T4F Series
PAPER N. 9
a.a. 2023/2024

Trento BioLaw Selected Student Papers

IA e profili di copyright nella raccolta dati: The New York Times vs. Open AI

ROSSELLA BORELLA, VALENTINA CAMPANA, SOFIA LABBRI, FRANCESCA PALMISANO

Jean Monnet EUROPEAN CENTRE UNIVERSITY OF TRENTO
Co-funded by the Erasmus+ Programme of the European Union
UNIVERSITÀ DI TRENTO Facoltà di Giurisprudenza

I paper sono stati selezionati a conclusione del corso *Diritto e Intelligenza Artificiale* a.a. 2023-2024, organizzato all'interno della Cattedra Jean Monnet "T4F – TrAIning 4 Future. Artificial Intelligence and EU Law", coordinato presso l'Università di Trento dal docente Carlo Casonato.

IA e Profili di Copyright nella Raccolta Dati:

The New York Times Company vs. Open AI

Rossella Borella, Valentina Campana, Sofia Labbri, Francesca Palmisano[*]

ABSTRACT: Questo studio esamina il caso NYT v. OpenAI e Microsoft, esplorando le implicazioni legali dell'uso di opere protette da copyright nei dataset di addestramento dell'intelligenza artificiale (IA). Con il rilascio di modelli IA generativi, che necessitano di dataset vastissimi, emerge la questione dell'utilizzo di contenuti protetti da copyright in assenza del consenso dei titolari. Il paper offre una panoramica sui Large Language Models (LLM), sulle criticità legate alla raccolta dati, sulla normativa vigente negli Stati Uniti e su quella dell'Unione europea. L'analisi si conclude con possibili evoluzioni normative e dottrinali per risolvere la tensione tra protezione del diritto d'autore e sviluppo dell'IA.

PAROLE CHIAVE: Intelligenza Artificiale; Diritto d'Autore; Raccolta Dati; Fair Use; Text and Data Mining; AI Act.

AI AND COPYRIGHT PROFILES IN DATA COLLECTION: THE NEW YORK TIMES COMPANY V. OPEN AI

ABSTRACT: This study examines the case of NYT v. OpenAI and Microsoft, exploring the legal implications of using copyrighted works in artificial intelligence (AI) training datasets. With the release of generative AI models, which require vast datasets, the issue arises regarding the use of copyrighted content without the consent of rights holders. The paper provides an overview of Large Language Models (LLMs), the challenges related to data collection, as well as the applicable legal frameworks in the United States and the European Union. The analysis concludes with potential regulatory and doctrinal developments aimed at resolving the tension between copyright protection and AI advancement.

KEYWORDS: Artificial Intelligence; Copyright; Data Collection; Fair Use; Text and Data Mining; AI Act.

SOMMARIO: 1. Introduzione – 2. Caso The New York Times Company vs. Open AI – 3. General Purpose AI: Large Language Model – 4. Problemi nella Raccolta Dati – 4.1. Il Consenso del Titolare di Diritto d'Autore – 4.2. La Trasparenza – 4.3. Business Model Based on Mass Copyright Infringement – 5. Disciplina di riferimento – 5.1. US Copyright e il Fair Use – 5.2. Normativa Europea – 5.2.1. La Direttiva 79/2019: Eccezioni di Text e Data Mining – 5.2.2. AI Act – 5.3. Orientamenti dottrinali e altre proposte – 6. Conclusione.

1. Introduzione

Il presente contributo, a partire dal *caso NYT v. OpenAI and Microsoft*[1], analizza un tema che solo di recente è stato portato all'attenzione della giurisprudenza, ma che si delinea già come nevralgico per le importanti ricadute che avrà sulla disciplina del diritto d'autore e sulla concezione stessa di opera artistica come frutto creativo dell'ingegno umano[2]. Con quella che Torrance e Tomlinson in «*Training is everything: Artificial*

[*] *Studentesse dell'Università degli Studi di Trento, Facoltà di Giurisprudenza. Email: rossella.borella@studenti.unitn.it, valentina.campana-1@studenti@unitn.it, sofia.labbri@studenti.unitn.it, francesca.palmisano@unitn.it.*

[1] *The New York Times Company v. Microsoft Corporation et al.*, Complaint, United States District Court, Southern District of New York, Case 1:23-cv-11195, paragraph 54 (filed Dec. 27, 2023).

[2] Come oggetto del diritto d'autore, definito dall'articolo 1, Legge n. 633 del 22 aprile 1941, Protezione del diritto d'autore e di altri diritti connessi al suo esercizio.

