggio delle risorse idriche. Infine, l'intelligenza artificiale potrebbe aprire la strada a nuove frontiere scientifiche e tecnologiche, individuando connessioni e schemi nascosti nei dati che sfuggono all'analisi umana.

Regolamentare l'uso delle pubblicazioni accademiche per l'addestramento dell'intelligenza artificiale non sarà un compito semplice. Richiederà un ampio confronto pubblico che coinvolga tutte le parti interessate: ricercatori, editori, enti finanziatori, sviluppatori di IA, legislatori. Non solo: sarà fondamentale anche il contributo dei cittadini, per individuare un equilibrio che tuteli i valori fondanti della ricerca accademica, senza rinunciare alle straordinarie opportunità offerte dall'IA. Solo un processo partecipativo e inclusivo potrà portare alla definizione di regole condivise, capaci di salvaguardare l'integrità del sapere. **Del resto, il movimento per l'accesso aperto è nato proprio dall'idea che la conoscenza sia un bene comune, motore del progresso scientifico e tecnologico: in questa nuova era digitale dovrà sapersi evolvere, restando fedele ai suoi principi di apertura e adattandosi alle sfide del presente.**