XV CONVEGNO ANNUALE
DELL'ASSOCIAZIONE ITALIANA DEI PROFESSORI UNIVERSITARI
DI DIRITTO COMMERCIALE "ORIZZONTI DEL DIRITTO COMMERCIALE"

***"IMPRESA E MERCATI: NUMERI E* COMPUTER SCIENCE"**
Roma, 23-24 febbraio 2024

CARLO MEO

ASSEGNISTA DI RICERCA PRESSO L'UNIVERSITÀ DI ROMA "LA SAPIENZA"

Diritto di riproduzione e intelligenza artificiale "generativa". Spunti alla luce dello scontro tra gli editori giornalistici e OpenAI.

SOMMARIO: 1. Introduzione. – 2. L'IA generativa di testo. – 3. Estrazione di testo e di dati. – 4. Riproduzioni temporanee. – 4.1. Temporaneità. – 4.2. Utilizzo legittimo. – 4.3. Parte integrante ed essenziale del procedimento – 4.4. Transitorietà ed accessorietà. – 4.5. Mancanza di rilievo economico autonomo. – 4.6. Considerazioni di sintesi. – 5. I confini del diritto di riproduzione. – 6. Conclusioni.

1. Introduzione.

Nel corso del 2023 numerosi editori giornalistici in tutto il mondo hanno deciso di impedire l'accesso ai propri contenuti ad OpenAI, l'operatore noto per aver sviluppato il servizio di intelligenza artificiale ChatGPT. "*Crawler*" automatici "rastrellano" infatti da tempo i siti giornalistici per raccogliere dati con cui addestrare sistemi di intelligenza artificiale.

Diversi sono gli argomenti adoperati dagli editori per giustificare questa scelta. ChatGPT – si dice – manipola le informazioni e finisce per offrire agli utenti notizie false o imprecise. Il sistema si appropria, poi, dei contenuti prodotti da altri, senza includere citazioni o rinvii alle fonti. La principale preoccupazione dell'industria giornalistica sembra però un'altra. ChatGPT potrebbe diventare un diretto concorrente delle testate. Anziché leggere i giornali cartacei o digitali, gli utenti potrebbero interrogare il "*chatbot*" sulle notizie di interesse, ottenendo una sintesi delle informazioni contenute nei vari siti. Ciò azzererebbe il traffico sulle pagine dei giornali, privando gli editori degli introiti derivanti dalla pubblicità e dalla raccolta dei dati.

Lo strumento giuridico invocato dagli editori per bloccare OpenAI è il diritto d'autore. Gli articoli pubblicati sul sito – si dice – sono coperti da

diritti esclusivi. Gli editori hanno dunque il diritto di impedirne l'uso ai sistemi di intelligenza artificiale[1].

Questa impostazione ha sollevato forti discussioni, anche al di fuori del campo dell'editoria giornalistica. L'esigenza di ottenere il consenso per l'uso di ogni contenuto adoperato nell'addestramento dell'IA infatti rischia di rendere sostanzialmente impossibile lo sviluppo di prodotti come ChatGPT. Sistemi in grado di elaborare immense quantità di dati e di offrire un quadro di sintesi su temi complessi devono, però, essere visti come risultati positivi dell'innovazione tecnologica. L'IA ha, del resto, già dimostrato le sue enormi potenzialità in alcuni campi, come la ricerca scientifica, l'analisi statistica e la programmazione elettronica.

Lo scontro tra editori e OpenAI ha acceso anche un dibattito in dottrina intorno alla questione se il diritto d'autore attribuisca effettivamente ai titolari dei diritti il potere di opporsi all'utilizzo delle opere da parte dell'intelligenza artificiale (d'ora in avanti, IA). Negli USA, una corrente di pensiero diffusa ritiene applicabile all'addestramento dell'IA il principio del "*fair use*"[2]. Viceversa, in Europa sembra più diffusa la convinzione che l'uso dei testi da parte degli algoritmi sia incompatibile con l'attuale disciplina sul diritto d'autore. Gran parte dei contributi affronta quindi il tema principalmente da una prospettiva *de iure condendo*. Questa divergenza nell'atteggiamento della dottrina mette in luce il rischio che si vengano a formare in questo campo soluzioni radicalmente diverse sul piano comparatistico.

---

[1] Sul punto sono numerose le controversie attualmente in corso, specialmente negli USA, anche al di fuori del settore dell'editoria giornalistica. Si v. tra i casi più recenti *Basbanes et al. v. Microsoft Corp., et al.,* U.S. District Court for the Southern District of New York, 1:23-cv-10211 (filed Jan. 5, 2024); *The New York Times Co. v. Microsoft Corp., OpenAI*, U.S. District Court for the Southern District of New York, 1:23-cv-11195 (filed Dec. 27, 2023); *Sancton v. Microsoft Corp., OpenAI*, U.S. District Court for the Southern District of New York, 1:23-cv-10211 (filed Nov. 21, 2023); *Mike Huckabee v. Meta Platforms, Inc., Bloomberg L.P., Bloomberg Finance, L.P., Microsoft Corporation, and The EleutherAI Institute*, U.S. District Court for the Southern District of New York, 1:23-cv-09152 (filed Oct. 17, 2023); *Authors Guild v. Open AI*, U.S. District Court for the Southern District of New York, 1:23-cv-8292 (filed Sept. 19, 2023); *Chabon v. OpenAI Inc.*, U.S. District Court for the Northern District of California, 3:23-cv-04625-PHK (filed Sept. 8, 2023).

Per il momento, non risulta che sia stata ancora raggiunta una decisione definitiva negli USA sulla violazione dei diritti da parte dell'IA generativa. Alcuni profili del problema sono stati affrontati in maniera sommaria dalle corti in alcune ordinanze iniziali. Si v. ad es. *Kadrey v. Meta Platforms, Inc.,* 20 novembre 2023, 23-cv-03417-VC (N.D. Cal. Nov. 20, 2023); *Andersen v. Stability AI, ltd.,* 30 ottobre 2023, 23-cv-00201-WHO (N.D. Cal. Oct. 30, 2023). V. anche l'ordinanza di rigetto della richiesta di giudizio sommario *Thomson Reuters, v. Ross Intelligence, Inc.,* 28 settembre 2023, 20-cv-613-SB (D. Del. Sep. 28, 2023).

[2] V. ad es. LINDBERG, *Building and using generative models under US copyright law,* in *Rutgers Business L. Rev.,* 2023, 1; SAG, *Copyright safety for generative AI,* in *Houston L. Rev.,* 2023, 104 e ss.; LEMLEY, PRINTER, *Fair Learning,* in *Texas L. Rev.,* 2021, 744 e ss.

Il lavoro parte dall'idea che, nel diritto dell'impresa, si debba tenere conto, nei limiti del possibile, dell'internazionalizzazione dei rapporti e, quindi, anche dell'esigenza di evitare che, a livello comparatistico, si vengano a formare sistemi "a doppia velocità". Di qui l'idea di analizzare il problema non soltanto per individuare possibili riforme future del sistema europeo, ma ancor prima per chiedersi fin dove si possa arrivare in base alle regole attualmente in vigore nel nostro ordinamento. Ci si chiede, cioè, se mediante uno sforzo interpretativo non si possa già giungere, *de lege lata,* a soluzioni che consentano, almeno in certa misura, lo sviluppo dell'IA generativa.

Il lavoro affronta dunque la questione se ed in quali circostanze nel sistema europeo le forme di sfruttamento delle opere poste in essere dall'IA generativa ricadano nell'ambito dei diritti esclusivi. Il problema è esaminato con riferimento all'IA generativa di testo. Resta invece ai margini del lavoro l'analisi dei problemi sollevati dall'IA creatrice di immagini, per la quale si pongono questioni, in parte, diverse. Infine, il contributo non si occupa dell'ulteriore questione se i contenuti generati dall'IA siano protetti dal diritto d'autore[3].

Per rispondere alla domanda da cui parte il presente lavoro occorre, in realtà, affrontare una pluralità di problemi. Le principali questioni sembrano le seguenti: a) se la creazione di copie dei testi per l'addestramento dell'IA sia oggetto del diritto di riproduzione; b) se la raccolta di documenti dal *web* sia compatibile con i diritti sulle banche dati; c) se la generazione di testo da parte dell'IA violi i diritti di riproduzione e quelli di elaborazione sulle opere usate nella fase di addestramento. Il *paper* affronta la questione a). Per le questioni b) e c) si deve rinviare ad un successivo sviluppo del lavoro.

2. L'IA generativa di testo.

Nel concetto di IA generativa rientrano tutti i sistemi di IA che, a seguito di un addestramento realizzato su vaste quantità di dati, acquistano la capacità di produrre contenuti, come testi, immagini, sequenze di programmazione, brani musicali, ecc. ChatGPT si fonda su un c.d. "*large language model"* (LLM), cioè un sistema caratterizzato dalla capacità di generare linguaggio e di rispondere così agli *input* degli utenti.

Il processo che porta un'IA come ChatGPT a comunicare comincia con la raccolta di documenti dal *web.* Meccanismi automatici di ricerca (*bot* o *web-crawler*) passano in rassegna la rete, copiano testi presenti sulle pagine aperte al pubblico (articoli giornalistici, voci enciclopediche, raccolte di dati, libri, blog, ecc) e archiviano le copie generate in banche dati (c.d. *dataset*).

[3] Anche questo problema è attualmente oggetto di discussione nel campo del giornalismo digitale. Si v. al riguardo TRAPOVA, MEZEI, *Robojournalism – A copyright study on the use of artificial intelligence in the European news industry,* in *GRUR Int.,* 2022, 589.

I documenti contenuti nel *dataset* vengono poi sottoposti agli algoritmi dell'IA. In questa fase, i testi vengono utilizzati per effettuare esercizi di vario genere, come il riempimento di brani incompleti, la sistemazione di frasi disordinate o la continuazione di frasi tronche. Questi esercizi sono risolti dagli algoritmi per tentativi e, ad ogni errore, il sistema aggiorna i propri dati precedenti. In questo senso, l'IA "apprende" strada facendo la struttura ricorrente delle frasi e i principi fondamentali del linguaggio, cioè le regole grammaticali e sintattiche. L'IA registra poi anche il modo in cui vengono combinate le parole[4].

Con questo procedimento, realizzato su una immensa quantità di testi, l'IA costruisce una mappa statistica pressoché completa delle possibili relazioni tra le parole che compongono una lingua. Essa è così in grado di orientarsi nel linguaggio e di generare testi originali, diversi da quelli utilizzati per l'addestramento[5]. In particolare, quando riceve una domanda dall'utente (c.d. *input* o *prompt*), l'IA individua le parole chiave della domanda e costruisce la risposta con le parole che, sulla base dei modelli appresi, risultano statisticamente più adatte a proseguire il discorso iniziato dall'utente. L'IA parte, cioè, dall'*input* e prevede le parole successive più probabili. Poi ripete parola per parola questo processo, fino ad arrivare alla costruzione di un testo completo.

Fin qui, la generazione di testo sembra consistere sostanzialmente in un meccanismo di individuazione della parola più probabile data la parola precedente. Un sistema del genere, però, non sarebbe di per sé in grado di assicurare che si arrivi ad una frase di senso compiuto e tanto meno che la risposta sia complessivamente coerente con la domanda originaria[6].

Questo problema è superato grazie al fatto che l'IA è anche in grado di tenere conto del "contesto" della conversazione nelle proprie risposte. In

---

[4] Per i sistemi in esame, l'addestramento avviene in genere senza supervisione, cioè su dati privi di "etichettatura" (c.d. *unsupervised learning*). Si v. sul punto GAO, LIU, *Representation learning and NLP,* in *Represenation learning for natural language processing,* edito da Liu, Lin, Sun, Springer, 2023, 8. Un intervento si ha comunque dopo il processo per apportare correttivi generali ed indirizzare la fase generativa, ad es., adeguando alcuni parametri, escludendo certi tipi di *output*, ecc. (c.d. "*fine tuning*"). V. CALLISON-BURCH, *Understanding artificial intelligence and its relationship to copyright,* Testimonianza scritta al US House of Representatives Judiciary Committee, 2023, disponibile al sito: https://docs.house.gov/meetings/JU/JU03/20230517/115951/HHRG-118-JU03-Wstate-Callison-BurchC-20230517.pdf.

[5] Non c'è quindi un processo di vera e propria comprensione del testo da parte dell'IA: LUCCHI, *ChatGPT: a case study on copyright challenges for generative artificial intelligence systems,* in *Eur. J. of risk regulation,* 2023, 5.

[6] Il sistema finirebbe poi anche per rispondere in maniera sempre uguale ad un medesimo *input.* Questo problema è risolto dall'IA, in genere, attraverso l'introduzione di una componente "random" nella generazione del testo. In altri termini, alcune delle parole usate vengono selezionate non sulla base di un calcolo statistico, ma attraverso meccanismi di individuazione casuale. Il che attribuisce varietà ed offre, in fase di *training,* anche la possibilità di ampliare le "conoscenze" del sistema.

realtà, nella fase di addestramento, l'IA non si limita a contare le volte che una certa parola segue un'altra parola. Essa riesce anche ad individuare le espressioni chiave del testo che legge e a captare gli schemi linguistici che ricorrono in presenza di quelle espressioni. In altri termini, l'IA individua diverse tipologie di conversazioni possibili e adatta le statistiche registrate a questi diversi "contesti" comunicativi. Essa sa dunque modificare il proprio modo di parlare a seconda che le venga richiesta un'informazione d'attualità, un racconto fantastico, una conversazione divertente oppure una teoria scientifica. L'IA è così in grado di conversare in maniera coerente sia da un punto di vista contenutistico che stilistico[7].

La coerenza delle risposte è poi raggiunta anche grazie al fatto che, quando genera il testo, l'IA non individua il termine più probabile soltanto sulla base dell'ultima parola utilizzata, ma tiene conto di tutte le parole usate nella conversazione con l'utente (i.e. l'*input* e le parole già generate). Il testo prodotto dall'IA è quindi coerente anche con il "contesto" della specifica conversazione[8].

---

[7] In ciò sta una delle maggiori innovazioni dei sistemi di IA generativa, definiti sistemi "*transformer*", rispetto ai precedenti meccanismi. Si v. per una efficace descrizione sul punto SAG, *Copyright safety for generative AI,* in *Houston Law Review,* 2023: "*one of the key differences between transformers and the prior state of the art, recurrent neural networks ("RNNs"), is that rather than looking at each word sequentially, a transformer first notes the position of the words. The ability to interpret these "positional encodings" makes the system sensitive to word order and context, which is useful because a great deal of meaning depends on sequence and context. Positional encoding is also important because it facilitates parallel processing: this in turn explains why throwing staggering amounts of computing power at LLMs works well for transformers, whereas the returns to scale for RNNs were less impressive. Transformers were also a breakthrough technology because of their capacity for "attention" and "self- attention." In simple terms, in the context of translation, this means that the system pays attention to all the words in source text when deciding how to translate any individual word. Based on the training data, the model learns which words in which contexts it should pay more and less attention to. Through "self- attention" the system derives fundamental relationships from input data and thus learns, for example, that "programmer" and "coder" are usually synonyms, and that "server" is a restaurant waiter in one context and a computer in another*".

