Rivoluzione AI e copyright: regole umane per lettori artificiali

di **Antonella De Robbio**

18 AGOSTO 2025

https://ilbolive.unipd.it/it/news/societa/rivoluzione-copyright-regole-umane-lettori

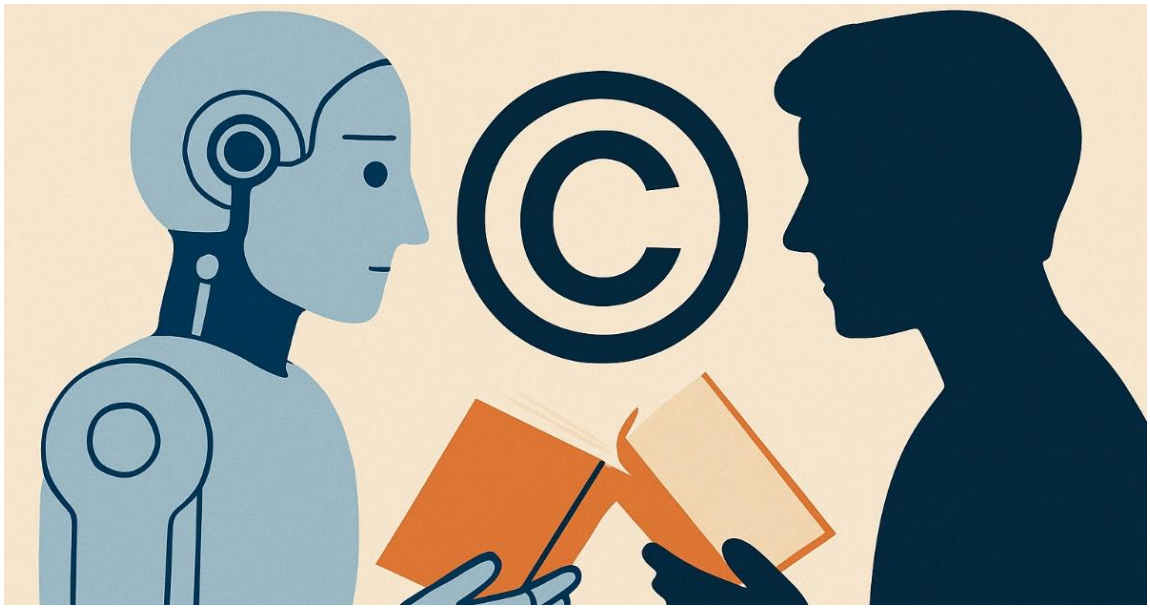

Immagine generata dall'AI

L'avvento e la rapida diffusione dei Large Language Models (LLM) sollevano questioni complesse riguardo l'uso dei dati per il loro addestramento, in particolare l'utilizzo di materiale scientifico – articoli di riviste accademiche, paper di conferenze, capitoli di libri. Un aspetto critico emerge quando gli editori, che spesso acquisiscono i diritti di pubblicazione dagli autori attraverso cessioni "for free", implementano meccanismi di "opt-out" per prevenire l'uso dei contenuti nell'addestramento degli LLM, di cui detengono il copyright. Tale fenomeno solleva interrogativi fondamentali sulle implicazioni legali – anche alla luce delle normative sul Text and Data Mining TDM –, etiche ed economiche, con il potenziale di ridefinire il panorama della pubblicazione scientifica e dell'accesso alla conoscenza.

Gli LLM richiedono immense quantità di dati testuali per apprendere modelli linguistici e generare risposte coerenti e informative. Lo scorso giugno ad esempio la piattaforma di social media Reddit ha intentato causa contro la società di intelligenza artificiale Anthropic, accusandola di aver "raccolto" illegalmente i commenti di milioni di utenti per addestrare il proprio chatbot Claude. Reddit sostiene che, nonostante il divieto esplicito, Anthropic avrebbe utilizzato bot automatizzati per accedere ai contenuti, addestrando intenzionalmente il modello sui dati personali degli utenti senza il loro consenso. Reddit ha già stipulato accordi di licenza con Google, OpenAI e altre aziende che pagano per

poter addestrare i propri sistemi di intelligenza artificiale sui commenti pubblici degli oltre 100 milioni di utenti giornalieri di Reddit.

Quanto alle pubblicazioni scientifiche, ricche di termini specialistici e dotate di struttura logica e informazioni verificate, esse rappresenta una risorsa di inestimabile valore per migliorare l'accuratezza e la pertinenza di questi modelli in ambiti tecnici. L'addestramento degli LLM su vasti corpus scientifici potrebbe inoltre accelerare scoperte e innovazione, in quanto i modelli possono sintetizzare rapidamente informazioni da migliaia di paper, identificare pattern, formulare ipotesi e persino generare nuove idee che un essere umano impiegherebbe anni a elaborare. Da questo punto di vista, l'uso massivo di dati, anche se controverso sul fronte dei diritti, potrebbe rivelarsi fondamentale per sbloccare il potenziale dell'IA. Intere "roccaforti" costruite sul copyright potrebbero essere scosse. D'altro canto editori, agenzie di stampa e singoli autori vedono minacciato il loro modello di business. La legislazione attuale, spesso pensata per un'era pre-AI, fatica a tenere il passo con la velocità con cui l'IA "divora" e rielabora l'informazione e sarà cruciale vedere come i futuri casi legali, sulla scia della Sentenza statunitense *Bartz v. Anthropic PBC,* le nuove normative (AI Act in Europa in correlazione alla Direttiva Europea sul Il Text and Data Mining - TDM) e le negoziazioni tra creatori di contenuti e aziende di AI plasmeranno il panorama dei diritti d'autore.