Intelligence, Copyright, and fair training» chiamano «democratizzazione dell'IA»[3] si è assistito infatti, negli ultimissimi anni, ad un uso inedito dei sistemi di IA: con il rilascio da parte di OpenAI di generatori di grafica, di testo, o chatbot estremamente *user-friendly*, si è, infatti, iniziato ad utilizzare sistemi di intelligenza artificiale per svolgere compiti che, prima della loro diffusione, erano dominio della creatività individuale. Perché i sistemi possano assolvere a queste funzioni è però necessario utilizzare una gamma vastissima di brani musicali, immagini e testi scritti nei set di addestramento dell'IA.
Il problema giuridico che si è iniziato a profilare è legato al fatto che molto spesso tali opere sono protette dal diritto d'autore. Le corti e i legislatori, nazionali e sovranazionali, sono quindi chiamati ad individuare, tramite atti normativi e interpretazione giurisprudenziale di clausole generali, un bilanciamento tra i vari diritti, considerando, da un lato, l'importanza del riconoscimento e della tutela di un diritto di proprietà sulle opere ad autori e artisti, dall'altro la rilevanza della creazione di modelli di intelligenza artificiale addestrati su un range di dati sufficientemente ampio e completo da poter fornire output capaci di contribuire al progresso della società nel suo complesso. Nelle pagine che seguono si procederà dunque - a seguito di un breve *excursus* sul funzionamento tecnico dei modelli di intelligenza artificiale - ad un'analisi dei problemi che sorgono nella creazione dei *dataset* per l'addestramento degli algoritmi, in particolare rispetto al consenso sull'utilizzo di opere protette da *copyright* e rispetto al tema della trasparenza, intesa come conoscibilità dei dati utilizzati per l'addestramento. Si sottolinea, inoltre, come la tensione tra regolamentazione del diritto d'autore e sviluppo dei modelli di IA rischi di tradursi nella liberalizzazione di modelli di business basati sulla violazione sistematica della *ratio* alla base del diritto d'autore.
Verrà poi analizzata, in un'ottica comparata, a partire dalle soluzioni che i diversi ordinamenti potrebbero adottare per risolvere il *caso NYT v. OpenAI*, la disciplina di riferimento vigente negli Stati Uniti e la normativa europea - con un particolare focus sulle eccezioni di *text and data mining*[4] e sul nuovo Regolamento europeo sull'intelligenza artificiale (d'ora in poi: AI Act)[5] . Si concluderà, dunque, con l'esposizione di alcune letture dottrinali della tensione tra diritto d'autore e creazione dei dataset per l'addestramento dei modelli di IA, con dei riferimenti alle recentissime evoluzioni proposte da *OpenAI* per la gestione di questo fenomeno.

2. Caso The New York Times Company vs. OpenAI

Il 27 dicembre 2023 *The New York Times Company* fa causa a *Microsoft Corporation* e *OpenAI:* ritiene che siano stati lesi i diritti di copyright attraverso l'utilizzo di articoli e contenuti del New York Times per il training di modelli di IA generativa. L'accusa verte sull'utilizzo di tali contenuti senza alcun accordo tra le parti, nonostante i precedenti tentativi di negoziazione, e, di conseguenza, senza il consenso dei titolari del diritto d'autore. L'avvocato della *New York Times Company*, nell'elaborare la propria posizione, prende in considerazione i dataset utilizzati per i modelli di GPT e sottolinea il gran numero di articoli del *New York Times* che compaiono in essi, al punto che il dominio NYTimes.com nel *webtext dataset* compare nella lista dei primi 15 domini presenti per volume. L'accusa, inoltre, attraverso degli esempi di conversazione con

[3] A.W. TORRANCE, B. TOMLINSON, *Training Is Everything: Artificial Intelligence, Copyright, and Fair Training*, in *arXiv.org*, 2023, disponibile al link https://arxiv.org/pdf/2305.03720 (ultima consultazione 13/10/2024).
[4] Direttiva (UE) 2019/790 del 17 aprile 2019 sul diritto d'autore e sui diritti connessi nel mercato unico digitale e che modifica le direttive 96/9/CE e 2001/29/CE, artt. 3 e 4.
[5] Regolamento (UE) n. 2024/1689 del Parlamento e del Consiglio del 13 giugno 2024 che stabilisce regole armonizzate sull'intelligenza artificiale e modifica alcuni atti legislativi dell'unione (Regolamento sull'intelligenza artificiale).

ChatGPT, mostra come vengano citati nelle risposte gli articoli del *New York Times*, sia riportati parola per parola sia con l'aggiunta di false affermazioni ("allucinazioni"). *Microsoft* e *OpenAI* vengono accusati di "*willful infringement*", violazione volontaria: la *New York Times Company* ritiene che essi fossero a conoscenza o avrebbero dovuto conoscere le possibili conseguenze dell'utilizzo di contenuti coperti da *copyright*. L'avvocato dell'accusa chiede un risarcimento per i danni economici subiti e per la lesione del diritto d'autore, la distruzione dei dati e dei contenuti del *New York Times* dai modelli di training e la possibilità di sottoporre il caso a una giuria.

3. General Purpose AI: Large Language Model

A livello tecnico, la fase che assume maggiore rilievo nella realizzazione dei *Large Language Model* (LLM) è quella di *training*. I modelli rilevanti a livello di *copyright* sono i modelli che generano linguaggio e quelli che si occupano della generazione di immagini. Questi sono dei modelli linguistici di grandi dimensioni usati per la generazione di linguaggio in ambito generale[6]. Per poter operare vengono addestrati tramite enormi quantità di dati, da cui l'aggettivo "*large*", apprendendo miliardi di parametri. In quanto modelli di *machine learning* funzionano grazie all'individuazione di regolarità statistiche: in particolare gli *LLM* prendono in ingresso un testo e "predicono" ripetutamente la parola immediatamente successiva. Per ottenere questi risultati vengono usati i *"word embeddings*", ossia le rappresentazioni delle parole o frasi che vengono "trasformate" in vettori numerici. In sostanza si tratta di una rappresentazione numerica dei token/parole che "cerca di cogliere" la relazione tra le stesse. Per questo, le parole simili tra di loro vengono poste con una distanza minore all'interno dello spazio. Si ritiene che questi modelli assorbano la conoscenza della sintassi e della semantica in modo implicito, ma contemporaneamente acquisiscono imprecisioni o pregiudizi presenti nei dataset di addestramento[7]. Esempi di modelli linguistici "di grandi dimensioni" sono i modelli *GPT di OpenAI*, oggetto della controversia, *PaLM di Google* e *LLaMa di Meta*.
Tali modelli possono assorbire enormi quantità di dati, spesso da Internet, ma anche da fonti come il Common Crawl[8], che comprende oltre 50 miliardi di pagine web, e *Wikipedia*, che ha circa 57 milioni di pagine. Il cuore dei modelli di *LLM* sta quindi nella grande quantità, e non qualità, dei dati su cui l'addestramento si basa. Nel caso in esame, il dataset di addestramento per GPT includerebbe un corpus interno costituito da OpenAI chiamato "*Webtext*", che comprenderebbe i contenuti testuali di 45 milioni di link postati dagli utenti di "*Reddit*", tra i quali rientrerebbero contenuti raccolti dal *Times*[9].