[8] Beninteso, tutto ciò non garantisce che le risposte fornite da un'IA siano corrette. L'IA riesce a rispondere alle domande che riguardano temi specifici soltanto grazie al fatto che durante l'addestramento ha incontrato le parole chiave relative al tema in questione e ha associato a queste parole un determinato contesto comunicativo, delle espressioni ricorrenti, ecc. Ovviamente, tutto ciò non garantisce però che la risposta sia corretta. Tanto più che, una volta concluso l'addestramento, l'IA perde accesso generalmente al *dataset* di riferimento. Sicché il sistema non ha neppure la possibilità di sottoporre le informazioni ad un controllo di attendibilità.

3. *Estrazione di testo e di dati.*

Il procedimento fin qui sintetaticamente descritto intercetta il diritto d'autore in varie fasi. Nella maggior parte dei casi, i testi presenti su Internet sono contenuti coperti dal diritto d'autore in qualità di "opere letterarie" (art. 2 l. aut.)[9]. Questo è senz'altro il caso degli articoli giornalistici raccolti dai siti *web* dei quotidiani. Gli articoli sono anche oggetto del diritto connesso degli editori sull'utilizzo online delle pubblicazioni giornalistiche, recentemente introdotto con l'art. 15 dir. 2019/790 (recepito in Italia nell'art. 43-*bis* l. aut.).

Per addestrare un'IA, in genere, i testi vengono scaricati da Internet, vengono tradotti in un formato adatto agli strumenti di lettura automatizzata e riversati in un *dataset*[10]. I *file* contenuti nel *dataset* possono poi eventualmente essere anche riprodotti, in tutto o in parte, per facilitare lo svolgimento degli esercizi di addestramento da parte dell'IA. Tutti questi passaggi comportano la creazione di copie dell'opera, e la creazione di copia è oggetto sia del diritto di riproduzione degli autori (art. 13 l. aut.) che del diritto connesso degli editori di giornale (art. 43-*bis* l. aut.)[11]. A prima vista, la raccolta dei testi è dunque un atto di sfruttamento soggetto all'esclusiva dei titolari dei diritti[12].

---

[9] NORDEMANN, PUKAS, *Copyright exceptions for AI training data – will there be an international level playing field?,* in *J. IP Law and Practice,* 2022, 973, secondo cui la vasta maggioranza dei documenti che attualmente alimentano il processo di addestramento è coperta dal diritto d'autore. Fanno comunque eccezione le opere cadute in pubblico dominio e quelle prive di carattere "creativo", come i testi composti da formule matematiche, i testi normativi, i manuali di istruzioni per l'uso di prodotti, ecc. Si v. sul tema BERTANI, *Diritto d'autore europeo,* Giappichelli, 2011, 105 e ss.

[10] Si v. sul tema lo studio commissionato dalla Comm. Europea, *Study on copyright and new technologies: copyright data management and artificial intelligence,* SMART 2019/0038, 2022, 182 e ss.

[11] Teoricamente l'esame del testo può anche avvenire con forme di "accesso diretto", vale a dire senza la intermediazione di una copia: il sistema individua il testo su Internet e lo sottopone ad analisi direttamente "alla fonte". In questo caso, non c'è una riproduzione, fatta salva forse la creazione di copie effimere nel processo di *training*. Comunque, a quanto risulta, questo metodo di analisi può trovare applicazione per sistemi semplici o che versano in una fase iniziale di sviluppo. Per la creazione di sistemi più complessi si segue, generalmente, un processo diverso, fondato sulla creazione di copie stabili. Non è escluso comunque che la tecnica dell'analisi "diretta" possa presto trovare applicazione anche ai sistemi più complessi. Si v. sul tema le considerazioni svolte nello studio finanziato dalla Comm. Eur., TRIAILLE, DE MEEÛS D'ARGENTEUIL, DE FRANCQUEN, *Study on the legal framework of text and data mining,* 2014, 31 e 47. V. anche MONTAGNANI, AIME, *Il text and data mining e il diritto d'autore,* in *AIDA,* 2017, 382 e STAMATOUDI, *Text and data mining*, in *New Developments in EU and International Copyright Law,* edito da Stamatoudi, Wolters Kluwer, 2016, 1261.

[12] Il problema della creazione di copie per il *training* è al centro della maggior parte delle controversie statunitensi in corso in tema di IA. V. *Kadrey v. Meta Platforms, Inc.,* 20 novembre 2023, 23-cv-03417-VC (N.D. Cal. Nov. 20, 2023); *Andersen v. Stability AI, ltd.,* 30

Nel dibattito pubblico sul tema si obietta che i testi vengono copiati da siti aperti al pubblico o, comunque, sono acquisiti tramite abbonamenti di lettura. In queste circostanze, gli autori non potrebbero pretendere di bloccare la successiva riproduzione dei testi da parte dei *crawler*. In realtà però, come è noto, il caricamento di un'opera su un sito *web* non produce l'esaurimento del diritto di riproduzione. Questo resta quindi esercitabile nei confronti delle forme di riutilizzo digitale dell'opera.

V'è anche chi sostiene che il titolare che accetta di caricare un articolo su un sito è consapevole che il suo testo potrà essere usato per attività di estrazione e accetta questa eventualità. In questo senso, si potrebbe dire che, con il caricamento, egli sta autorizzando implicitamente la copia della propria opera per fini di analisi. In realtà, però, gran parte dei titolari dei diritti è del tutto inconsapevole delle forme di utilizzazione computazionale che avvengono sul *web*. Quelli che ne sono consapevoli spesso si adoperano per bloccare l'estrazione, ad es., ponendo restrizioni al *download* seriale di contenuti dal sito. Peraltro, la mancanza di restrizioni tecniche alla copia del sito può derivare anche da un problema di costi o dalla difficoltà di costruire un efficace sistema di protezione. Non è dunque scontato che dietro il caricamento di un contenuto su un sito aperto al pubblico vi sia un'implicita accettazione dell'autore circa l'estrazione della propria opera. E non sembra quindi convincente leggere la pubblicazione digitale di un testo come una rinuncia dell'autore a far valere i propri diritti sul riutilizzo dell'opera[13].

La creazione di copie a fini di addestramento è, dunque, in linea di principio, attività soggetta ai diritti esclusivi. Resta, però, da chiedersi se queste riproduzioni non possano rientrare in una delle eccezioni al diritto d'autore. Qui viene, innanzitutto, in rilievo l'eccezione sull'estrazione di testo e di dati introdotta dalla dir. 2019/790 (art. 3 e art. 4, recepiti dagli artt. 70-*ter* e ss. l. aut.)[14]. Con l'espressione "estrazione di testo e di dati" si intende "*qualsiasi tecnica di analisi automatizzata volta ad analizzare testi e dati in formato digitale avente lo scopo di generare informazioni, inclusi, a titolo non esaustivo, modelli, tendenze e correlazioni*" (art. 2, dir. 2019/790). Come già visto, l'addestramento dell'IA è un processo di analisi volto a costruire

ottobre 2023, 23-cv-00201-WHO (N.D. Cal. Oct. 30, 2023); *Thomson Reuters, v. Ross Intelligence, Inc.*, 28 settembre 2023, 20-cv-613-SB (D. Del. Sep. 28, 2023).

[13] In questo senso, si v. OTTOLIA, Big data *e innovazione computazionale,* Giappichelli, 2017, 33 e ss. e, con riferimento alle copie "cache", HUGENHOLTZ, *Caching and copyright. The right of temporary copying*, in *EIPR,* 2000, 490 e ss. V. anche CORTE DI GIUSTIZIA, 16 novembre 2016, C-301/15, *Soulier e Doke,* par. 37 e ss., in cui si afferma che un consenso implicito dell'autore può darsi soltanto laddove questi sia stato previamente informato della futura utilizzazione della sua opera da parte dei terzi e degli strumenti di cui dispone per vietarla.

[14] L'eccezione si applica tanto al diritto d'autore quanto al diritto connesso degli editori di recente introduzione. V. art. 3 par. 1 e art. 4, par. 1 della dir. 2019/790.

regressioni statistiche sull'uso del linguaggio. Sembra dunque poter rientrare nella definizione di "estrazione di testo e di dati"[15].

Ai sensi di questa disciplina, gli organismi di ricerca e le istituzioni di tutela del patrimonio culturale possono liberamente riprodurre le opere per fini di estrazione (art. 70-*ter* l. aut.), a condizione che l'operazione sia effettuata a fini di ricerca scientifica e che le riproduzioni riguardino solo documenti cui l'ente abbia legalmente accesso. Sono espressamente inclusi in questa categoria di documenti quelli presenti in siti aperti al pubblico senza restrizioni[16]. Se necessario, le copie prodotte possono essere conservate con adeguati meccanismi di sicurezza. La raccolta di testi dal *web* e la creazione di un *dataset* per lo sviluppo dell'IA generativa sono dunque attività libere se poste in essere, ad es., da un'Università per scopi meramente scientifici[17].

---

[15] In questo senso si esprime la maggior parte della dottrina. Si v. ad es. GEIGER, IAIA, *The forgotten creator: towards a statutory remuneration right for machine learning of generative AI,* in *Computer Law & Security review,* 2023; MARGONI, KRETSCHMER, *A deeper look into the EU text and data mining exceptions: harmonization, data ownership and the future of technology,* in *GRUR Int.,* 2022, 685. Qualche dubbio al riguardo è sollevato da NORDEMANN, PUKAS, *Copyright exceptions for AI training,* cit., 974, secondo cui le correlazioni estratte dall'IA non sono accessibili all'uomo, ma possono essere usate soltanto dalla macchina stessa. Si dubita quindi che il sistema sia effettivamente volto a "generare informazioni". L'osservazione è condivisa anche dallo European Writers' Council: si v. lo *Statement on the trilogue negotiations of the AI Act proposal and on the urgently needed reform of the text and data mining exception Art. 4 of the CSDM directive 2019/790 (EU),* 26 luglio 2023, disponibile al sito: https://europeanwriterscouncil.eu/23ewc_on_aiact/. Va detto però che la definizione di estrazione non contiene alcun riferimento all'utilizzo delle informazioni dopo la raccolta. Essa pare quindi prescindere dalla questione se le informazioni generate siano direttamente comprensibili dagli utenti oppure richiedano, a tal fine, un ulteriore passaggio tecnologico. L'applicabilità all'IA della disciplina sull'estrazione parrebbe poi confermata dal testo provvisorio di regolamento europeo sull'intelligenza artificiale (c.d. AI Act, v. il testo pubblicato da Politico, disponibile al sito: http://www.openfuture.eu/wp-content/uploads/2023/12/231206GPAI_Compromise_proposalv4.pdf). Qui infatti si prevede il dovere dei sistemi di IA di adottare meccanismi adeguati a rispettare i limiti dell'eccezione di estrazione. L'eccezione è considerata quindi applicabile all'IA. Nello stesso senso, si v. KELLER, *A first look at the copyright relevant parts in the final AI Act compromise,* in *Kluwer Copyright Blog,* 11 dicembre 2023, disponibile al sito: https://copyrightblog.kluweriplaw.com/2023/12/11/a-first-look-at-the-copyright-relevant-parts-in-the-final-ai-act-compromise/.

[16] V. considerando 14: "*la nozione di accesso legale dovrebbe essere intesa nel senso che comprende l'accesso ai contenuti sulla base di una politica di accesso aperto o di accordi contrattuali, quali abbonamenti, tra i titolari dei diritti e gli organismi di ricerca o gli istituti di tutela del patrimonio culturale, o mediante altri mezzi legali*".

[17] La disposizione prevede anche che "*i titolari dei diritti sono autorizzati ad applicare misure atte a garantire la sicurezza e l'integrità delle reti e delle banche dati in cui sono ospitate le opere o altri materiali. Tali misure non vanno al di là di quanto necessario per il raggiungimento di detto obiettivo*". Le misure non devono, cioè, compromettere l'applicazione dell'eccezione (v. anche considerando 16). Per una panoramica sui possibili problemi applicativi sollevati

La riproduzione a fini di estrazione è, in linea di principio, consentita anche agli organismi diversi da quelli di ricerca, come le imprese e gli enti pubblici (art. 4 dir. e art. 70-*quater* l. aut.). Anche qui, l'operazione deve riguardare documenti cui l'ente abbia legalmente accesso[18]. Per queste ipotesi di estrazione, però, si prevede che i titolari possano effettuare una "riserva". In tal caso, l'eccezione non si applica (sistema c.d. *opt-out*).

Per i testi caricati su siti aperti al pubblico, la riserva deve essere espressa in maniera tale che i sistemi di lettura automatizzata possano "captarla", ad es., adoperando metadati o dando indicazione nelle condizioni d'uso del sito (art. 4, par. 3 dir. 2019/790 e cons. 18). Si tratta di un onere non molto gravoso, specialmente per le imprese che effettuano investimenti significativi per il funzionamento del proprio sito *web*, come gli editori giornalistici o letterari[19]. V'è dunque la possibilità che si verifichi un ricorso generalizzato all'*opt-out* da parte di questi operatori. Tanto più che gli editori producono grandi quantità di testi ed hanno, quindi, più da guadagnare dalla concessione di eventuali licenze[20]. Per i sistemi di IA a

---

da questa disposizione v. GEIGER, FROSIO, BULAYENKO, *Text and data mining in the proposed copyright reform: making the EU ready for an age of Big Data?,* in *IIC,* 2018, 836 e ss.

Ai sensi del considerando 11, "*in linea con l'attuale politica di ricerca dell'Unione, che incoraggia le università e gli istituti di ricerca a collaborare con il settore privato, gli organismi di ricerca dovrebbero beneficiare di una tale eccezione anche nel caso in cui le loro attività di ricerca siano svolte nel quadro di partenariati pubblico-privato. Gli organismi di ricerca e gli istituti di tutela del patrimonio culturale dovrebbero continuare a essere i beneficiari dell'eccezione, ma dovrebbero anche poter fare affidamento sui loro partner privati per effettuare l'estrazione di testo e di dati, anche utilizzando i loro strumenti tecnologici*". Sembrerebbe dunque ammessa l'estrazione effettuata dai partner commerciali degli enti di ricerca. Naturalmente, a condizione che siano rispettati i limiti previsti dall'art. 3 e, quindi, che l'estrazione sia effettuata esclusivamente per scopi di ricerca scientifica e che sia effettuata su testi cui gli enti di ricerca abbiano legalmente accesso.

[18] Le copie possono essere poi conservate per il tempo necessario ai fini dell'estrazione (art. 4, par. 2, dir. 2019/790).

[19] D'altra parte, l'*opt-out* può anche essere esercitato in maniera "collettiva" attraverso le *collecting societies.* È quanto accaduto con la francese SACEM, la quale ha recentemente esercitato l'*opt-out* per conto di tutti i titolari rappresentati. V. la dichiarazione del 12 ottobre 2023, al sito https://societe.sacem.fr/en/news/our-society/sacem-favour-virtuous-transparent-and-fair-ai-exercises-its-right-opt-out.