La Johns Hopkins University Press, che pubblica ogni anno circa 150 nuovi volumi firmati da docenti e studiosi nei campi della sanità pubblica, delle scienze, dell'istruzione superiore e delle discipline umanistiche, con un catalogo di circa 3.000 titoli, ha deciso ad esempio di concedere in licenza i propri libri per l'addestramento dei modelli di intelligenza artificiale. In un'email agli autori, l'editore ha spiegato che i ritorni economici per i singoli titoli saranno modesti – inferiori ai 100 dollari per licenza – ma che il ricavo complessivo potrebbe risultare significativo per sostenere la missione non-profit della casa editrice, soprattutto in un momento di contrazione del mercato accademico. La JHUP non ha reso pubblica la cifra complessiva che si aspetta di ottenere dall'accordo, né ha rivelato il nome dell'azienda di intelligenza artificiale partner; tuttavia l'operazione segue l'esempio di altri grandi editori accademici, che hanno già dimostrato come iniziative del genere possano generare entrate rilevanti. La Oxford University Press ha avviato collaborazioni simili, mentre lo stesso ateneo ha siglato un accordo quinquennale con OpenAI; la Cambridge University Press sta ancora valutando possibili licenze, consentendo però agli autori di rinunciare a eventuali progetti legati all'intelligenza artificiale; la MIT Press, infine, pur senza annunci ufficiali ha confermato di essere stata contattata da diverse aziende e di aver chiesto il parere degli autori prima di procedere.

La sentenza Bartz v. Anthropic PBC: luce verde al Fair Use, stop alla pirateria

La sentenza nel caso *Bartz v. Anthropic PBC*, emessa il 25 giugno 2025 dalla Corte Distrettuale degli Stati Uniti per il Distretto Settentrionale della California (numero di fascicolo 3:23-cv-08577-WAL), rappresenta un momento cruciale di questo dibattito: la decisione, data dal Giudice William Alsup, ha esaminato due aspetti distinti ma correlati dell'addestramento dei Large Language Models (LLM) di Anthropic, la società dietro l'AI Claude.

Il giudice ha stabilito che l'uso di libri **acquistati legalmente** da parte di Anthropic per addestrare i suoi modelli di intelligenza artificiale rientra pienamente nella dottrina del "fair use" (uso lecito) secondo la legge sul copyright statunitense. Questa parte della sentenza è particolarmente significativa perché qualifica tale pratica come "*uso quintessenzialmente trasformativo*" dei contenuti, e sembra destinata a influenzare profondamente le pratiche delle aziende di AI e il modo in cui i titolari dei diritti d'autore tuteleranno in futuro le proprie opere.

In pratica la corte equipara l'atto di un'AI che impara da un testo, acquistato legalmente, al processo di apprendimento e sintesi che avverrebbe in un contesto umano, come uno studente che studia un libro o un insegnante che lo utilizza per formare i propri alunni, senza che ciò costituisca una violazione del diritto d'autore. Tuttavia la sentenza ha anche chiarito un'altra questione fondamentale: il giudice ha specificato che la copia e l'archiviazione di milioni di libri piratati in una "libreria centrale" da parte di Anthropic costituiva una chiara violazione del *copyright* e non poteva essere considerata "fair use". Questa parte del caso non è stata risolta in via definitiva: l'entità dei danni derivanti da questa condotta sarà determinata in un processo separato, previsto per dicembre 2025, il giudice ha fermamente condannato l'uso di contenuti ottenuti illecitamente, aprendo la strada nel prossimo futuro a possibili risarcimenti significativi per gli autori.

La decisione traccia una linea importante, dando il via libera all'utilizzo di materiale protetto da copyright a condizione che questo sia stato acquisito in modo legale. Si auspica che questa sentenza, apripista per quanto riguarda il sistema statunitense, possa offrire un precedente utile anche per l'Europa, contribuendo a chiarire un quadro normativo in evoluzione e a promuovere l'innovazione responsabile nel campo dell'IA.