[6] In *Modello linguistico di grandi dimensioni*, in Treccani.it – Vocabolario Treccani online, 2023, disponibile al link https://www.treccani.it/vocabolario/neo-modello-linguistico-di-grandi-dimensioni_(Neologismi)/ (ultima consultazione 29/09/2024).
[7] VISUAL STORYTELLING TEAM AND MADHUMITA MURGIA, *Generative AI*, settembre 2023, disponibile al link https://ig.ft.com/generative-ai/ (ultima consultazione 29/09/2024).
[8] https://commoncrawl.org/ (ultima consultazione 29/09/2024)
[9] https://www.wired.it/article/chatgpt-reddit-addestramento-dati/ (ultima consultazione 13/10/2024) e https://www.agendadigitale.eu/mercati-digitali/new-york-times-contro-chatgpt-ecco-tutti-i-diritti-in-gioco/ (ultima consultazione 13/10/2024).

4. Problemi nella raccolta dati

Come emerge dalla narrazione del caso, diversi sono i problemi legati non solo al prodotto finale (output) realizzato o realizzabile dai sistemi di intelligenza artificiale, ma anche alla raccolta dati (input), che vengono utilizzati per l'addestramento dei modelli. I profili di problematicità legati all'input spaziano, riguardando sia questioni attinenti al diritto d'autore che alla trasparenza nella creazione delle banche dati, sia nella fase di raccolta dei materiali che nella fase di uso delle banche dati stesse. I prossimi paragrafi si focalizzano in particolare sull'analisi dei problemi legati al consenso dei titolari del diritto d'autore, alla trasparenza delle banche dati e ai modelli di business costruiti a partire da queste violazioni.

4.1. Il consenso del titolare di diritto d'autore

Il primo problema riguarda la mancanza di consenso del titolare di diritto d'autore nella creazione e nel conseguente utilizzo di banche dati costituite dalle opere dei titolari. Come già evidenziato infatti nel caso *NYT v. Open AI*, l'attore lamenta l'utilizzo da parte del convenuto di banche dati costituite dai propri articoli, senza il suo consenso. L'assenza del consenso da parte dell'autore rappresenta una violazione del principio cardine del diritto d'autore, il quale conferisce al titolare il diritto esclusivo di sfruttamento e distribuzione dell'opera. Negli Stati Uniti, tale principio è sancito dal Copyright Act del 1976[10]. In particolare, la sezione 17 U.S.C. § 106[11] attribuisce ai titolari del copyright una serie di diritti esclusivi, tra cui la riproduzione dell'opera, la creazione di opere derivate, la distribuzione di copie e la loro esecuzione o esposizione pubblica. L'esercizio di questi diritti richiede il consenso del titolare e ogni utilizzo non autorizzato costituisce una violazione del copyright. Il Digital Millennium Copyright Act (DMCA)[12], inoltre, fornisce ulteriori protezioni per la gestione dei contenuti digitali, rafforzando il quadro giuridico in materia. Pertanto, l'inclusione di opere coperte da copyright in dataset di intelligenza artificiale senza il consenso dei titolari può violare i diritti esclusivi di riproduzione e distribuzione. Secondo quanto emerge dal documento contenente l'atto di citazione, OpenAI sarebbe stato infatti a conoscenza della mancanza di consenso per l'utilizzo dei suddetti articoli a scopo di training, non solo grazie al tentativo di accordo avanzato dall'attore[13],ma anche in considerazione del fatto

[10] *U.S. Copyright Act,* 17 U.S.C. § 101 et seq. (1976).

[11] *U.S. Copyright Act*, 17 U.S.C. § 106 (1976) - Exclusive rights in copyrighted works: *Subject to sections 107 through 122, the owner of copyright under this title has the exclusive rights to do and to authorize any of the following:*

(1) to reproduce the copyrighted work in copies or phonorecords;

(2) to prepare derivative works based upon the copyrighted work;

(3) to distribute copies or phonorecords of the copyrighted work to the public by sale or other transfer of ownership, or by rental, lease, or lending;

(4) in the case of literary, musical, dramatic, and choreographic works, pantomimes, and motion pictures and other audiovisual works, to perform the copyrighted work publicly;

(5) in the case of literary, musical, dramatic, and choreographic works, pantomimes, and pictorial, graphic, or sculptural works, including the individual images of a motion picture or other audiovisual work, to display the copyrighted work publicly; and

(6) in the case of sound recordings, to perform the copyrighted work publicly by means of a digital audio transmission.