[20] MANSANI, *Le eccezioni per estrazioni di testo e di dati, didattica e conservazione del patrimonio culturale,* in *AIDA,* 2019, 13; DUCATO, STROWEL, *Ensuring text and data mining: remaining issues with the EU copyright exceptions and possible ways out,* CRIDES Working Paper series, 1/2021, 13. C'è da dire che, secondo alcuni, il controllo sul rispetto della riserva da parte dell'IA è molto difficile (se non addirittura impossibile) per i titolari. Sicché la tutela dell'*opt-out* rischia di essere sostanzialmente vanificata. Si v. *Study on copyright and new technologies: copyright data management and artificial intelligence,* SMART 2019/0038, 2022, 201. A questo proposito, il testo provvisorio di regolamento europeo sull'IA (AI Act) contiene il dovere per i sistemi di IA di predisporre "*a sufficiently detailed summary about the content used for training*". Secondo alcuni, si tratterebbe di una previsione volta proprio ad agevolare l'esercizio dell'*opt-out.* Si v. in tal senso GEIGER, IAIA, *The forgotten creator,* cit., 4 e ss. e QUINTAIS, *Generative AI, copyright and the AI Act,* in *Kluwer Copyright Blog,* 9 maggio 2023, disponibile al sito:

carattere commerciale, la disciplina in tema di estrazione ha dunque un'efficacia piuttosto limitata[21].

Questa conclusione è rafforzata da un'altra considerazione. La creazione del *dataset* è un'attività complessa e può richiedere notevoli investimenti. Talora, ad occuparsi di questa fase del procedimento è un soggetto specializzato, diverso dall'impresa che sviluppa l'IA. È quanto accaduto nel caso di ChatGPT, che, almeno in una prima fase, è stato sviluppato da OpenAI sulla base di un *dataset* prodotto da Common Crawl. In un caso del genere, c'è quindi un'impresa che crea le copie, le immette in un *database* e poi trasferisce il *database* al programmatore dell'IA per l'addestramento. Come è noto, il trasferimento delle copie da un soggetto all'altro è oggetto di diritti diversi da quello di riproduzione, vale a dire dei diritti di distribuzione e di comunicazione al pubblico. L'eccezione

---

https://copyrightblog.kluweriplaw.com/2023/05/09/generative-ai-copyright-and-the-ai-act/.

[21] Da questo punto di vista, l'UE ha adottato un approccio più restrittivo rispetto a quello seguito da altri ordinamenti, in cui il "*text and data mining*" è tendenzialmente consentito anche per fini commerciali. Si v. ad es. sulla soluzione adottata in Giappone, UENO, *The Flexible Copyright Exception for "Non-Enjoyment" Purposes – Recent Amendment in Japan and Its Implication*, in *GRUR Int.,* 2022. V. anche per una comparazione tra sistema giapponese e direttive europee: DERWAMAN, *Text and data mining exceptions in the development of generative Ai models: what the EU member States could learn from the Japanese "nonenjoyment" purposes?,* in *J. of world IP,* 2023, 1. Nel Regno Unito, il governo ha avviato i lavori su un "*code of practice on copyright and AI*" volto ad elaborare soluzioni per semplificare l'acquisizione di licenze per l'estrazione a fini commerciali. Si v. le informazioni al sito: https://www.gov.uk/guidance/the-governments-code-of-practice-on-copyright-and-ai. Negli USA, la giurisprudenza ha talora qualificato l'uso delle opere per fini di estrazione come ipotesi di "*fair use*". L'uso è stato infatti considerato "trasformativo", i.e. "*one that communicates something new and different from the original or expands its utility, thus serving copyright's overall objective of contributing to public knowledge*": *Authors Guild v. Google, Inc.*, 804 F. 3d 202 (2d Cir. 2015), 214. V. anche *Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014). V. anche *Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 644–45 (4th Cir. 2009), relativo ad un sistema di analisi automatizzata del testo per finalità anti-plagio. Per un confronto della giurisprudenza statunitense con il diritto europeo si v. SCALZINI, *L'estrazione di dati e di testo per finalità commerciali dai contenuti degli utenti. Algoritmi, proprietà intellettuale e autonomia negoziale*, in *AGE,* 2019, 413 e ss. Per una panoramica comparatistica sul *text and data mining* v. anche OTTOLIA, *L'*opt-out commons *nella nuova disciplina del* data mining, in *Il diritto d'autore nel mercato unico digitale,* a cura di Cogo, in *Giur it.,* 2022, 1255.

L'approccio dell'UE è considerato eccessivamente restrittivo dalla maggior parte della dottrina europea. Si v., ad es., il Position Statement del Max Planck di Monaco, *Artificial intelligence and intellectual property law,* 9 aprile 2021, 7. V. anche DUCATO, STROWEL, *Ensuring text and data mining*, cit., 13 e ss.; MARGONI, KRETSCHMER, *A deeper look into the EU text and data mining exceptions,* cit., 685 e ss.; GEIGER, FROSIO, BULAYENKO, *Text and Data Mining in the Proposed Copyright Reform*, cit.; ROSATI, *Copyright as an Obstacle or an Enabler? A European Perspective on Text and Data Mining and its Role in the Development of AI Creativity*, in *Asia Pacific Law Review,* 2019, 199; GHIDINI, BANTERLE, *A critical view on the European Commission's proposal for a directive on copyright in the Digital Single Market,* in *Giur. comm.,* 2018, 961 e ss.

sull'"estrazione" riguarda però esclusivamente il diritto di riproduzione. Essa consente, quindi, soltanto la creazione delle copie, non il loro successivo trasferimento, che resta soggetto agli altri diritti esclusivi. Nell'ipotesi, piuttosto comune, in cui il creatore del *dataset* sia un soggetto diverso dal programmatore resta aperto il problema di raccogliere le autorizzazioni dei titolari delle opere coinvolte [22].

4. Riproduzioni temporanee.

La direttiva 2019/790 aggiunge, comunque, che "*vi possono essere anche casi di estrazione di testo e di dati* […] *in cui le riproduzioni effettuate rientrano nell'eccezione obbligatoria per gli atti di riproduzione temporanea di cui all'articolo 5, paragrafo 1, della direttiva 2001/29/CE, che dovrebbe continuare ad applicarsi alle tecniche di estrazione di testo e di dati che non comportino la realizzazione di copie al di là dell'ambito di applicazione dell'eccezione stessa*" (considerando 9).

---

[22] In questo senso, si v. FLYNN, GEIGER, QUINTAIS, MARGONI, SAG, GUIBAULT, CARROLL, *Implementing user rights for research in the field of artificial intelligence,* American University Washington College of Law research paper, 2020, 13; FRANCESCHELLI, MUSOLESI, *Copyright in generative deep learning,* in *Data & Policy,* 2022, 7 e ss; *Study on copyright and new technologies: copyright data management and artificial intelligence,* cit., 183. L'eccezione ai diritti sul trasferimento delle copie non pare potersi considerare implicita nell'eccezione al diritto di riproduzione. Il che si desume specialmente dall'art. 5, par. 4 della dir. 2001/29, in cui si afferma che "*quando gli Stati membri possono disporre un'eccezione o limitazione al diritto di riproduzione in virtù dei paragrafi 2 e 3 del presente articolo, essi possono anche disporre un'eccezione o limitazione al diritto di distribuzione di cui all'articolo 4 nella misura giustificata dallo scopo della riproduzione permessa*". Il che induce a pensare che, in mancanza di espressa previsione da parte del legislatore nazionale, l'eccezione al diritto di riproduzione si applichi soltanto alla creazione della copia. Peraltro, per l'"estrazione" la possibilità che gli Stati membri estendano l'eccezione alla distribuzione delle copie non è nemmeno contemplata nella direttiva.

Inoltre, come è noto, al "ritrasferimento" di un *file* digitale non si applica il principio dell'esaurimento. Il principio è infatti limitato dalla dir. 2001/29 al trasferimento di supporti "tangibili". V. considerando 29 dir. 2001/29: "*la questione dell'esaurimento del diritto non si pone nel caso di servizi, soprattutto di servizi «on-line». Ciò vale anche per una copia tangibile di un'opera o di altri materiali protetti realizzata da un utente di tale servizio con il consenso del titolare del diritto. Perciò lo stesso vale per il noleggio e il prestito dell'originale e delle copie di opere o altri materiali protetti che sono prestazioni in natura. Diversamente dal caso dei CD-ROM o dei CD-I, nel quale la proprietà intellettuale è incorporata in un supporto materiale, cioè in un bene, ogni servizio «on-line» è di fatto un atto che dovrà essere sottoposto ad autorizzazione se il diritto d'autore o i diritti connessi lo prevedono*". Il punto è stato recentemente confermato dalla Corte di Giustizia. Si v. CORTE DI GIUSTIZIA, 19 dicembre 2019, C-263/18, *Tom Kabinet,* par. 53 e ss.. L'applicazione dell'esaurimento alla diffusione digitale è comunque oggetto di discussione in dottrina. Si v. ad es. SGANGA, *Digital exhaustion after Tom Kabinet: a non-exhausted debate,* in *EU Internet law in the Digital Single Market,* edito da Synodinou, Jougleux, Markou, Prastitou-Merdi, Springer, 2021, 35 e ss.; MEZEI, *Copyright exhaustion,* Cambridge, 2018, 92 e ss.

Alla creazione di copie per fini di estrazione commerciale resta dunque potenzialmente applicabile l'eccezione di cui all'art. 5, par. 1 dir. 2001/29 (recepito in Italia con art. 68-*bis* l. aut.). Questa disposizione contiene un'eccezione per gli atti di riproduzione che soddisfano le seguenti condizioni cumulative: a) sono temporanei; b) sono eseguiti all'unico scopo di consentire la trasmissione in rete tra terzi con l'intervento di un intermediario oppure un utilizzo legittimo di un'opera; c) sono parte integrante ed essenziale di un procedimento tecnologico; d) sono transitori o accessori al procedimento; e) sono privi di rilievo economico proprio[23].

4.1. *Temporaneità.*

Il carattere "temporaneo" richiesto dalla condizione a) viene inteso in giurisprudenza nel senso che gli atti di riproduzione devono essere destinati alla cancellazione dopo un certo periodo di tempo[24]. Nel caso dell'IA, il requisito è in genere soddisfatto dai primi atti di *download* dei testi dal *web*. Queste copie servono infatti soltanto come base per creare delle ulteriori copie da inserire nel *dataset* in un formato leggibile dagli algoritmi. Una volta realizzate queste seconde copie, le prime divengono obsolete e possono essere eliminate. Lo stesso vale per le eventuali copie prodotte durante gli esercizi di addestramento. Queste sono infatti quasi sempre riproduzioni effimere, volte a facilitare l'esercizio dell'IA e sono automaticamente eliminate alla conclusione dell'esercizio stesso o, comunque, alla chiusura dell'addestramento[25].

Più complessa è invece la situazione delle copie che compongono il *dataset*. Qui può, innanzitutto, accadere che il *dataset* venga interamente eliminato alla fine della procedura di addestramento dell'IA. Il requisito della "temporaneità" sarebbe, in tal caso, senz'altro rispettato[26]. Potrebbe, però, anche accadere che il *dataset* sia conservato per un tempo più lungo, ad es., per consentire controlli o aggiornamenti periodici del sistema. La questione se copie come queste siano "temporanee" non è risolta

---

[23] La disposizione non è espressamente richiamata tra le eccezioni applicabili al diritto degli editori di giornali di cui all'art. 15 dir. 2019/790. Essa pare comunque applicabile al diritto in questione. L'art. 15 dir. 2019/790 afferma infatti che "*gli Stati membri riconoscono agli editori di giornali stabilito in uno Stato membro i diritti di cui all'articolo 2* [...] *della direttiva 2001/29/CE per l'utilizzo online delle loro pubblicazioni di carattere giornalistico da parte di prestatori di servizi della società dell'informazione*. Il contenuto del diritto è definito con un rinvio all'art. 2 dir. 2001/29 (i.e. il diritto di riproduzione degli autori), il quale è soggetto all'eccezione di cui all'art. 5 par. 1.

[24] CORTE DI GIUSTIZIA, 5 giugno 2014, C-360/13, *Public Relations,* par. 26.

[25] VESALA, *Developing artificial intelligence-based content creation: are EU copyright and antitrust law fit for purpose?,* in *IIC,* 2023, 361.

[26] SCHÖNBERGER, *Deep copyright: Up - and downstream questions related to artificial intelligence (AI) and machine learning (ML),* in *Droit d'auteur 4.0,* edito da De Werra, 2018, 145.

espressamente dalla direttiva. La *ratio* del requisito di temporaneità è però quella di consentire le riproduzioni indispensabili allo sviluppo di un procedimento tecnologico ed evitare di "liberalizzare" copie che invece vanno oltre questo obiettivo. La fase di revisione è un passaggio indispensabile per assicurare il buon funzionamento di un procedimento. Se le copie necessarie per la revisione non fossero consentite, la disposizione non riuscirebbe a realizzare efficacemente la propria funzione. Tali copie sembrano dunque doversi considerare, a loro volta, compatibili con il requisito di "temporaneità", sempre che la loro durata sia strettamente collegata alla realizzazione di controlli e aggiornamenti.

Fin qui si è trattato di casi in cui la produzione del *dataset* è collegata alla realizzazione di un unico procedimento di addestramento. Si è detto, però, che la realizzazione di un buon *dataset* è un'attività complessa che può richiedere investimenti anche notevoli. Nella maggior parte dei casi, quindi, il *dataset* è costruito per essere utilizzato in una pluralità di procedimenti, anche futuri[27]. Talora, esso è addirittura sviluppato da operatori specializzati che intendono offrirlo ai terzi sul mercato. In tutti questi casi, il *dataset* è prodotto per durare. E ciò esclude la possibilità di applicare l'art. 5, par. 1[28].

Un'eccezione a questo discorso può aversi, tutt'al più, nel caso in cui il contenuto del *dataset* venga periodicamente rinnovato. Può accadere, infatti, che i documenti all'interno del *dataset* siano sostituiti dopo un certo tempo con testi nuovi, per assicurare diversità al materiale d'addestramento. In queste circostanze, il *dataset* in quanto tale è stabile, ma le copie che lo compongono hanno durata limitata[29]. In ogni caso, come l'eccezione di estrazione, anche l'art. 5, par. 1 si riferisce soltanto al diritto di riproduzione. La disposizione può quindi consentire la creazione del *dataset*, ma non il suo eventuale trasferimento da un'impresa che lo produce ad una che lo utilizza. Questo passaggio resta soggetto ai diritti di distribuzione o di comunicazione al pubblico.

4.2. Utilizzo legittimo.

Le copie temporanee devono poi essere volte soltanto a consentire una trasmissione in rete tra terzi oppure un utilizzo legittimo delle opere (condizione b). Il procedimento dell'IA non è un'ipotesi di trasmissione tra

---

[27] LEMLEY, PRINTER, *Fair Learning,* in *Texas L. Rev.,* 2021, 753.

[28] In questo senso, MARGONI, KRETSCHMER, *A deeper look into the EU text and data mining exceptions,* cit., 693. SENFTLEBEN, *Compliance of national TDM rules with international copyright law: an overrated nonissue?,* in *IIC,* 2022, 1483. V. anche FRANCESCHELLI, MUSOLESI, *Copyright in generative deep learning,* in *Data & Policy,* 2022, e17.