Il fenomeno addestramento implicito e "Opt-Out" editoriale

In risposta al prelievo massivo di contenuti, molti editori stanno introducendo clausole o tecnologie di **"opt-out"** (come file robots.txt o *watermark* digitali) per impedire o scoraggiare il *crawling* dei loro contenuti. Qui la questione risiede anche nel fatto che gli editori accademici ottengono il materiale scientifico dagli autori – spinti dalla necessità di pubblicare per la progressione di carriera, il riconoscimento accademico e la possibilità di ottenere finanziamenti – senza alcun compenso economico diretto. A ciò si aggiunge che il fondamentale processo di revisione paritaria (*peer-review*), garanzia di qualità e validità scientifica, viene svolto gratuitamente da altri ricercatori e accademici, in uno spirito di servizio alla comunità. Nonostante questa base di acquisizione gratuita sia dei contenuti che dei servizi essenziali di validazione, gli editori generano profitti consistenti attraverso gli abbonamenti – spesso estremamente onerosi per università e istituzioni – e i costi per l'accesso ai singoli articoli. In questo contesto, la loro pretesa di imporre un "opt-out" per l'addestramento dei modelli di intelligenza artificiale su contenuti che non hanno pagato, o di chiedere un ulteriore compenso per tale uso, diventa altamente problematica. Tale posizione degli editori è vista da molti come un ostacolo significativo al potenziale dell'IA di accelerare scoperte scientifiche, lo sviluppo di nuove terapie e, più in generale, il progresso della conoscenza.

Il punto cruciale della sentenza è proprio il concetto di uso trasformativo e la legalità dell'acquisizione. Il giudice William Alsup ha dato ragione alla società guidata da Dario

Amodei, ritenendo che l'uso dei testi di Andrea Bartz, Charles Graeber e Kirk Wallace Johnson – i tre autori promotori della causa – rientri nel cosiddetto "fair use". Se l'AI "impara" dai testi come farebbe una persona, trasformando le informazioni in nuove capacità e risposte senza riprodurre fedelmente l'originale, e se i dati di partenza sono stati ottenuti in modo legittimo (acquistando i diritti o le copie fisiche), allora la legge, secondo questa interpretazione, tende a favorire l'innovazione.

La distinzione che ne consegue, tra la digitalizzazione per la messa a disposizione pubblica (che sarebbe una violazione chiara) e l'uso per l'addestramento (che è un processo interno all'AI) è fondamentale per capire il ragionamento dietro questa illuminante sentenza. Lo stop per Anthropic, a livello di violazione diretta del *copyright*, si giocherà sulla provenienza illegale dei dati, ma questo vale per tutti, sia AI che umani.

Parallelamente al dibattito sul copyright nell'addestramento dell'IA, si sta assistendo a un nuovo e strategico fronte di collaborazione tra aziende tecnologiche e biblioteche, che mira a superare le attuali controversie legali e le limitazioni dei dati. La crescente fame di dati di alta qualità e la minore affidabilità dei contenuti online o dei "dati sintetici" generati dalle stesse IA, sta infatti portando istituzioni come l'Università di Harvard e la Boston Public Library ad aprire i loro vasti archivi. Queste collezioni, che includono quasi un milione di libri dal XV secolo in 254 lingue, oltre a giornali e documenti governativi di dominio pubblico, offrono agli LLM una base di conoscenza storicamente ricca, linguisticamente diversificata (con un focus sulle lingue europee come tedesco, francese, italiano, spagnolo e latino) e legalmente meno problematica, come riconosciuto dalla stessa Authors Guild, precedentemente impegnata in cause contro le aziende AI. Questo ricorda in parte quanto accaduto ai tempi del Google Book Project, che a sua volta aveva affrontato significative sfide legali legate al copyright nella digitalizzazione su vasta scala. Una sinergia che non solo fornisce dati cruciali per migliorare l'accuratezza e le capacità di ragionamento dei modelli AI, ma beneficia anche le biblioteche con finanziamenti (ad esempio da Microsoft e OpenAI) che accelerano la digitalizzazione di patrimoni culturali unici e preziosi, riaffermando anche nell'era digitale il ruolo centrale di questi custodi della conoscenza. Si sottolinea, tuttavia, la necessità di affrontare con responsabilità la presenza di contenuti obsoleti o potenzialmente dannosi, da teorie scientifiche superate a narrazioni storiche problematiche, fornendo linee guida per un utilizzo consapevole dei dati. Un approccio legale e collaborativo che rappresenterebbe un passo significativo verso la democratizzazione della creazione di nuovi modelli AI basati su fonti affidabili.

Come evidenziato nel precedente articolo su addestramento AI e contenuti Open Access, se il movimento dell'Accesso Aperto aveva previsto la libera accessibilità della ricerca e l'uso di metadati per sistemi automatizzati, non aveva anticipato l'assimilazione massiva di interi corpus testuali da parte di sistemi di AI commerciali. L'utilizzo di contenuti scientifici – spesso frutto di lavoro non direttamente remunerato e sottoposto a revisione gratuita – da parte di queste grandi entità solleva urgenti interrogativi sull'equità, la sostenibilità dell'ecosistema della ricerca e il futuro stesso del progresso scientifico. Assicurare che l'innovazione dell'IA proceda nel rispetto dei principi di apertura e del riconoscimento del valore intellettuale è l'imperativo collettivo che ci attende, per garantire che la conoscenza rimanga un bene comune a beneficio di tutti.