[12] Digital Millennium Copyright Act, 17 U.S.C. (1998).

[13] *The New York Times Company v. Microsoft Corporation et al.*, Complaint, United States District Court, Southern District of New York, Case 1:23-cv-11195, par. 54 (filed Dec. 27, 2023).

che il New York Times ha storicamente negoziato accordi di licenza con altre piattaforme tecnologiche per consentire l'utilizzo dei propri contenuti, accordi che in questo caso OpenAI e Microsoft non hanno rispettato, procedendo senza autorizzazione[14].

4.2. La trasparenza

Il secondo problema è quello della trasparenza, intesa in una duplice accezione: il problema, infatti, si divide in conoscibilità dei dati di *training* e conoscibilità dell'algoritmo. Per quanto riguarda la conoscibilità dei dati di *training* bisogna tenere conto che normalmente i dati non vengono raccolti da una sola fonte, ma da molteplici sorgenti e vengono integrati anche con banche dati già costituite. Si avrebbe quindi una conoscibilità completa dei dati se si potessi accedere, e di conseguenza controllare, a ogni fonte e singola banca dati usata. Inoltre, bisogna tenere conto sia della singola informazione che del rapporto in cui si trova con gli altri materiali. Lo scenario poi qui si differenzia qualora si faccia ricorso a banche dati aperte o a banche dati chiuse. Per quanto, dal punto di vista giuridico, la banca dati aperta sia trasparente, nel senso che si possono controllare i materiali usati, in realtà ci si interfaccia con un problema di natura tecnica visto che molto spesso si tratta di un numero di materiali tali che non possono essere effettivamente controllati da un utente qualunque. Nel caso in cui, invece, la banca dati sia chiusa, non è possibile risalire ai contenuti, come accade nel caso in esame. Secondo una ricerca della *Stanford University* e della *Princeton University*[15], *OpenAI* avrebbe una trasparenza pari al 20%, una percentuale piuttosto bassa, ma purtroppo nella media. Nonostante si classifichi nella ricerca come una delle tre compagnie più trasparenti, in realtà soffre molto in quella categoria denominata "*Upstream*" dai ricercatori, nella quale sono considerati *«the ingredients and processes involved in building a foundation model, such as the computational resources, data, and labor used to build foundation models»*[16]. Secondariamente, il problema della trasparenza riguarda l'impossibilità di avere una riconducibilità certa tra un testo generato e un testo magari anche parzialmente contenuto in un dataset: in questo caso si passa dalla trasparenza del dataset alla trasparenza dell'algoritmo, per cui non si potrebbe sapere quale particolare file ha portato alla generazione del dato contenuto. Resta quindi aperto il dibattito in merito a chi sia effettivamente gravato dal dovere di trasparenza tra chi crea la banca dati (o una sua parte) e che invece crea il modello di intelligenza artificiale.

4.3. Modelli di business basati su sistematiche infrazioni del diritto d'autore

Il terzo problema riguarda la violazione del *copyright* nello sviluppo di un *business model*, il danno economico subito dal *New York Times* e il guadagno derivante dalla condotta illecita di *OpenAI* e *Microsoft.*
Gli strumenti della *Generative AI* vengono spesso implementati nel processo di *business*, automatizzando molte attività che prima richiedevano l'intervento di esseri umani, riducendo i costi e incrementando

[14] *Ibidem.*
[15] STANFORD UNIVERSITY'S CENTER FOR RESEARCH ON FOUNDATION MODELS (CRFM) AND INSTITUTE ON HUMAN-CENTERED ARTIFICIALINTELLIGENCE (HAI), *The Foundation Model Transparency Index - A comprehensive assessment of the transparency of foundation model developers,* disponibile al link https://crfm.stanford.edu/fmti/ (ultima consultazione 17/10/2024).
[16] *Ibidem.*

l'efficienza e la produttività. I modelli di *Generative AI* hanno così un impatto significativo in tutti i settori industriali, ma comportano nuovi rischi e sfide per gli imprenditori.
Tali problematiche emergono proprio nel caso analizzato precedentemente. Il problema sorge nel momento in cui il dataset utilizzato per l'addestramento di *GPT* contiene una serie di *link* che rimandano agli articoli del *New York Times,* senza che i titolari dei diritti di copyright abbiano espresso il loro consenso o abbiano ricevuto un equo compenso. Nell'atto di citazione, la *New York Times Company* sottolinea come questa condotta di *OpenAI* e *Microsoft* le abbia arrecato un danno economico portando avanti una concorrenza sleale[17]. *ChatGPT* riportando citazioni di articoli, permette la lettura di essi senza che si debba pagare l'abbonamento al *New York Times*, provocando così una diminuzione degli abbonati al giornale e un danno economico alla compagnia. D'altra parte, invece, *OpenAI* e *Microsoft* ottengono da questa condotta un vantaggio economico: non pagano i diritti al *New York Times*, guadagnano dalle loro piattaforme e dagli abbonamenti sottoscritti dagli utenti e possono lucrare sulla diffusione del modello da loro sviluppato e dalla sua successiva integrazione in altri e diversi sistemi di IA.
A questa accusa della *New York Times Company, OpenAI e Microsoft*, però, rispondono sostenendo che la loro condotta non rientri in tale fattispecie, in quanto affermano che i dataset contenenti i *link* che rimandano agli articoli siano stati utilizzati per scopi di ricerca scientifica e non per scopi commerciali.