[29] VESALA, *Developing artificial intelligence-based content creation: are EU copyright and antitrust law fit for purpose?,* in *IIC,* 2023, 361 e ss.

terzi. Resta dunque da vedere se le copie realizzate in questo procedimento possano dirsi funzionali a realizzare un utilizzo "legittimo" delle opere.

Secondo la Corte, un utilizzo è da considerare legittimo in due casi: se è autorizzato dal titolare o se non rientra tra le attività riservate dalla legge ai titolari[30]. Nel caso in esame, l'utilizzo che il procedimento punta a realizzare è l'analisi delle opere per l'apprendimento del linguaggio da parte dell'IA. Come già visto, in genere, questo utilizzo non viene autorizzato dai titolari. Non resta, allora, che domandarsi se l'analisi automatizzata di un testo rientri tra le attività che la legge sottopone all'esclusiva dell'autore.

La questione non è risolta espressamente dalla legge. In genere, la mera "fruizione" di un'opera non è considerata oggetto dei diritti d'autore. Così la lettura, l'analisi e lo studio di un'opera letteraria sono tradizionalmente considerate utilizzazioni libere. Secondo gran parte della dottrina, ciò dovrebbe valere anche per le forme di "lettura" automatizzata poste in essere da una macchina[31].

Quest'impostazione è però criticata da alcuni autori. La lettura artificiale sarebbe radicalmente distinta dalla lettura umana. La macchina non si limita a "fruire" dell'opera come un qualunque essere umano, ma estrae dall'analisi un valore "informativo", che viene poi generalmente reimpiegato in attività commerciali. L'analisi in questione non potrebbe essere qualificata come ipotesi di "mero godimento", ma sarebbe a tutti gli effetti "utilizzazione economica" dell'opera[32].

Quest'ultima tesi appare oggi smentita dalla disciplina, già richiamata, sull'estrazione di testo e di dati di cui alla direttiva 2019/790. L'obiettivo dichiarato di queste eccezioni è quello di liberalizzare le ipotesi di analisi computazionale che rispettano una serie di condizioni. Le eccezioni si applicano, però, soltanto alle "riproduzioni" necessarie

---

[30] CORTE DI GIUSTIZIA, 26 aprile 2017, C-527/15, *Stichting Brein,* par. 65 e ss.; CORTE DI GIUSTIZIA, 4 ottobre 2011, cause riunite C-403/08, C-429/08, *FA Premier League,* par. 168; CORTE DI GIUSTIZIA, 17 gennaio 2012, C-302/10, *Infopaq,* par. 44 e ss.

[31] Si v. HILTY, RICHTER, *Position statement of the Max Planck Institute on the modernisation of European copyright rules,* 2017, 2 e ss.; MUSSO, *Eccezioni e limitazioni ai diritti d'autore nella direttiva UE n. 790/2019,* in *Dir. Informazione e dell'informatica,* 2020, 411 e ss.; TRIAILLE, DE MEEÛS D'ARGENTEUIL, DE FRANCQUEN, *Study on the legal framework of text and data mining,* cit., 31; LITMAN, *The exclusive right to read,* in *Cardozo Art & Ent. Law J.,* 1994, 29.

[32] In questo senso pare di poter leggere OTTOLIA, *L'utilizzo computazionale dell'opera dell'ingegno in Internet,* in *AIDA,* 2014, 386 e ss. La tesi è poi sviluppata ulteriormente in OTTOLIA, Big data *e innovazione computazionale*, cit., 19 e ss. V. più recentemente OTTOLIA, *L'*opt-out commons *nella nuova disciplina del* data mining, cit., 1254 e ss. Si v. anche sul punto SARTI, *Diritti esclusivi e circolazione dei beni,* Giuffrè, 1996, 359 e ss., in cui l'A. sottolinea che l'argomento del "mero godimento" non può essere utilizzato per giustificare utilizzazioni che aumentano la concorrenzialità delle organizzazioni produttive. Il riferimento è specialmente allo sfruttamento del *software.* Secondo SERVANZI, *Le estrazioni di testo e di dati,* in *NLCC,* 2022, 1152, questo ragionamento sarebbe applicabile anche all'uso delle opere nel *text and data mining*.

all'analisi. Non all'analisi in quanto tale. Da ciò si desume che, secondo il legislatore, l'analisi artificiale del testo non è oggetto di esclusiva e, pertanto, non richiede apposite eccezioni[33].

D'altra parte, il diritto d'autore si fonda sul principio per cui la protezione si estende solo alla "espressione", cioè al modo in cui l'idea creativa è espressa dall'autore, alla "forma" dell'opera. Non sono invece tutelate le idee e le informazioni contenute nell'opera[34]. L'autore non può

---

[33] A conferma di questa lettura si v. il considerando 9 dir. 2019/790/UE: "*L'estrazione di testo e di dati può essere effettuata anche in relazione a semplici fatti o dati non tutelati dal diritto d'autore, nel qual caso non è richiesta alcuna autorizzazione in base alla legislazione sul diritto d'autore. Vi possono essere anche casi di estrazione di testo e di dati che non comportano atti di riproduzione o in cui le riproduzioni effettuate rientrano nell'eccezione obbligatoria per gli atti di riproduzione temporanea* […]". Non convince dunque la critica, mossa da alcuni, secondo cui l'eccezione di estrazione avrebbe l'effetto di estendere indirettamente il diritto d'autore fino a coprire atti di mera "lettura" prima sottratti all'esclusiva. Sul punti si v. ad es. MARGONI, KRETSCHMER, *A deeper look into the EU text and data mining exceptions,* cit., 693. Va detto che l'art. 3 dir. 2019/790 include espressamente tra le attività liberalizzate anche le "estrazioni". Questo però si spiega con il fatto che la disposizione introduce anche un'eccezione al diritto *sui generis* sulle banche dati, il quale, come è noto, ha proprio ad oggetto l'estrazione di informazioni.

Su questi temi, una qualche ambiguità si potrebbe creare se fosse adottato il regolamento sull'intelligenza artificiale (AI Act) nel testo provvisorio che è attualmente in circolazione. Qui si legge infatti, nel considerando 60i, che "*text and data mining techniques may be used extensively in this context for the retrieval and analysis of such content, which may be protected by copyright and related rights. Any use of copyright protected content requires the authorization of the rightholder concerned unless relevant copyright exceptions apply*". Il che sembrerebbe presupporre che anche la mera analisi sia attività riservata. Per le ragioni esposte *infra* nel testo, questa lettura "estensiva" dell'esclusiva non pare però compatibile con i principi generali del diritto d'autore.

In Francia, è stata recentemente presentata una proposta di riforma volta, tra l'altro, a sottoporre l'addestramento dell'IA al diritto esclusivo dei titolari. V. art. 1 della *Proposition de loi* n. 1630 - *Visant à encadrer l'intelligence artificielle par le droit d'auteur*, pres. 12 settembre 2023: "*l'intégration par un logiciel d'intelligence artificielle d'œuvres de l'esprit protégées par le droit d'auteur dans son système et a fortiori leur exploitation est soumise aux dispositions générales du présent code et donc à autorisation des auteurs ou ayants droit*". La proposta prevede, addirittura, che i diritti sui contenuti generati dall'IA siano assegnati ai titolari dei diritti sulle opere che hanno reso possibile il risultato creativo. V. art. 2: "*lorsque l'œuvre est créée par une intelligence artificielle sans intervention humaine directe, les seuls titulaires des droits sont les auteurs ou ayants droit des œuvres qui ont permis de concevoir ladite œuvre artificielle*". La proposta è stata criticata duramente in dottrina: si v. ad es. GEIGER, IAIA, *The forgotten creator,* cit., 7 e ss.

[34] Il principio è codificato in diverse disposizioni dell'ordinamento. Nella l. aut. nazionale, si trova espresso all'art. 2, n. 8-9, con riferimento ai *software* e alle banche dati. A livello europeo, è richiamato all'art. 1 della dir. 91/250 sui programmi per elaboratore e agli artt. 3 e 5 della dir. 96/9 sulla protezione delle banche dati. Il principio è anche ribadito dal considerando 9 della dir. 2019/790, in cui si legge che "*l'estrazione di testo e di dati può essere effettuata anche in relazione a semplici fatti o dati non tutelati dal diritto d'autore, nel qual caso non è richiesta alcuna autorizzazione in base alla legislazione sul diritto d'autore*". Il principio che sottrae le idee alla tutela autoriale ha trovato poi applicazione anche nella giurisprudenza della Corte di Giustizia. Si v. recentemente CORTE DI GIUSTIZIA, 11 giugno

impedire l'utilizzo, anche per fini commerciali, delle informazioni, degli insegnamenti e dei concetti tratti dalla propria opera[35]. L'analisi computazionale ha la funzione di estrarre dall'opera informazioni che sarebbero inaccessibili attraverso la normale elaborazione umana[36]. Riconoscere all'autore il potere di impedire l'analisi computazionale significa sostanzialmente impedire ai terzi lo sfruttamento di queste informazioni. Significa, quindi, estendere l'esclusiva a campi che la legge intende sottrarre al controllo dell'autore[37].

Si potrebbe obiettare che l'analisi computazionale estrae dati non tanto dal contenuto, ma dal testo dell'opera. Essa sarebbe quindi pur sempre un'utilizzazione della "forma espressiva" dell'opera. E, come tale, dovrebbe rientrare nell'ambito dell'esclusiva[38]. Occorre intendersi, però, sul significato del concetto di "espressione" utilizzato dal legislatore. La forma espressiva è il veicolo che consente all'autore di comunicare le proprie idee e di trasmettere emozioni, sentimenti e riflessioni. Sono questi gli elementi che spingono il pubblico a "consumare" un libro o un articolo e che, quindi, determinano l'esistenza di un mercato dell'opera letteraria. Da questo punto di vista, la "forma espressiva" deve effettivamente essere soggetta al controllo dell'autore. La forma di un'opera non è però soltanto veicolo di comunicazione della personalità del suo autore. Essa ha anche una ricca componente di informazioni: il testo contiene indicazioni sulla grammatica, sul significato delle parole, sul modo in cui esse vanno utilizzate, sulla frequenza delle soluzioni espressive, ecc. Se queste informazioni fossero oggetto dell'esclusiva, si giungerebbe al risultato assurdo di assegnare ad un autore il potere di opporsi all'uso altrui di una certa lingua. La

---

2020, C-833/18, *Brompton Bicycle,* par. 27; CORTE DI GIUSTIZIA, 22 dicembre 2010, C-393/09, *BSA,* par. 48 e ss.; CORTE DI GIUSTIZIA, 2 maggio 2012, C-406/10, *SAS,* par. 31 e ss. Infine, il principio è presente nelle convenzioni internazionali in materia di diritto d'autore. Si v. art. 9 par. 2 TRIPs e art. 2 WCT.

[35] In questo senso, si v. M. BERTANI, *Diritto d'autore europeo,* cit., 109, secondo cui il principio per cui il diritto d'autore non si estende al sapere teorico si spiega, tra l'altro, con l'esigenza di consentire l'innovazione concorrente e il progresso culturale.

[36] GRANIERI, *Il data mining nella disciplina del diritto d'autore e la strategia europea sui dati,* in *AIDA,* 2022, 24; ROSSI, *Opere dell'ingegno come dati: il text and data mining nella direttiva 2019/790,* in *AIDA,* 2019, 235.

[37] In questo senso, si v. anche LEMLEY, PRINTER, *Fair Learning,* in *Texas L. Rev.,* 2021, 750 e MARGONI, KRETSCHMER, *A deeper look into the EU text and data mining exceptions,* cit., 689. Il punto sembra condiviso anche dalla giurisprudenza statunitense che si è occupata degli usi "trasformativi" delle opere dell'ingegno in campo digitale. Si v. ad es. il caso *Authors Guild v. Google* (2015) F. 3d 202, in cui si legge, con riferimento alle funzioni di ricerca nelle opere e agli *snippet*, che "*the copyright resulting from the Plaintiffs' authorship of their works does not include an exclusive right to furnish the kind of information about the works that Google's programs provide to the public. For substantially the same reasons, the copyright that protects Plaintiffs' works does not include an exclusive derivative right to supply such information through query of a digitized copy*". V. anche *Thomson Reuters, v. Ross Intelligence, Inc.,* 28 settembre 2023, 20-cv-613-SB (D. Del. Sep. 28, 2023), 24.

[38] SOBEL, *Artificial intelligence's fair use crisis,* in *Col. J. of law & the arts,* 2017, 46 e ss.

componente "informativa" del testo cade dunque fuori dall'ambito dell'esclusiva.

Questa distinzione tra "forma" come veicolo di comunicazione e forma come veicolo di informazione è stata, a lungo, superflua. La raccolta di informazioni sulla lingua da parte del lettore di un libro è, infatti, attività inscindibile dalla fruizione della espressione creativa dell'autore. Ciò giustifica, ad es., che la vendita di un romanzo ad uno studente interessato ad usarlo per esercitarsi con la lingua sia attività soggetta all'esclusiva dell'autore. I due aspetti diventano invece scindibili nel caso dell'IA: gli algoritmi che "leggono" il testo non percepiscono l'espressione creativa dell'autore e non capiscono le sue scelte creative; essi si limitano ad estrapolare le informazioni statistiche sull'uso del linguaggio che sono racchiuse nelle scelte espressive. In sostanza, l'IA fa uso dell'espressione soltanto come veicolo di informazioni, non come veicolo di comunicazione creativa[39]. La lettura artificiale del testo non sembra dunque potersi qualificare come attività riservata all'autore.

L'idea che l'addestramento, in sé, sia un utilizzo legittimo potrebbe essere, ancora, contestata sulla base del fatto che, una volta concluso il procedimento, l'IA potrebbe teoricamente utilizzare il linguaggio appreso per produrre testi in violazione del diritto d'autore[40]. E questo, ovviamente, non sarebbe un utilizzo "legittimo" delle opere dell'ingegno. Le copie temporanee non sarebbero allora qui realizzate "all'unico scopo" di consentire un utilizzo legittimo e l'art. 5, par. 1 non sarebbe applicabile all'IA generativa. In sostanza, l'obiezione in esame si fonda sull'idea che l'art. 5, par. 1 non possa essere applicato ad un utilizzo legittimo che può eventualmente portare ad un successivo utilizzo illegittimo.

La tesi non considera, però, che, in realtà, qualsiasi utilizzo dell'opera, anche legittimo, è astrattamente idoneo a causare successivi utilizzi illegittimi: la lettura di un testo dal sito può portare l'utente a scaricare una copia del testo sul computer da diffondere poi su altri siti; la trasmissione lecita di un videoclip su una piattaforma può essere utilizzata per effettuare registrazioni abusive; e via dicendo. In sostanza, ad un procedimento di utilizzo legittimo può sempre seguire un diverso procedimento illegittimo. Dire che l'art. 5, par. 1 è applicabile solo quando

---

[39] Considerazioni simili paiono espresse in BORGHI, KARAPAPA, *Non-display uses of copyright works: Google books and beyond,* in *Queen Mary J. of IP,* 2011, 44 e ss.; LEMLEY, PRINTER, *Fair Learning,* in *Texas L. Rev.,* 2021, 749. V. anche le riflessioni sulla forma espressiva in SPEDICATO, *Interesse pubblico e bilanciamento del diritto d'autore,* Giuffrè, 2013, 156 e ss.

[40] L'argomento sembra sollevato dall'editore nel caso *Thomson Reuters, v. Ross Intelligence, Inc.,* 28 settembre 2023, 20-cv-613-SB (D. Del. Sep. 28, 2023), 19. V. anche sul punto FRANCESCHELLI, MUSOLESI, *Copyright in generative deep learning,* in *Data & Policy,* 2022, 8; VESALA, *Developing artificial intelligence-based content creation,* cit., 362; KARAPAPA, *Defences to copyright infringement,* Oxford UP, 2020, 112 e ss.

non vi sia alcun rischio di successivi utilizzi illegittimi significa allora sostanzialmente dire che la disposizione non può mai trovare applicazione.

È vero, per altro verso, che il caso dell'IA presenta una particolarità rispetto agli esempi sopra richiamati. L'eventuale violazione del diritto d'autore nella generazione di testo da parte dell'IA non è un procedimento del tutto autonomo rispetto a quello di addestramento, ma è uno sviluppo o, comunque, un'applicazione dello stesso procedimento. In altre parole, utilizzo legittimo e utilizzo illegittimo si verificano qui nell'ambito di procedimenti tecnologici collegati e consequenziali. Si pone dunque la questione se l'art. 5, par. 1 sia applicabile quando il successivo utilizzo illegittimo dell'opera provenga da un'applicazione o da uno sviluppo del procedimento legittimo.

La questione non è risolta espressamente dalla direttiva. Come riconosciuto dalla Corte di Giustizia, l'obiettivo dell'art. 5, par. 1 è però quello di "*consentire e assicurare lo sviluppo ed il funzionamento di nuove tecnologie, nonché mantenere un giusto equilibrio tra i diritti e gli interessi dei titolari di diritti e degli utilizzatori delle opere protette che intendano beneficiare di tali tecnologie*". L'IA può avere innumerevoli applicazioni utili e virtuose. Dire che l'art. 5, par. 1 non si applica in questo campo significa, di fatto, impedire la realizzazione dei risultati di benessere generale che queste linee di innovazione potrebbero realizzare. Se letta in questo senso, la disposizione finirebbe dunque per dare assoluta prevalenza agli interessi degli autori rispetto all'interesse generale. Il che non pare coerente con l'obiettivo di realizzare un giusto equilibrio tra gli interessi in gioco. Tanto più che, se l'IA, una volta sviluppata, violasse i diritti d'autore ricopiando un'opera, l'autore resterebbe comunque legittimato ad esercitare i suoi diritti di riproduzione e di comunicazione al pubblico nei confronti dei testi generati dall'IA. Per assicurare tutela ai titolari contro gli utilizzi illegittimi, non c'è allora bisogno di impedire del tutto lo sviluppo del procedimento.

L'art. 5, par. 1 pare dunque doversi piuttosto leggere nel senso che un procedimento, come quello dell'IA, che abbia come risultato un utilizzo "legittimo" delle opere può rientrare nell'eccezione, a prescindere dal fatto che lo stesso procedimento possa poi avere anche alcune applicazioni illegittime, le quali ultime restano invece soggette alla comune disciplina protettiva del diritto dell'autore.

4.3. Parte integrante ed essenziale del procedimento.

La condizione c) (i.e. che le copie siano parte integrante ed essenziale del procedimento) viene intesa nella giurisprudenza europea nel senso che gli atti di riproduzione devono risultare necessari affinché il procedimento

funzioni efficacemente e devono essere interamente compiuti nell'ambito del procedimento stesso[41].

Le copie funzionali alla costruzione di un *dataset* sono necessarie per il buon funzionamento dell'addestramento, visto che questa non può prescindere dal contatto con i testi per apprendere il linguaggio umano[42].

Più articolata è invece la risposta alla domanda se il *dataset* sia realizzato interamente "all'interno" del procedimento. Il requisito pare doversi intendere nel senso che il procedimento non possa servirsi di copie prodotte in precedenza per altri fini, ma soltanto di copie prodotte in occasione del procedimento stesso. Pertanto, il requisito non è soddisfatto se il *dataset* viene creato in vista di una pluralità di applicazioni future o, addirittura, per essere venduto. Lo è invece nel caso in cui il *dataset* sia costruito appositamente per uno specifico processo di addestramento.

4.4. Transitorietà e accessorietà.

Ai sensi della condizione d), le copie devono essere poi, alternativamente, transitorie o accessorie al procedimento. La disposizione è piuttosto ambigua, in quanto il concetto di "transitorietà" sembra già implicito nel requisito della "temporaneità". La Corte sembra però intendere i concetti in maniera diversa: temporaneità significa che la copia deve avere una durata limitata; "transitorietà", invece, significa che le copie devono essere cancellate automaticamente una volta esaurito il proprio ruolo nel procedimento[43]. In quest'ottica, possono esserci quindi copie "temporanee" e non "transitorie": è il caso in cui per le copie è previsto un termine di durata, ma la loro cancellazione richiede un intervento manuale dell'uomo.

Nel campo dell'IA, possono essere soggette ad eliminazione automatica le copie iniziali prodotte dal *download* dei testi dal *web* e quelle effimere eventualmente necessarie per lo svolgimento di esercizi di *training.* Lo stesso non vale sempre per il *dataset,* la cui eliminazione dipende dall'esito del procedimento ed è dunque, in genere, rimessa alle scelte dei programmatori.

---

[41] CORTE DI GIUSTIZIA, 17 gennaio 2012, C-302/10, *Infopaq,* par. 30; CORTE DI GIUSTIZIA, 5 giugno 2014, C-360/13, *Public Relations,* par. 28.

[42] Affinché il requisito sia rispettato, basta che le copie rendano il processo più efficace. Si v. CORTE DI GIUSTIZIA, 5 giugno 2014, C-360/13, *Public Relations,* par. 35 e ss. Sul tema, v. anche GUGLIELMETTI, *Riproduzione e riproduzione temporanea,* in *AIDA,* 2002, 35 e ss.

[43] CORTE DI GIUSTIZIA, 16 luglio 2009, C-5/08, *Infopaq,* par. 62; CORTE DI GIUSTIZIA, 5 giugno 2014, C-360/13, *Public Relations,* par. 40: "*un atto può essere qualificato come «transitorio» esclusivamente qualora la sua durata sia limitata a quanto necessario per il buon funzionamento del procedimento tecnologico utilizzato, restando inteso che tale procedimento deve essere automatizzato in modo tale da cancellare detto atto in maniera automatica, senza intervento umano, nel momento in cui è esaurita la sua funzione tesa a consentire la realizzazione di un siffatto procedimento*".

Per queste copie, occorre dunque chiedersi se ricorra il requisito alternativo dell'"accessorietà". Anche il significato di "accessorietà" non è affatto chiaro. Il concetto sembra infatti sovrapporsi con il requisito per cui le copie devono essere parte integrante del procedimento. La Corte interpreta il requisito di "accessorietà" nel senso che le copie, in quanto tali, non devono avere "*né un'esistenza né una finalità autonome rispetto al procedimento*" [44]. In quest'ottica, fra i due requisiti c'è allora una differenza. Come già visto, "essere parte integrante del procedimento" si riferisce alla "nascita" della copia, nascita che deve avvenire in occasione del procedimento. "Accessorietà" si riferisce invece alla "vita" successiva della copia. La copia prodotta deve poter essere usata soltanto per il procedimento e non deve poter esistere al di fuori del procedimento[45]. Per soddisfare il requisito, il *dataset* deve dunque essere costruito con *standard* o altre misure tecnologiche che ne vincolino l'uso al solo addestramento dell'IA.

4.5. Mancanza di rilievo economico autonomo.

Infine, l'eccezione richiede che le copie siano prive di rilievo economico proprio (condizione e). Secondo la Corte, ciò significa che la copia temporanea non deve essere in grado di generare un vantaggio economico "distinto e separabile" rispetto al vantaggio economico realizzato con l'utilizzo legittimo delle opere che è l'esito del procedimento[46]. La Corte, in genere, considera soddisfatto il requisito se

---

[44] CORTE DI GIUSTIZIA, 5 giugno 2014, C-360/13, *Public Relations,* par. 47 e ss.

[45] GUGLIELMETTI, *Riproduzione e riproduzione temporanea,* in *AIDA,* 2002, 35.

[46] CORTE DI GIUSTIZIA, 17 gennaio 2012, C-302/10, *Infopaq,* par. 50 e ss. Qui la Corte afferma che ha rilievo economico proprio anche la riproduzione che comporta "*una modifica dell'oggetto riprodotto, quale esistente al momento dell'avvio del procedimento tecnologico interessato, poiché i suddetti atti sono, in tal caso, diretti a facilitare non già il suo utilizzo, ma l'utilizzo di un oggetto diverso*". L'affermazione della Corte potrebbe essere letta nel senso che qualsiasi modifica nel formato o nel "linguaggio" della copia sia una modifica dell'opera e ricada fuori dall'ambito dell'eccezione. In tal caso, la riproduzione temporanea realizzata ai fini dell'addestramento si collocherebbe fuori dall'ambito di applicazione dell'art. 5, par. 1, visto che il procedimento in questione comporta generalmente la traduzione delle opere in un linguaggio digitale, per consentire la lettura degli algoritmi.

Questa lettura della sentenza non pare però convincente. Quasi sempre i procedimenti tecnologici telematici si fondano sulla creazione di copie in formati diversi da quelli originari. D'altra parte, la creazione di una copia temporanea serve, in molti casi, al solo scopo di cambiare forma all'opera e di consentirne così usi altrimenti impossibili. Se ogni cambiamento di formato o di "linguaggio" fosse sufficiente ad escludere l'applicazione dell'art. 5 par. 1, il diritto d'autore finirebbe per precludere numerosi procedimenti tecnologici innovativi. Il che pare entrare in conflitto con la *ratio* di fondo dell'art. 5, par. 1, cioè la tutela dell'efficienza dinamica del mercato digitale (CORTE DI GIUSTIZIA, 5 giugno 2014, C-360/13, *Public Relations,* par. 23). D'altra parte, una discriminazione tra procedimenti fondati su copie identiche all'originale e procedimenti fondati su copie basate su formati diversi non pare giustificata dall'esigenza di tutelare gli interessi degli

l'unica utilità che la copia può generare consiste in una maggiore efficienza del procedimento in cui è inserita. Il che si verifica quando la copia è inseparabile dal procedimento. In altri termini, la copia è priva di rilievo economico quando non può essere né condivisa con terzi né usata in altri procedimenti.

Inteso in questo senso, il requisito appare senz'altro soddisfatto nel caso in cui le copie siano transitorie. Queste copie sono automaticamente distrutte una volta esaurito il loro ruolo nel procedimento. Esse non si prestano quindi ad essere condivise con i terzi. Potrebbero, tutt'al più, essere adoperate per il funzionamento di altri sistemi tecnologici controllati dal creatore del procedimento principale. Questi sistemi andrebbero però sincronizzati con quello principale per essere contemporanei, avere la stessa velocità di calcolo e disporre della stessa durata. Appare piuttosto improbabile che ciò sia fattibile per procedimenti completamente diversi da quello principale. È più verosimile che si tratti di procedimenti ancillari e collegati a quel procedimento. In questo caso, il vantaggio economico prodotto dalla copia sarebbe quindi pur sempre riconducibile al procedimento principale.

Nel campo dell'IA possono allora essere considerati privi di rilievo economico i *download* iniziali che siano soggetti a forme di cancellazione automatica, nonché le copie effimere eventualmente prodotte durante il *training*.

Come già detto, le copie non transitorie possono ancora essere giustificate se risultano accessorie. E ciò si verifica quando le copie possono essere utilizzate soltanto all'interno del procedimento. Anche queste copie dovrebbero, allora, essere incapaci di produrre un ricavo autonomo rispetto al procedimento in cui sono usate. In questo senso, però, la "mancanza di rilievo economico proprio" sarebbe una mera ripetizione di requisiti già esistenti e sarebbe quindi una condizione del tutto inutile.

Per risolvere questo problema occorre una precisazione sul significato di "accessorietà". Si è detto che "accessorietà" significa che la copia deve poter essere usata soltanto per un procedimento. Uno stesso procedimento può essere, però, realizzato più volte o da più soggetti diversi. Per es., un'impresa potrebbe ripetere più volte l'addestramento della propria IA oppure potrebbero esserci più imprese che applicano lo stesso tipo di procedimento per addestrare IA diverse. Il requisito può

---

autori. Questi interessi giustificherebbero, anzi, conclusioni opposte: la produzione di copie identiche all'originale è infatti verosimilmente più problematica per gli autori rispetto alla creazione di una copia in un formato pensato per la lettura artificiale.

Sembra dunque preferibile ritenere che la Corte abbia inteso qui escludere l'applicazione dell'art. 5, par. 1 ai casi in cui la copia modifichi la "sostanza" dell'opera, producendo, cioè, un'elaborazione che incide sul suo contenuto espressivo. In questo caso, in effetti, consentire la copia potrebbe finire per "liberalizzare" atti di elaborazione che gli ordinamenti nazionali generalmente riservano all'autore.

essere allora letto in due modi. L'accessorietà al procedimento può essere intesa, in primo luogo, nel senso che la copia deve poter essere utilizzata soltanto nello specifico procedimento in cui è creata. Se fosse intesa in questo senso, però, una copia accessoria non potrebbe proprio essere utilizzata al di fuori di un determinato procedimento e non ci sarebbe, quindi, alcuna possibilità che essa generi un ricavo autonomo rispetto al procedimento stesso. L'accessorietà implicherebbe necessariamente mancanza di rilievo economico autonomo delle copie. E quest'ultimo requisito sarebbe, appunto, inutile.

In alternativa, l'accessorietà al procedimento può essere intesa nel senso che la copia deve poter essere utilizzata soltanto in un certo tipo di procedimento. In quest'ottica, è accessoria una copia che può essere utilizzata sia nel procedimento in cui è creata che in altre esecuzioni dello stesso procedimento poste in essere dal produttore o da altri operatori. Una copia del genere è suscettibile di più applicazioni ed è, quindi, anche in grado di generare un vantaggio economico indipendente rispetto al singolo procedimento in cui è creata. In quest'ottica, c'è allora una differenza tra "accessorietà" e "mancanza di rilievo economico": una copia "accessoria" può essere usata in una pluralità di procedimenti dello stesso tipo; una copia "priva di rilievo economico" può essere utilizzata soltanto nello specifico procedimento in cui è creata.

Da tutto questo discorso deriva che, per rispettare il requisito e), un *dataset* non deve soltanto essere incompatibile con utilizzi diversi dall'addestramento dell'IA. Deve, invece, addirittura, avere caratteristiche tali da poter essere applicato soltanto nello sviluppo di una determinata IA. Non basta allora che il *dataset* sia realizzato con un formato standard applicabile, in generale, a tutti i processi di addestramento. Il produttore dovrà costruire il *dataset* con formati, matrici o misure tecnologiche che assicurino che le copie siano utilizzabili soltanto dai propri algoritmi e che diventino obsolete dopo il loro primo utilizzo.