5. Disciplina di riferimento

Per analizzare i problemi precedentemente esposti e cercare di risolvere le questioni attinenti ai profili di *copyright* nella raccolta dati, risulta utile affacciarsi alla normativa sul diritto d'autore e contestualmente alla legislazione in materia di nuove tecnologie e intelligenza artificiale. Trattandosi di un caso che ha luogo negli Stati Uniti, è utile prendere in analisi la normativa statunitense in tema di diritto d'autore, rappresentata dal *Copyright Act del 1976*. Allo stesso modo risulta proficuo, e di migliore comprensione e analisi delle problematiche emerse, affacciarsi alla normativa europea in un'ottica di comparazione e traslazione del caso di studio nel contesto europeo.

5.1. US Copyright e il Fair Use

Il caso di studio *NYT v. Open AI* vede il dibattito tra attore e convenuto concentrarsi sulla disciplina di *fair use* nel contesto dell'utilizzo di opere protette da *copyright* per addestrare sistemi di intelligenza artificiale. Il *fair use* è un principio giuridico che negli Stati Uniti consente l'utilizzo limitato di materiale protetto da copyright senza che sia necessario il permesso del titolare dei diritti, in presenza di determinate circostanze. Si tratta di una clausola di eccezione al diritto d'autore, disciplinata dall'art. §107 del *Copyright Act del 1976*[18]. Nel

[17] *The New York Times Company v. Microsoft Corporation et al.*, Complaint, United States District Court, Southern District of New York, Case 1:23-cv-11195, par. 54 (filed Dec. 27, 2023).
[18] Section 107, Copyright Act (17 U.S.C. § 107), Limitations on exclusive rights: «Fair use Notwithstanding the provisions of sections 106 and 106A, the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include— (1) the purpose and character of the use, including whether such use is of a commercial nature or is for

caso citato, infatti, il convenuto sostiene che l'utilizzo dei contenuti del NYT per addestrare sistemi di IA costituisca un "uso trasformativo", che non sostituisce l'uso originale dell'opera, giustificato da un interesse generale. Questo argomento potrebbe essere sostenuto con riferimento al principio del "*fair use*", che consente l'utilizzo di opere protette da copyright per scopi quali critica, cronaca, insegnamento, ricerca, e altro ancora, purché l'utilizzo rispetti determinati criteri[19]. Quando si valuta se un uso costituisca un uso trasformativo ai fini del "*fair use*", la giurisprudenza considera i seguenti fattori[20]: (1) la natura e lo scopo dell'uso: se l'uso è per scopi commerciali o senza scopo di lucro, se è di natura trasformativa (come critica, commento, parodia, ricerca), e se contribuisce in modo significativo al dibattito pubblico o alla creatività; (2) la natura dell'opera protetta: se l'opera originale è di natura più creativa o più informativa; (3) la quantità e la sostanzialità dell'uso: se è stata utilizzata una quantità minima dell'opera originale e se è stata utilizzata solo la quantità necessaria per raggiungere lo scopo trasformativo; (4) l'effetto sull'uso di mercato dell'opera originale: se l'uso trasformativo concorre con il mercato dell'opera originale o se ne diluisce il valore commerciale[21]. La posizione del *NYT* è che l'utilizzo dei suoi contenuti per addestrare sistemi di IA costituisca una violazione del copyright e che tale uso non sia giustificato dall'eccezione di "*fair use*"[22]. Inoltre, sostengono che nonostante l'uso possa essere considerato trasformativo, rimane comunque un utilizzo commerciale dei loro contenuti senza autorizzazione[23]. Questo riflette la distinzione tra fair use trasformativo e uso commerciale, ampiamente discussa in letteratura[24].

Il fatto che *OpenAI* abbia precedentemente pattuito compensi per l'utilizzo di materiale protetto da copyright con altre fonti, come *Alex Springer* e *Associated Press*, potrebbe indicare una volontà da parte di *OpenAI* di rispettare i diritti dei titolari del *copyright* e di negoziare per l'utilizzo dei loro contenuti[25].

Il convenuto solleva la possibilità di ricadere nell'eccezione di *fair use* in quanto l'utilizzo dei materiali protetti da *copyright* sarebbe volto al *training* di un modello a scopo di ricerca e sviluppo tecnologico[26]. Tuttavia, OpenAI immette successivamente il modello sul mercato, rendendo difficile sostenere che l'uso sia senza scopo di lucro[27]. Come discusso in dottrina, l'equilibrio tra uso commerciale e uso trasformativo è essenziale per comprendere i limiti del fair use[28]. Dunque, rivendica un uso che sarebbe senza scopo di lucro e di natura trasformativa, di ricerca, che niente ha a che vedere con lo scopo perseguito invece dal titolare dei diritti, che

---

nonprofit educational purposes; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and (4) the effect of the use upon the potential market for or value of the copyrighted work. The fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the above factors*».* https://www.govinfo.gov/content/pkg/USCODE-2010-title17/pdf/USCODE-2010-title17-chap1-sec107.pdf (ultima (ultima consultazione 29/09/2024).

[19] *The New York Times Company v. Microsoft Corporation et al.*, Complaint, United States District Court, Southern District of New York, Case 1:23-cv-11195, par. 8 (filed Dec. 27, 2023).

[20] See *ivi* nota 13.

[21] *Ivi*, par. 8.

[22] *Ivi*, par. 9.

[23] *Ivi*, par. 8-9.

[24] M.B. NIMMER, D. NIMMER, *Nimmer on Copyright*, LexisNexis, discute ampiamente la relazione tra uso commerciale e trasformativo nell'ambito del *fair use*.

[25] *Ivi*, par. 13.

[26] *Ivi*, par. 8.

[27] *Ivi*, par. 6.

[28] N.W. NETANEL, *Copyright's Fair Use Doctrine and the Transformation of the American Economy*, in *Harvard Law Review*, 106, 1993.

si occupa di giornalismo indipendente. L'eccezione potrebbe essere fatta salva, se non fosse che il modello dopo essere stato addestrato venga poi commercializzato. Non si può considerare un uso a scopo di ricerca dal momento in cui lo stesso modello viene immesso sul mercato e si realizzano servizi è possibile sottoscrivere abbonamenti che generano chiaramente un'entrata per il *provider.*
Inoltre, non si può considerare un uso minimo e necessario per lo scopo trasformativo: il training di modelli IA richiede grandi quantità di dati, e *OpenAI* avrebbe utilizzato integralmente gli articoli del *NYT*, inclusi contenuti riservati agli abbonati[29]. Pertanto, l'utilizzo da parte di *OpenAI* non rientra nemmeno sotto il terzo criterio di *fair use*, che riguarda la quantità e la sostanzialità dell'uso[30]. Per l'addestramento *OpenAI* utilizza non solo una grande quantità di articoli, ma di questi riprende l'intero testo. Non si tratta infatti di una minima quantità di dati necessaria per raggiungere lo scopo trasformativo: questi modelli richiedono un grande numero di dati per essere addestrati e in questo caso è chiaro che il materiale venga reperito senza porsi alcun tipo di limite sostanziale. Infatti, le opere utilizzate non sembrerebbero soltanto quelle sezioni di articoli reperibili online senza bisogno di sottoscrivere un abbonamento per il *New York Times*, dunque parti già condivise dal titolare con il pubblico, bensì articoli integrali, senza aver ottenuto alcun tipo di consenso o sottoscritto alcun abbonamento con il titolare dell'opera.
Infine, il modello GPT, pur essendo un sistema di linguaggio generico, può essere applicato a diversi settori che concorrono direttamente o indirettamente con il mercato del *NYT*, come evidenziato dal querelante[31].

5.2. Normativa europea

Visto l'interesse globale dei profili emersi dal caso in analisi, si ritiene utile procedere alla disamina della normativa europea in materia. Laddove un caso analogo si presentasse in un Paese dell'Unione europea, le parti in causa potrebbero far riferimento sia alla *direttiva 790/2019* sul diritto d'autore e sui diritti connessi nel mercato unico digitale, sia all'AI Act, il Regolamento europeo che disciplina l'intelligenza artificiale approvato il 13 giugno 2024.

5.2.1. La Direttiva 790/2019: Eccezioni di Text e Data Mining

La direttiva 790/2019 sul diritto d'autore modifica le direttive precedenti in materia, 96/9/CE e 2001/29/CE, e le adegua allo sviluppo odierno delle nuove tecnologie e dei nuovi strumenti dell'intelligenza artificiale.
Viene recepita nell'ordinamento italiano con l'emanazione del *D.lgs. n. 177/2021* e apporta modifiche alla *Legge n. 633/1941* sul diritto d'autore. La Direttiva cerca di bilanciare le nuove esigenze di ricerca e sviluppo e le spinte tecnologiche innovative con la tutela del diritto d'autore.
Per l'analisi che si sta effettuando rilevano gli articoli 3 e 4 della Direttiva. L'articolo 3 introduce un'eccezione per le «riproduzioni e le estrazioni effettuate da organismi di ricerca e istituti di tutela del patrimonio culturale ai fini dell'estrazione, per scopi di ricerca scientifica, di testo e di dati da opere o altri materiali cui essi hanno legalmente accesso». Calando tale articolo nel caso pratico, applicare questa eccezione comporterebbe il non riconoscimento della violazione del copyright e la possibile perdita della causa da parte

[29] *Ivi*, par. 86 e 89.
[30] *Ivi*, par. 86.
[31] *Ivi*, par. 5.

del *New York Times,* in quanto lo sviluppo dei modelli di *IA Generativa* tramite l'utilizzo di materiali coperti dal diritto d'autore sarebbe stato funzionale a una ricerca in ambito scientifico.
L'articolo 4, invece, introduce una eccezione o limitazione ai diritti di *copyright* ai fini di estrazione di testo e dati che si applica a tutte le imprese che hanno legalmente accesso a una banca dati, a condizione che non vi sia un'esplicita riserva dei diritti. Il problema principale riguarda la definizione e l'applicazione di questa riserva. Il testo normativo, infatti, fa riferimento ad una riserva espressa dei diritti, che può essere realizzata anche in modo automatico. Dichiarata l'intenzione di avvalersi della riserva, i titolari del diritto d'autore possono, ma non sono obbligati ad, adottare misure per garantirne il rispetto. Nel caso del *New York Times*, il *paywall*, i termini di servizio e la modifica del *robot.txt* servono proprio a indicare l'esplicita riserva dei diritti di riproduzione ed estrazione ai fini di attività di *Text and Data mining (TDM)*. Mancando un'ulteriore specificazione, sarà necessario attendere la giurisprudenza europea su analoghi casi o l'intervento del legislatore per poter comprendere come una questione simile a quella in esame possa essere risolta nel contesto europeo.