4.6. Considerazioni di sintesi.

In sintesi, l'eccezione di cui all'art. 5, par. 1 può essere applicata alle copie funzionali all'addestramento dell'IA se ricorrono le seguenti circostanze:

- le copie iniziali, prodotte al momento del *download,* e quelle generate durante il *training* sono eliminate automaticamente dopo la creazione delle versioni dei testi da inserire nel *dataset;*
- il *dataset* è costruito in occasione di uno specifico procedimento di addestramento.
- il *dataset* non è utilizzabile al di fuori di tale procedimento e non può essere offerto sul mercato;

- il *dataset* viene conservato solo per il tempo strettamente necessario a consentire l'addestramento dell'IA ed eventuali revisioni del sistema.

La disposizione non consente dunque alle imprese di IA di "rifornirsi" dai terzi. L'eccezione, infatti, non copre l'ipotesi in cui un operatore sviluppi il *dataset* per offrirlo ai programmatori di IA. Resta quindi sostanzialmente ostacolata la nascita di un mercato dei *dataset.* Cade poi anche fuori dall'eccezione il caso in cui il *dataset* sia creato da un'impresa su commissione del programmatore. In sostanza, il programmatore di IA deve occuparsi di tutta l'attività "a monte" dell'addestramento, e cioè della ricerca delle fonti, dello scaricamento dei testi, dell'adattamento dei file e della costruzione del *database.* Il che impedisce di beneficiare delle efficienze che derivano dall'emersione di imprese specializzate nelle fasi iniziali della catena e comporta un'inefficiente duplicazione delle reti di raccolta dei dati e dei sistemi di elaborazione dei *dataset.*

Inoltre, l'impresa che sviluppa l'IA non può costruire un *dataset* stabile e compatibile con una pluralità di procedimenti diversi. L'operazione di raccolta e di adattamento dei dati deve quindi sostanzialmente essere ripetuta per ogni procedimento. Questa soluzione porta con sé un'enorme duplicazione di costi. E ciò a tacere del problema di impatto "ambientale" derivante dalla ripetizione di operazioni potenzialmente dispendiose anche sul piano energetico.

In questo campo, l'art. 5, par. 1 non riesce dunque a realizzare efficacemente la sua funzione di conciliare la tutela del diritto d'autore con le esigenze dell'innovazione tecnologica. Nel momento in cui l'art. 5, par. 1 viene adottato, all'inizio degli anni Duemila, la tecnologia telematica è sostanzialmente concepita come un'immensa biblioteca di informazioni da ricercare, elaborare e scambiare. In quest'ottica, per consentire l'innovazione digitale basta, in effetti, "liberalizzare" le copie effimere necessarie alla navigazione degli utenti, alla visualizzazione dei contenuti e alla trasmissione delle informazioni.

Lo scenario è però cambiato radicalmente negli ultimi anni. La tecnologia digitale non è più soltanto un "oceano" passivo di informazioni su cui navigare. Con l'IA, la tecnologia digitale è diventata, a sua volta, un protagonista attivo della comunicazione telematica. L'IA è capace di esaminare i dati esistenti e di elaborarli, producendo informazioni e soluzioni nuove. E questo consente di raggiungere risultati che sarebbero impensabili in un mondo in cui protagonisti attivi della comunicazione sono soltanto gli utenti umani. Affinché questo nuovo protagonista del digitale possa funzionare è, però, necessario che le informazioni esistenti siano ad esso "comprensibili". Da un punto di vista tecnico, questo significa che i contenuti della rete devono essere copiati e trasformati in

un linguaggio accessibile alla macchina. E, come si è visto, la liberalizzazione delle copie effimere non è più sufficiente a questo fine.

In sintesi, l'art. 5, par. 1 è una disposizione pensata per un mondo che è ormai cambiato ed è oggi incompatibile con la realtà del mercato digitale. Ci sarebbero allora, teoricamente, le premesse per un tentativo di interpretazione evolutiva della disposizione. Questa via non sembra, però, percorribile. V'è, in primo luogo, il problema che l'art. 5, par. 1 si fonda sul requisito che le copie siano "temporanee", mentre l'IA ha bisogno di copie (almeno in certa misura) "permanenti". Per adeguare la disposizione a questa esigenza, si dovrebbe assegnare al termine "temporaneo" un significato che questo non può proprio assumere[47]. Un'interpretazione evolutiva dell'eccezione richiederebbe poi che le esigenze della realtà attuale non siano state già prese in considerazione dal legislatore. Tuttavia, come già visto, la direttiva 2019/790 ha espressamente disciplinato la materia attraverso le eccezioni per l'estrazione di dati, che realizzano una liberalizzazione limitata degli atti di riproduzione. Gli ostacoli che l'IA affronta nell'attuale quadro normativo non dipendono allora tanto da un ritardo nell'aggiornamento delle regole, ma sembrano il frutto di una vera e propria scelta del legislatore. Una lettura dell'art. 5, par. 1 nel senso che siano consentite senza limiti tutte le riproduzioni necessarie per addestrare un'IA si porrebbe in contraddizione con questa scelta.

5. *I confini del diritto di riproduzione.*

L'idea che il sistema delle eccezioni al diritto di riproduzione non offra soluzioni soddisfacenti per l'IA è condivisa da gran parte della dottrina. E ciò ha contribuito a rafforzare la convinzione, da tempo diffusa, che, per riconciliare il diritto d'autore con gli obiettivi generali di efficienza dinamica del mercato digitale, sia necessaria una modifica delle direttive[48].

Alcuni autori propongono di espandere le eccezioni al diritto d'autore, vuoi introducendo una clausola generale di esenzione, sulla falsariga del "*fair use*" statunitense, vuoi aggiungendo una specifica

---

[47] Sul punto v. anche BORGHI, KARAPAPA, *Copyright and mass digitization,* Oxford UP, 2013, 59; GHIDINI, *Proprietà intellettuale e innovazione digitale. Dalla "interferenza antitrust" a un nuovo paradigma?,* in *Giur. Comm.,* 2023, 367 e ss.

[48] Si v. ad es. HUGENHOLTZ, *The new copyright directive: text and data mining (articles 3 and 4),* in *Kluwer Copyright Blog,* 24 luglio 2019; FROSIO, *Should we ban generative AI, incentivize it or make it a medium for inclusive creativity?,* su ssrn.com, 2023, 12; ROSSI, *Opere dell'ingegno come dati*, cit., 247 e ss.; MONTAGNANI, AIME, *Il text and data mining e il diritto d'autore*, cit.; CHRISTENSEN, *A European solution for text and data mining in the development of creative artificial intelligence,* in *Stockholm IP L. Rev.,* 2021, 18 e ss.; CASO, *Il conflitto tra diritto d'autore e ricerca scientifica nella disciplina del* text and data mining *della direttiva sul mercato unico digitale,* Trento Law and Technology Research Group, 2020.

eccezione che consenta tutte le forme di riproduzione necessarie per il funzionamento di sistemi tecnici innovativi, come l'IA[49].

In base ad un'altra impostazione, anziché intervenire sulle eccezioni, occorrerebbe modificare il diritto di riproduzione, limitandolo alle copie che sono destinate alla distribuzione ai consumatori finali. Dovrebbero essere, cioè, sottratte all'esclusiva tutte le copie c.d. "intermedie", cioè le copie che sono prodotte per essere adoperate all'interno di un processo tecnologico[50].

Secondo alcuni autori, questo risultato sarebbe, peraltro, già raggiungibile a livello interpretativo. Per riproduzione di un'opera si intende la moltiplicazione della stessa in copie. Il concetto di copia non è però definito nelle direttive e deve essere interpretato alla luce dell'obiettivo del diritto di riproduzione di assegnare all'autore un controllo sulla diffusione commerciale degli esemplari presso il pubblico. "Copie" sarebbero, allora, soltanto gli esemplari che sono destinati, in un modo o nell'altro, alla circolazione presso i consumatori finali; non quelli volti ad agevolare il funzionamento di procedimenti tecnici, che nulla hanno a che vedere con la "vita" commerciale dell'opera. La creazione di queste ultime copie e il loro eventuale trasferimento non richiederebbero quindi alcuna autorizzazione da parte dei titolari[51].

---

[49] Si v. tra gli altri GEIGER, IAIA, *The forgotten creator,* cit., 10 e ss.; GHIDINI, *Proprietà intellettuale e innovazione digitale,* cit., 367 e ss. V., con riferimento in generale, agli usi digitali HUGENHOLTZ, SENFTLEBEN, *Fair use in Europe: in search of flexibilities,* IVIR Report, Amsterdam, 2011; HEARGRAVES, *Digital opportunity,* Report, Maggio 2011, 46 e ss.; GEIGER, IZYUMENKO, *Towards a European "fair use" grounded on the freedom of expression,* in *American Univ. Int. L. Rev.,* 2019, 1 e ss.

[50] V. MARGONI, KRETSCHMER, *A deeper look into the EU text and data mining exceptions,* cit., 693 e ss.; DUCATO, STROWEL, *Ensuring text and data mining,* cit., 24 e ss., secondo cui l'espansione del diritto di riproduzione alle copie "intermedie" non è compatibile con gli obiettivi di fondo del diritto d'autore. V. in senso simile BORGHI, KARAPAPA, *Copyright and mass digitization,* Oxford UP, 2013, 52 e ss. e 153 e ss., in cui gli A. esprimono proposte simili, effettuando anche un confronto dettagliato con la disciplina generale in tema di uso dei dati. In generale, l'idea che la tradizionale esclusiva sulla riproduzione richieda un ripensamento nel mondo digitale è da tempo diffusa a livello internazionale. Si v. ad es. LITMAN, *Real copyright reform,* in *Iowa L. Rev.,* 2010, 41 e ss.; LEMLEY, *Dealing with overlapping copyrights on the Internet,* in *Univ. Daytona L. Rev.,* 1997, 22 e ss.; SPADA, *La proprietà intellettuale nelle reti telematiche,* in *Riv. dir. civ.,* 1998, 636 e ss.

[51] Si v. SCHÖNBERGER, *Deep copyright: Up - and downstream questions*, cit., 13; *Study on copyright and new technologies: copyright data management and artificial intelligence,* cit., 182 e ss.; SENFTLEBEN, *Compliance of national TDM rules with international copyright law*, cit., 1483 e ss., in cui la questione è affrontata dal punto di vista delle convenzioni internazionali in materia di diritto d'autore. L'idea che le copie meramente tecniche cadano al di fuori dell'esclusiva è, comunque, presente nella dottrina europea anche prima della rivoluzione dell'IA. Si v. ad es. HUGENHOLTZ, *Caching and copyright. The right of temporary copying*, in *EIPR,* 2000, 482 e ss.; SCHIUMA, *Diritto d'autore e normativa europea,* in *Treccani – diritto online,* 2009; MUSSO, *Diritto di autore sulle opere dell'ingegno letterarie ed artistiche,* in *Comm. Scialoja-Branca,* Zanichelli, 2008, 212 e ss. Sull'esigenza di effettuare una valutazione di carattere

Quest'approccio è stato, a sua volta, criticato. Le tesi in esame fanno dipendere la qualifica di "copia" dallo scopo per cui un esemplare è prodotto. Il diritto di riproduzione è oggi definito dalle direttive come "*il diritto esclusivo di autorizzare o vietare la riproduzione diretta o indiretta, temporanea o permanente, in qualunque modo o forma, in tutto o in parte*". La definizione, molto ampia, non contiene alcun riferimento alla funzione per cui una copia viene prodotta. Il fatto che essa sia realizzata per uno scopo commerciale o per uno scopo tecnico sarebbe dunque irrilevante[52]. Peraltro, le copie oggetto delle eccezioni di estrazione e di riproduzione temporanea sono sempre copie "intermedie" rispetto a procedimenti tecnici. Se fosse vero che queste copie si collocano fuori dall'esclusiva, le relative eccezioni sarebbero disposizioni inutili. L'idea che la nozione di "copia" dipenda dallo scopo della riproduzione sembra, dunque, incompatibile con le attuali regole sul diritto d'autore.

Di queste tesi appare però condivisibile l'idea che il concetto di "copia" debba essere interpretato tenendo conto degli obiettivi dell'esclusiva sulla riproduzione. Le ricostruzioni sopra richiamate partono dall'idea che il diritto di riproduzione abbia la funzione di assegnare all'autore il controllo sulla successiva circolazione delle copie sul mercato[53]. Storicamente, il diritto di riproduzione si giustifica

---

"teleologico" in tema di riproduzioni digitali, v. ROMANO, *L'opera e l'esemplare nel diritto della proprietà intellettuale,* CEDAM, 2001, 186 e ss.

In tema di IA, la soluzione viene spesso sostenuta richiamando la distinzione che viene effettuata negli Stati Uniti tra utilizzi dotati di "*expressive purposes*" e utilizzi dotati di "*non-expressive purposes*". Questi ultimi vengono talora qualificati come "*transformative uses*". Da ciò parte della dottrina trae la conclusione che le copie intermedie realizzate per l'addestramento siano giustificate dalla clausola di "*fair use*". V. in questo senso, tra gli altri, SAG, *Copyright safety for generative AI,* in *Houston L. Rev.,* 2023, 104 e ss.; LEMLEY, PRINTER, *Fair Learning,* in *Texas L. Rev.,* 2021, 744 e ss.. V. per un'opinione, in parte, diversa SOBEL, *Artificial intelligence's fair use crisis,* in *Col. J. of law & the arts,* 2017, 46 e ss. La questione è oggetto dei procedimenti in corso. V. ad es. *Thomson Reuters, v. Ross Intelligence, Inc.,* 28 settembre 2023, 20-cv-613-SB (D. Del. Sep. 28, 2023). Una parte della dottrina statunitense ha adoperato argomenti simili per sostenere che gli esemplari privi di "*expressive purposes*" si collocano del tutto al di fuori del diritto di riproduzione. CARROLL, *Copyright and the progress of science: why text and data mining is lawful,* in *U.C. Davis L. Rev.,* 2019, 894 e ss., in cui si distingue tra "*copies that count*" e "*copies that don't count*". V. in senso simile QUANG, *Does training AI violate copyright law?,* in *Berkeley Tech. L. J.,* 2021, 1407 e ss.

[52] OTTOLIA, *L'utilizzo computazionale dell'opera dell'ingegno in Internet*, cit., 394 ss. Il punto è confermato dalla giurisprudenza europea. V. CORTE DI GIUSTIZIA, 4 ottobre 2011, cause riunite C-403/08, C-429/08, *FA Premier League,* 159; CORTE DI GIUSTIZIA, 24 marzo 2022, C-433/20, *Austro Mechana,* par. 16 e ss.