5.2.2. L'AI Act

Attualmente l'AI Act non offre una soluzione efficace a questo tipo di problema. Per quanto ci siano degli obblighi in materia di trasparenza rispetto ai dati di *training*, si tratta sempre di disposizioni normative che riguardano i sistemi, mentre sono poche le disposizioni che riguardano i modelli. Nello specifico ad oggi le disposizioni che riguardano i modelli a finalità generale sono gli articoli 53-54 (sezione 2) e gli allegati XI e XII. Tra gli obblighi previsti per i fornitori di modelli di IA per finalità generali vi è quello di fornire una «descrizione dettagliata» delle «informazioni sui dati utilizzati per l'addestramento»[32]. Il problema è che non vengono specificate quali informazioni siano necessarie e soprattutto non si crea un obbligo di trasparenza tale da imporre al fornitore di rendere accessibile il dataset di addestramento per poter controllare eventuali violazioni. Sulla stessa linea si pone il Considerando 107 che indica come necessaria una sintesi dei materiali usati, senza che però questo possa effettivamente, nonostante le speranze evidentemente riposte nel considerando, aiutare i titolari del diritto d'autore. L'accesso diretto infatti resta l'unico modo che si ha per poter verificare se vi siano materiali protetti da diritto d'autore, in quali quantità e da dove questi provengano. Inoltre, il Considerando 105, facendo esplicito riferimento alle normative in materia di diritto d'autore riporta la necessità da parte del titolare del diritto d'autore di sottrarsi espressamente all'utilizzo delle proprie opere per finalità di *training*, ma al contempo indica delle "modalità adeguate" non meglio specificate. Si rinvia quindi a quanto precedentemente detto in materia di "adeguatezza tecnica" delle misure di "sottrazione" e tutela. Mancando infatti un'adeguata definizione i titolari del diritto d'autore ricorrono più a misure tecniche, piuttosto che a misure giuridiche. Questa previsione, inoltre, si pone in contrasto con l'*animus* del diritto d'autore che prevederebbe più che un *opt-out* del titolare una tutela che spetterebbe al fornitore rispettare.
Bisogna tenere inoltre in considerazione che l'articolo 13 si riferisce solo molto latamente ai dati utilizzati nel training e se anche l'allegato IV indica di inserire nella documentazione tecnica (ex Articolo 11) informazioni sull'etichettatura e altre informazioni, questa non è pubblica, bensì riservata ad eventuali controlli delle autorità. Si potrebbe quindi concludere che nonostante vi sia stato un tentativo di tutela questo è comunque

[32] Regolamento (UE) 2024/1689, Allegato XI.

parziale e sembrerebbe derivare da un bilanciamento degli interessi più spostato sugli interessi delle aziende produttrici che non sui titolari del diritto d'autore.

5.3. Orientamenti dottrinali e altre proposte

Come precedentemente delineato, l'affermarsi dell'utilizzo diffuso di modelli e sistemi di intelligenza artificiale, ha portato alla luce nuovi fenomeni, tra cui alcuni anche di rilevanza costituzionale. Molti dei profili di problematicità emersi riguardano l'individuazione di un bilanciamento tra diritti legati ai dati raccolti o utilizzati dall'IA (si pensi al diritto alla privacy rispetto ai dati personali o al diritto d'autore sulle opere frutto del proprio ingegno) e i benefici che l'utilizzo dei sistemi di IA implementati grazie a tali dati hanno sulla società e sul suo progresso (si pensi agli orizzonti aperti dall'utilizzo dell'IA in ambito medico o di ricerca scientifica, artistica, *ecc.*). La necessità di individuare un bilanciamento tra interessi giuridicamente rilevanti risulta evidente anche dall'analisi quivi svolta. Se da un lato, infatti, gli ordinamenti tutelano il diritto d'autore, dall'altro risultano ormai chiari gli importanti benefici che l'IA può portare alla società in moltissimi campi del sapere. Questi progressi verrebbero drasticamente ridotti o, per lo meno, rallentati, laddove si decidesse di addestrare gli algoritmi solamente con opere di pubblico dominio. Come sostenuto in dottrina[33], si potrebbe decidere di svolgere la fase di *training* dei modelli di IA solo su dati liberamente utilizzabili perché non protetti dal diritto d'autore. Tuttavia, se ci si dovesse basare solo su queste opere, si escluderebbero molti dati rilevanti e, soprattutto, si privilegerebbero alcune culture a scapito di altre, alzando di conseguenza il rischio di ottenere output viziati da bias ed errori. Chi sostiene che addestrare l'IA anche su opere protette da *copyright* sia legittimo fa leva sull'argomentazione che il loro uso per l'addestramento è transitorio, privo di scopo di lucro e non interferisce in modo concreto con il diritto d'autore dei titolari. Si richiama, dunque, l'uso trasformativo per sostenere che tale uso rientri nel *fair use* statunitense, o più in generale, nel cd *fair training*[34], ossia nelle eccezioni- tipizzate (come in Unione Europea) o atipiche (come in US)- che nei vari ordinamenti assumono denominazioni diverse, previste per rendere legittima un'azione che altrimenti si configurerebbe come una violazione del diritto d'autore, ossia delle previsioni che permettono l'utilizzo di materiale protetto anche senza il permesso dell'autore in presenza di determinati requisiti. D'altro canto, il rischio che emerge laddove gli ordinamenti riconoscano un ampio margine di libertà nell'utilizzo di dati coperti dal diritto d'autore per l'addestramento dei modelli di intelligenza artificiale, è quello di svuotare di senso il diritto d'autore a favore di quelli che rischiano di essere dei modelli di *business* basati sulla violazione sistematica del *copyright* come originariamente ideato.