[53] Più precisamente, l'esclusiva sulla riproduzione è ricollegata in dottrina all'esigenza di assegnare all'autore il potere di decidere il numero degli esemplari in circolazione. Il che gli consente di esercitare un potere "monopolistico" sul mercato dell'opera, influenzando il prezzo degli esemplari. Si v. ROMANO, *L'opera e l'esemplare nel diritto della proprietà intellettuale,* CEDAM, 2001, 159; SARTI, *Diritti esclusivi e circolazione dei beni,* Giuffrè, 1996, 358 e ss., in cui, più precisamente, l'A. individua in via interpretativa l'esistenza di un

soprattutto con il fatto che è più semplice agire nei confronti delle stamperie abusive, rispetto ad agire nei confronti dei successivi atti di smercio delle copie. Atti, questi, che possono essere numerosi e difficili da individuare[54]. Peraltro, questa funzione non riguarda soltanto la distribuzione delle copie materiali, ma vale anche per molte altre forme di sfruttamento commerciale dell'opera. Basti pensare, ad es., alla registrazione abusiva di un'opera musicale. Questa può essere distribuita sul mercato sotto forma di CD, può essere messa in onda sulla radio o sulla televisione, può essere diffusa nei pubblici esercizi, ecc. Agendo "a monte" contro l'atto di registrazione, l'autore evita di dover agire contro tutte le possibili utilizzazioni della stessa.

Questo discorso potrebbe, in effetti, portare alla conclusione che le "copie", ai sensi del diritto di riproduzione, siano soltanto quelle destinate allo sfruttamento commerciale e, cioè, alla diffusione presso il pubblico. Tuttavia, il diritto di riproduzione ha subito un'evoluzione nel corso tempo.

A partire dagli anni '60, si diffondono le tecnologie di fotocopia e di registrazione analogica. Ciò semplifica la produzione di copie destinate al mercato, ma consente anche al grande pubblico di creare copie delle opere per uso personale a costi marginali. Il dilagare di queste copie "private" rischia di ridurre la domanda delle copie "commerciali". Con il tempo, il diritto di riproduzione finisce quindi per coprire anche le copie destinate ad usi "non commerciali". Data la difficoltà pratica di esercitare l'esclusiva nei confronti degli usi "privati", il diritto esclusivo in questo campo viene ben presto sostituito con diritti al compenso affidati alla gestione delle *collecting societies.* L'obiettivo è, in ogni caso, quello di proteggere il mercato dell'opera dalla concorrenza data dalla diffusione dell'attività privata di copia. Per il diritto d'autore sono dunque copie anche quelle destinate ad uno sfruttamento "domestico" [55].

Con la rivoluzione digitale, poi, l'attività di copia diventa ancora più accessibile. A questo punto, chiunque può partire dalla copia di un'opera per produrre un numero infinito di esemplari a costi pressoché nulli. Con Internet diventa poi anche molto semplice trasmettere le copie al pubblico. Ogni copia può facilmente essere caricata e diffusa su diversi siti-*web*. Gli utenti di questi siti possono, a loro volta, creare nuove copie

---

diritto a stabilire la quantità di prodotti destinati al mercato. V. anche AUTERI, *Diritto di autore,* in AA.VV., *Diritto industriale,* Giappichelli, 2023, 717 e ss.

[54] SPOOR, *The copyright approach to copying on the Internet: (over)stretching the reproduction right?*, in *The future of copyright in the digital environment,* edito da Hugenholtz, Wolters Kluwer, 1996, 77.

[55] Su questa evoluzione si v. RICOLFI, *Il diritto d'autore,* in *Tr. dir. comm.,* diretto da Cottino, vol. II, CEDAM, 2001, 415; SARTI, *Copia privata e diritto d'autore,* in *AIDA,* 1992, 35 e ss. AUTERI, *Diritto di autore,* in AA.VV., *Diritto industriale,* Giappichelli, 2023, 717 e ss.; SPOOR, *The copyright approach to copying on the Internet: (over)stretching the reproduction right?,* cit., 70 e ss.

da diffondere, e così via. In sostanza, sul *web* una singola copia può dare luogo ad un'infinita serie di trasmissioni. Trasmissioni, peraltro, molto difficili da individuare e da contrastare. Ogni copia digitale porta quindi con sé un pericolo per i mercati "ufficiali" dell'opera. In questo contesto, il diritto di riproduzione viene esteso fino a coprire, come si è visto, anche le copie "intermedie". Queste copie hanno funzioni meramente tecniche e non sono destinate né alla diffusione commerciale né alla fruizione privata degli utenti. Eppure, nel mondo digitale esse possono essere facilmente distolte dalla loro funzione originaria e trasmesse al pubblico. Di qui l'esigenza di sottoporle al controllo dell'autore[56]. Queste copie vengono consentite soltanto in presenza di condizioni che ne rendano impossibile lo sfruttamento commerciale. È, appunto, il caso dell'art. 5, par. 1 sulle copie temporanee.

Il diritto di riproduzione ha dunque assunto nel tempo funzioni diverse. Il problema di fondo che esso affronta è, però, sempre lo stesso. Se fosse libera la produzione di copie, sarebbe impossibile per il titolare esercitare un controllo sul mercato dell'opera. Il mercato verrebbe facilmente "inondato" di esemplari non autorizzati e i diritti esclusivi sulle varie forme di sfruttamento commerciale sarebbero, di fatto, svuotati di efficacia. Il diritto di riproduzione serve quindi a proteggere i mercati che la legge riserva all'autore. In questo senso, si tratta di un diritto strumentale rispetto a tutti gli altri diritti esclusivi dell'autore.

È alla luce di questa funzione che pare doversi leggere il concetto di "copia": per il diritto d'autore "copia" non è soltanto l'esemplare prodotto per la distribuzione commerciale, ma qualsiasi esemplare che abbia le caratteristiche per entrare in concorrenza con le forme di sfruttamento economico che la legge riserva al titolare, indipendentemente dallo scopo per cui l'esemplare è originariamente creato[57].

---

[56] RICOLFI, *Il diritto d'autore*, cit., 418 e ss. L'estensione del diritto di riproduzione alle copie incidentali è oggetto di accese discussioni durante la negoziazione delle convenzioni internazionali in materia. Si v. per una ricostruzione SENFTLEBEN, *Compliance of national TDM rules with international copyright law,* cit., 1483 e ss. Sull'evoluzione della nozione di riproduzione con l'avvento del digitale si v. ROMANO, *L'opera e l'esemplare nel diritto della proprietà intellettuale,* CEDAM, 2001, 159 e ss. e 181 e ss., in cui si mette in luce il fatto che, con la tecnologia telematica, la distinzione tradizionale tra creazione della copia e sua successiva distribuzione viene meno. Fenomeno, questo, che diventa particolarmente evidente in materia di protezione dei programmi per elaboratore. Si v. anche SCHIUMA, *Diritto d'autore e normativa europea,* in *Treccani – diritto online,* 2009, secondo cui la disintermediazione dell'attività di copia resa possibile dalle tecnologie digitale giustifica un ripensamento nell'estensione del diritto di riproduzione.

[57] Si v. sul punto RICOLFI, *Il diritto d'autore,* cit., 415 e ss.; MUSSO, *Diritto di autore sulle opere dell'ingegno letterarie ed artistiche,* in *Comm. Scialoja-Branca,* Zanichelli, 2008, 205; HUGENHOLTZ, *Caching and copyright. The right of temporary copying*, cit., 482 e ss., anche per riferimenti ad opere precedenti. Si v. anche SENFTLEBEN, *Compliance of national TDM rules with international copyright law,* cit., 1497, secondo cui simili proposte di definizione del concetto di "copia" sono più volte emerse nel dibattito internazionale.

Tornando all'IA, si tratta dunque di capire se le copie che vengono prodotte per l'addestramento siano idonee a "minacciare" i mercati oggetto di esclusiva. A questo proposito, va detto che i sistemi di IA sono molto numerosi e presentano significative differenze. Il loro funzionamento è spesso segreto e non è dunque possibile conoscere con precisione ogni aspetto del procedimento. È perciò impossibile dare alla domanda una risposta definitiva, valida per ogni sistema di IA. Si può però svolgere qualche osservazione di taglio generale, da cui trarre criteri per l'applicazione al caso concreto.

Se nel *dataset* i testi sono conservati nel linguaggio e nel formato originario o, comunque, in un formato facilmente utilizzabile per fini diversi dall'addestramento, i testi sono potenzialmente utilizzabili sul mercato della diffusione dell'opera al pubblico. In questo caso, non c'è dubbio quindi che il *dataset* sia composto da "copie" ai sensi del diritto d'autore.

Nella maggior parte dei casi, però, l'addestramento dell'IA non avviene su *file* di questo genere. Il testo nella sua forma originaria contiene molti elementi che "disturbano" la comprensione degli algoritmi dell'IA, come gli articoli, la punteggiatura, le espressioni non chiare, gli errori, ecc. Peraltro, i formati leggibili per l'uomo non si prestano ad un'analisi efficace da parte degli algoritmi. Per ottimizzare il processo di estrazione di dati, innanzitutto, i testi devono essere suddivisi in unità più piccole, ad es., parole o espressioni. È il c.d. processo di "tokenizzazione" del testo. Le singole unità sono poi talora "asciugate" da alcune variabili che complicano la raccolta di dati. È ad es. il caso del verbo coniugato che può essere riportato ad una versione base, come il verbo all'infinito. Inoltre, a seconda dei casi, le unità di testo possono essere mantenute nell'ordine originario oppure "mescolate" con unità simili provenienti da altri testi. Generalmente, le parole vengono poi tradotte in vettori matematici per la lettura algoritmica. Questa complessa attività di intervento sul testo viene chiamata "normalizzazione" o "*pre-processing*"[58] ed è volta ad ottimizzare l'elaborazione dei dati da parte dell'algoritmo.

A prima vista, la conclusione raggiunta nel testo si scontra con l'art. 68 l. aut., in cui si legge che "*è libera la riproduzione di singole opere o brani di opere per uso personale dei lettori, fatta a mano o con mezzi di riproduzione non idonei allo spaccio o diffusione dell'opera nel pubblico*". Il che pare presupporre che, in linea di principio, anche le riproduzioni inidonee alla distribuzione commerciale siano "copie" per il diritto d'autore. In realtà, l'eccezione non pare incoerente con quanto sopra detto. Le copie dell'art. 68 non sono idonee al commercio per il mezzo con cui sono realizzate, ma sono pur sempre esemplari leggibili da parte del pubblico. In teoria, esse possono circolare tra gli utenti ed interferire, così, con la domanda delle copie commerciali. In questo senso, esse possono effettivamente rientrare nel concetto di "copia" sopra richiamato. La potenzialità lesiva di queste copie è poi, di fatto, molto ridotta e ciò giustifica l'esistenza di un'apposita eccezione.

[58] MONTAGNANI, AIME, *Il text and data mining e il diritto d'autore,* cit., 378.

Benintesto, il "*pre-processing*" può avvenire in modi diversi. Ci sono casi in cui la lavorazione è limitata e il testo resta quindi, almeno in certa misura, utilizzabile per fini diversi dall'apprendimento artificiale. In molti casi, invece, il testo è sottoposto ad un profondo processo di manipolazione, fino a trasformare la forma originaria delle opere in immense serie numeriche. Queste forme di intervento sottraggono alla copia i connotati tipici di un testo destinato alla fruizione del pubblico. C'è chi parla, al riguardo, di una vera e propria trasformazione dell'opera in "dato". Ed in effetti il *pre-processing* può anche essere visto come una prima forma di produzione di metadati sull'opera.

In questi ultimi casi, il file non può essere immediatamente utilizzato dagli utenti per accedere all'opera originaria. A tal fine, l'utente dovrebbe disporre della matrice di decodifica e dovrebbe conoscere con precisione tutte le fasi di manipolazione che il testo ha subito. Peraltro, anche se l'utente disponesse di queste informazioni, la ricompilazione di un *dataset* sarebbe difficile da realizzare con i comuni programmi operativi. Teoricamente, si potrebbe realizzare utilizzando specifici programmi di riconversione. Questa operazione, richiederebbe, comunque, notevole spazio di memorizzazione e strumenti tecnici dotati di elevata capacità di calcolo. Per l'applicazione di sofisticati sistemi di decodificazione è anche necessaria una preparazione tecnica in tema di programmazione e di analisi dei dati. Infine, i processi di normalizzazione sono molto vari, sicché l'uso di programmi di riconversione standard non può sempre garantire l'efficace trasformazione del file in un testo fruibile. Dal punto di vista del pubblico, un file soggetto a profondi processi di normalizzazione non sembra quindi potersi considerare, di per sé, sostituibile alle normali copie digitali.

Una volta assunta la forma "*pre-processed*", il file può essere utilizzato soltanto per l'analisi da parte degli algoritmi dell'IA. Questa è, a sua volta, una forma di utilizzazione commerciale del testo. Come già detto, però l'analisi computazionale ha ad oggetto soltanto l'opera come veicolo di informazioni, non l'espressione creativa dell'autore. Si tratta quindi di un'attività che cade fuori dall'ambito dell'esclusiva[59]. I file normalizzati, in quanto tali, non interferiscono allora con i mercati che la legge riserva all'autore. In questo senso, si potrebbe dire che tali file non rientrano nel concetto di "copia" ai sensi del diritto d'autore[60].

---

[59] In questo senso, pare di poter leggere il ragionamento dell'ordinanza *Thomson Reuters, v. Ross Intelligence, Inc.,* 28 settembre 2023, 20-cv-613-SB (D. Del. Sep. 28, 2023): "*if Ross's characterization of its activities is accurate, it translated human language into something understandable by a computer as a step in the process of trying to develop a "wholly new," albeit competing, product — a search tool that would produce highly relevant quotations from judicial opinions in response to natural language questions. This also means that Ross's final product would not contain or output infringing material*".

[60] Indicazioni in questo senso (riferite però al diritto connesso del produttore di fonogrammi) paiono potersi leggere in CORTE DI GIUSTIZIA, 29 luglio 2019, C-476/19,

Questa conclusione si espone ad alcune obiezioni. Se è vero che il file "normalizzato" non è direttamente accessibile per gli utenti, l'algoritmo dell'IA è però tendenzialmente in grado di risalire al testo originario. Esso ha infatti la capacità computazionale per individuare le matrici e per ricostruire le varie fasi di manipolazione e di interpolazione subite dal testo di base. Se programmata per questo scopo, un'IA è quindi in grado di ritrasformare un file normalizzato in un testo fruibile e di trasmetterlo al pubblico[61].

Si pone qui in sostanza la questione se rientri nel concetto di riproduzione la creazione di una copia di per sé inutilizzabile, che diventa però fruibile se sottoposta ad ulteriori passaggi tecnici.

Questo fenomeno non è, in realtà, una novità dell'IA. Nella diffusione dei contenuti c.d. *torrent*, la copia subisce una scomposizione in fase di trasmissione e viene poi "riassemblata" per consentire il *download* e l'accesso all'opera da parte degli utenti. D'altra parte, anche il CD audio, da solo, non è in grado di comunicare l'opera; se combinato con un apposito strumento di "lettura" diventa però a tutti gli effetti veicolo dell'opera. Queste situazioni vengono generalmente considerate riproduzioni ai sensi del diritto d'autore[62]. Il che depone a favore della tesi secondo cui anche un file "*pre-processed*" rientra nel concetto di copia.

Pelham, par. 31: "*quando un utente, nell'esercizio della libertà delle arti, preleva un campione sonoro da un fonogramma al fine di utilizzarlo, in una forma modificata e non riconoscibile all'ascolto, in una nuova opera, si deve ritenere che un utilizzo del genere non costituisca una «riproduzione», ai sensi dell'articolo 2, lettera c), della direttiva 2001/29*".