Dottrina recente, inoltre, individua un principio costituzionale di conoscibilità e comprensibilità dell'IA, che implicherebbe anche la conoscibilità dei dati utilizzati per la fase di training del modello.[35] Se da un lato l'imposizione di un simile obbligo renderebbe facilmente dimostrabile la violazione del diritto d'autore laddove il training avvenga senza il consenso del titolare del diritto, prova che ad oggi risulta estremamente complesso, se non impossibile, fornire, dall'altro porrebbe un carico in capo agli sviluppatori che rallenterebbe l'addestramento dei modelli. Svariate proposte di innovazioni sono state avanzate da più parti:

[33] A. GUADAMUZ, *A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs*, in *GRUR International*, 2, 2024, pp. 111-127.
[34] A.W. TORRANCE, B. TOMLINSON, *op. cit*.
[35] M. FASAN, *I principi costituzionali nella disciplina dell'intelligenza artificiale. Nuove prospettive interpretative,* in *DPCE online*, 1, 2022, pp. 181-199.

il 7 maggio 2024 OpenAI ha dichiarato di voler implementare uno strumento, chiamato *Media Manager*, tramite cui i titolari di diritti d'autore possono specificare l'oggetto del loro diritto e la loro volontà che le loro opere siano incluse o escluse dal training e dal learning dei modelli di intelligenza artificiale sviluppati dalla società.[36] OpenAI non ha però ancora chiarito il funzionamento preciso di questo strumento, per cui, ad oggi, si possono solo formulare delle ipotesi. Quello che risulta certo è che i meccanismi di *opt-out* già esistenti si collocano in un quadro frammentato e complesso. Ci si auspica quindi che OpenAI riesca a creare uno strumento efficace, che possa semplificare la realtà attuale, per esempio, permettendo ad altre aziende di utilizzare il suo *Media Manager*, cosicché la preferenza espressa dai detentori del diritto d'autore sia segnalata a più sviluppatori contemporaneamente.
In questa stessa direzione si collocano gli esperimenti sul processo di *unlearning,* che, laddove perfezionati, permetteranno ad un modello di rimuovere retroattivamente una specifica parte dei dati su cui è stata addestrato[37]. Ciononostante, appare chiaro che richiedere la rimozione dei propri dati, soprattutto dove già utilizzati, sia una strada complessa da percorrere e che, inoltre, pone in capo a chi lamenta una violazione del proprio diritto d'autore l'onere di attivarsi. Per questo motivo, le aziende, laddove consentano l'opzione di *opt-out*, tendono a predisporre come opzione di default quella di *opt-in.*[38]*,* realizzando una chiara azione di *nudging*.
Un approccio diverso e più tutelante del diritto d'autore- sostenuto da diverse voci, tra cui quella di Reid Southen[39]- sarebbe quello di predisporre come clausola predefinita quella di *opt-out* e permettere ai modelli di IA di utilizzare i dati coperti dal diritto d'autore solo laddove il titolare dello stesso abbia espressamente prestato il suo consenso.

6. Conclusione

In conclusione, questa ricerca ha investigato l'impatto dell'uso di dataset di addestramento soggetti a diritto d'autore nell'ambito dell'intelligenza artificiale, focalizzandosi sul concetto di *fair use* e sulle normative relative alla proprietà intellettuale. Attraverso un'analisi giuridica approfondita e l'esplorazione dei principi giuridici, emerge un quadro complesso che sottolinea l'importanza di trovare un equilibrio tra l'innovazione tecnologica e la protezione dei diritti degli autori. I risultati indicano che, pur riconoscendo i vantaggi derivanti dall'uso di dataset protetti da *copyright* per lo sviluppo e il perfezionamento degli algoritmi di intelligenza artificiale, è essenziale rispettare rigorosamente i limiti imposti dalla legge e garantire il rispetto dei diritti dei titolari di *copyright*. Inoltre, emerge la necessità di una maggiore coerenza e chiarezza nelle normative sulla proprietà intellettuale al fine di fornire orientamenti più precisi e aggiornati per gli sviluppatori e gli utenti di algoritmi di intelligenza artificiale.
Al fine di affrontare queste sfide in modo efficace, si raccomanda un approccio multidisciplinare che coinvolga esperti legali, tecnici ed etici. Solo attraverso un dialogo interdisciplinare e una collaborazione

[36] https://openai.com/index/approach-to-data-and-ai/ (ultima consultazione 13/10/2024).
[37] https://www.wired.it/article/artisti-formazione-chatgpt/?utm_source=pocket-newtab-it-it (ultima consultazione 13/10/2024).
[38] https://www.wired.com/story/how-to-stop-your-data-from-being-used-to-train-ai/ (ultima consultazione 13/10/2024).
[39] https://www.wired.it/article/artisti-formazione-chatgpt/?utm_source=pocket-newtab-it-it (ultima consultazione 13/10/2024).

stretta tra varie parti interessate sarà possibile sviluppare politiche e pratiche che bilancino in modo equo i diritti degli autori con l'innovazione tecnologica e il progresso scientifico.