[61] Considerazioni simili sono espresse, seppur con riferimento alle prime forme di sviluppo del mercato digitale, in GUGLIELMETTI, *Riproduzione e riproduzione temporanea,* in *AIDA*, 2002, 17 e ss.

[62] In realtà, la trasmissione *torrent* è stata affrontata dalla giurisprudenza europea soprattutto dal punto di vista della comunicazione al pubblico interattiva. Si v. CORTE DI GIUSTIZIA,14 giugno 2017, C-610/15, *Stichting Brein* e soprattutto CORTE DI GIUSTIZIA, 17 giugno 2021, C-597/19, *Mircom.* In quest'ultimo caso si affrontava, tra l'altro, la questione se l'utente che partecipa ad una trasmissione *torrent* commetta atti di violazione del diritto d'autore. La questione pregiudiziale faceva riferimento esclusivamente al diritto di comunicazione al pubblico. Nella sua motivazione, la Corte afferma inizialmente che "*alla Corte spetta, se necessario, riformulare le questioni che le sono sottoposte. Infatti, la Corte ha il compito di interpretare tutte le disposizioni del diritto dell'Unione che possano essere utili ai giudici nazionali al fine di dirimere le controversie di cui sono investiti, anche qualora tali disposizioni non siano espressamente indicate nelle questioni a essa sottoposte da detti giudici*". È interessante notare che la Corte non richiama a questo proposito le disposizioni sul diritto di riproduzione. Essa conclude nel senso che "*costituisce una messa a disposizione del pubblico, ai sensi di tale disposizione, il caricamento, a partire dall'apparecchiatura terminale di un utente di una rete tra pari (peer-to-peer) verso apparecchiature terminali di altri utenti di tale rete, dei segmenti, previamente scaricati da detto utente, di un file multimediale contenente un'opera protetta, benché tali segmenti siano utilizzabili da soli soltanto a partire da una determinata percentuale di scaricamento*".

I file normalizzati dell'IA presentano, però, delle peculiarità rispetto a queste situazioni. Le copie *torrent* e le riproduzioni meccaniche sono pur sempre copie prodotte allo scopo di realizzare la diffusione dell'opera al pubblico. La forma "intermedia" che si assegna alla copia è, anzi, proprio funzionale a rendere più efficace o più rapida la comunicazione del suo contenuto. In questi casi, si può dunque presumere che alla produzione e alla diffusione della copia facciano seguito atti di sfruttamento dell'opera riservati all'autore.

Lo stesso non pare potersi dire per i file normalizzati dell'IA. Questi file non sono costruiti per consentire la trasmissione al pubblico dei contenuti. Al contrario, il loro formato rende la comunicazione del contenuto più difficile. Questa comunicazione è pure tecnicamente possibile, ma richiede un'azione ulteriore che è avulsa rispetto alla finalità per cui il file è originariamente generato e che è volta proprio ad indirizzare la copia verso una finalità diversa. In queste circostanze, si pone, allora, quanto meno il problema se sia corretto intervenire al momento della produzione del file o piuttosto al momento in cui si verifica il passaggio tecnico che cambia la "vocazione" del file, rendendolo idoneo a ledere gli interessi dei titolari. E, in ogni caso, la riconversione del file è un passaggio del tutto eventuale, tanto più che allo stato si tratta di un'operazione che non sembra alla portata di un utente qualsiasi. Qui non si può quindi presumere che alla creazione del file segua il suo sfruttamento per fini di trasmissione dell'opera al pubblico. Infine, anche se l'utilizzo per scopi di trasmissione del testo fosse fattibile per tutti gli utenti della rete, resta il fatto che questi file sono prodotti per uno scopo diverso e legittimo, vale a dire quello di consentire l'analisi algoritmica. Rispetto alle altre copie, le quali, salvo rare eccezioni, sono sempre destinate ad utilizzi coperti da esclusiva, questi file sono suscettibili di diversi utilizzi, alcuni riservati e altri legittimi.

C'è poi anche un'altra considerazione che vale a distinguere il caso dell'IA dalle copie *torrent* e dalle riproduzioni meccaniche sopra richiamate. Proprio perché queste ultime sono copie destinate a trasmettere l'opera al pubblico, gli applicativi tecnici necessari per "leggerle" sono in genere alla portata di tutti. Pertanto, se la creazione delle copie fosse libera, per il titolare dei diritti sarebbe poi estremamente difficile agire nei confronti dei successivi atti di sfruttamento, che potrebbero anche essere posti in essere direttamente dagli utenti finali in maniera "decentrata". In sintesi, il pericolo che copie del genere ledano gli interessi dell'autore è piuttosto elevato e ciò giustifica l'anticipazione dell'esclusiva al momento della creazione della copia. Anche questo argomento non sembra valido, però, in tema di IA. Qui l'atto che rende fruibile l'opera per il pubblico richiede sempre l'intervento "centrale" di un'IA o, comunque, di un servizio digitale in grado di riconvertire il testo. Questo atto di conversione del file da parte di un'IA comporta la creazione

di una versione fruibile dell'opera ed è quindi sicuramente da riguardare come un atto di "copia" coperto dall'esclusiva. Peraltro, se alla decodifica segue la diffusione del testo agli utenti si ha anche un atto di comunicazione al pubblico, anch'esso oggetto di esclusiva. In sintesi, un'IA che fosse programmata per restituire agli utenti opere da leggere commetterebbe atti di violazione dei diritti d'autore. A fronte di un uso "abusivo" dei file normalizzati, il titolare non sarebbe allora costretto ad agire nei confronti di una massa dispersa di soggetti, ma potrebbe agire direttamente nei confronti dell'operatore dell'IA[63].

[63] Questa conclusione sembra valida soprattutto nel caso in cui l'IA sia programmata per replicare testi "memorizzati" in fase di addestramento. L'IA si comporterebbe qui in maniera non molto diversa da un qualunque sito di trasmissione di opere "pirata". Si tratta di un problema che è emerso nell'ambito delle controversie tra editori e ChatGPT. In particolare, certi editori affermano che i meccanismi statistici su cui si fonda ChatGPT fanno sì che il sistema replichi integralmente gli articoli di giornale usati in fase di addestramento, se l'utente fornisce come input le prime righe del testo. In questo modo, ChatGPT diventa un meccanismo per aggirare le restrizioni poste dagli editori per i consumatori non abbonati al sito.

Qualche dubbio sulla responsabilità dell'operatore di IA potrebbe porsi invece nell'eventualità in cui il servizio fosse programmato semplicemente per convertire *file* normalizzati forniti dagli utenti in testi facilmente fruibili dal pubblico. A prima vista, in questo caso, l'IA si limita a fornire la infrastruttura tecnica, mentre l'atto di *upload* viene effettuato dagli utenti del servizio. La fattispecie sembrerebbe allora avvicinarsi alle situazioni in cui un operatore offre in maniera meramente passiva un mezzo di comunicazione che viene poi adoperato dagli utenti per violare diritti d'autore. In realtà, se è vero che in questo caso l'iniziativa dell'atto di trasmissione è presa dall'utente, resta fermo il fatto che la comunicazione finale del testo fruibile viene effettuata dal sistema di IA. Comunque, appare anche discutibile qualificare come infrastruttura meramente passiva un sistema di IA appositamente programmato per offrire servizi di riconversione di file in testi fruibili dal pubblico. Sembra più convincente ritenere che un operatore che offre un'infrastruttura del genere al pubblico sia, a sua volta, responsabile della comunicazione resa possibile dal proprio servizio. In questo senso, con riferimento alle piattaforme "*peer to peer*" v. CORTE DI GIUSTIZIA, 14 giugno 2017, C-610/15, *Stichting Brein,* par. 36 e ss.: "*le opere così messe a disposizione degli utenti della piattaforma di condivisione online TPB sono state messe online su tale piattaforma non dagli amministratori di quest'ultima, bensì dai suoi utenti. Tuttavia detti amministratori, mediante la messa a disposizione e la gestione di una piattaforma di condivisione online, come quella di cui al procedimento principale, intervengono con piena cognizione delle conseguenze del proprio comportamento, al fine di dare accesso alle opere protette, indicizzando ed elencando su tale piattaforma i file torrent che consentono agli utenti della medesima di localizzare tali opere e di condividerle nell'ambito di una rete tra utenti (peer-to-peer). A tale riguardo* […] *senza la messa a disposizione e la gestione da parte dei suddetti amministratori di una siffatta piattaforma, le opere in questione non potrebbero essere condivise dagli utenti o, quantomeno, la loro condivisione su Internet sarebbe più complessa. Occorre pertanto considerare che, con la messa a disposizione e la gestione della piattaforma di condivisione online TPB, gli amministratori di quest'ultima offrono ai loro utenti un accesso alle opere di cui trattasi. Si può quindi ritenere che essi svolgano un ruolo imprescindibile nella messa a disposizione delle opere in questione*". V. anche in tema di piattaforme CORTE DI GIUSTIZIA, 22 giugno 2021, cause riunite C-682/18 e C-683/18, *Peterson.* In CORTE DI GIUSTIZIA, 26 aprile 2017, C-527/15, *Stichting Brein,* par. 53 si legge anche che il diritto di comunicazione al pubblico "*ricomprende la vendita di un lettore multimediale, come quello di cui al procedimento principale,*

Resta dunque aperta la questione se sia copia ai sensi del diritto d'autore una versione dell'opera che: a) non è immediatamente in grado di trasmettere l'espressione creativa dell'autore al pubblico; b) può dare luogo indirettamente ad alcuni utilizzi concorrenti con lo sfruttamento commerciale dell'opera; c) ha però come funzione principale quella di consentire attività di raccolta e di analisi di dati, che cadono fuori dal campo di applicazione del diritto d'autore.

Una prima possibile soluzione del problema è quella di interpretare il diritto di riproduzione nel senso che il rischio, anche remoto, che la creazione di un file possa portare ad una successiva interferenza con i mercati dell'opera basta a concludere che il suddetto file rientra nel campo dell'esclusiva.

Questa interpretazione solleva però qualche perplessità. L'analisi computazionale dell'IA consente di raccogliere e di utilizzare informazioni che non sarebbero verosimilmente accessibili con le normali tecniche di analisi. Anticipare la tutela dei titolari al momento della creazione dei file normalizzati significa rendere impossibile o, comunque, ostacolare enormemente l'utilizzo di queste informazioni. La tesi che include tali file tra le riproduzioni oggetto di esclusiva parte dunque dal presupposto che, in questo contesto, l'interesse di proteggere i titolari dei diritti sia l'unico obiettivo meritevole di tutela o, comunque, che l'interesse dei titolari sia prevalente rispetto all'interesse all'uso delle informazioni.

Si è già detto, però, che il diritto d'autore si fonda sul principio per cui la protezione riguarda solo l'espressione creativa dell'autore e non si estende alle idee e alle informazioni contenute nell'opera. Anche l'interesse alla libera circolazione delle informazioni è, quindi, considerato rilevante dalla disciplina del diritto d'autore ed è individuato come un limite alla protezione degli autori. L'interpretazione delle regole sul diritto d'autore deve dunque tenere conto di entrambi gli interessi e tentare di realizzare un bilanciamento in caso di conflitto. In quest'ottica, la lettura che estende il diritto di riproduzione a qualunque oggetto che possa interferire, anche solo lontanamente, con i mercati dell'opera andrebbe, quanto meno, sottoposta ad un vaglio di "proporzionalità". Si tratta, cioè, di chiedersi se non vi siano delle interpretazioni alternative altrettanto efficaci dal punto di vista dei titolari, ma meno restrittive nei confronti della raccolta e dell'utilizzo delle informazioni.

La lettura qui proposta, secondo cui i file fortemente normalizzati non devono essere inclusi tra le "copie" ai sensi del diritto di riproduzione, è sicuramente meno restrittiva dal punto di vista della circolazione delle informazioni. Resta, però, da chiedersi se essa sia anche in grado di

---

nel quale sono state preinstallate estensioni, disponibili su Internet, contenenti collegamenti ipertestuali a siti web liberamente accessibili al pubblico sui quali sono state messe a disposizione del pubblico opere tutelate dal diritto d'autore senza l'autorizzazione dei titolari di tale diritto".

assicurare ai titolari una tutela adeguata nel caso in cui la produzione dei file porti a forme di sfruttamento commerciale della espressione creativa dell'autore. Come accadrebbe, appunto, nel caso in cui un'IA venga programmata per decodificare dataset normalizzati e per trasmetterne i contenuti al pubblico. A questo proposito, si è già visto, però, che la lettura in esame non priva affatto di tutela i titolari contro gli eventuali usi "abusivi" dei file normalizzati. Il titolare resta infatti libero di esercitare i propri diritti di riproduzione e di comunicazione contro i successivi atti di divulgazione del testo. La differenza rispetto alla lettura più restrittiva sta nel fatto che qui il titolare non può agire al momento iniziale della produzione dei file normalizzati. A differenza di quello che accade in altri campi, però, qui lo spostamento in avanti dell'intervento non rende più difficile l'esercizio dei diritti. Il titolare infatti non è costretto ad agire nei confronti di una massa dispersa di utenti ma può rivolgere le sue pretese anche nei confronti dell'operatore di IA che offra agli utenti l'infrastruttura di decodifica e di trasmissione dei testi. La lettura che sottrae i dataset normalizzati all'esclusiva sembra allora realizzare un migliore bilanciamento degli interessi in gioco ed appare quindi preferibile[64].

Questa conclusione appare del resto confermata anche se si guarda al problema dal punto di vista degli obiettivi generali della disciplina sul diritto d'autore. La tesi secondo cui anche i file normalizzati sono copie si fonda sull'idea che l'obiettivo del diritto d'autore sia quello di assicurare la massima tutela possibile agli autori. Questa però non è l'unica possibile lettura della disciplina. Secondo un'altra parte della dottrina, il diritto d'autore e la proprietà intellettuale, in generale, sono discipline funzionalizzate al conseguimento di obiettivi di benessere collettivo[65]. Un

[64] Va detto che qui si potrebbe anche porre la questione se la trasformazione del testo in un codice di stampo matematico non sia piuttosto una traduzione dell'opera. Attività, questa, che è oggetto di un autonomo diritto esclusivo (art. 18, co. 1 l. aut.). La questione impone di chiedersi quale sia il significato del concetto di "traduzione" in questo campo. Il diritto di tradurre l'opera ha tradizionalmente lo scopo di consentire all'autore di beneficiare dell'espansione territoriale dei mercati dell'opera. La legge sembrerebbe riferirsi, quindi, alla traduzione in una lingua che consenta di raggiungere un nuovo pubblico; cioè, in sostanza, ad una lingua "parlata". Il punto sembra confermato dalla lettera dell'art. 18, co. 1, il quale riserva all'autore la traduzione dell'opera "*in altra lingua o dialetto*". Il che dovrebbe escludere l'applicazione del diritto alla normalizzazione del testo. D'altra parte, gli argomenti espressi nel testo con riferimento alla portata del diritto di riproduzione sembrano validi anche per il diritto di traduzione.

[65] LIBERTINI, *Tutela e promozione delle creazioni intellettuali e limiti funzionali della proprietà intellettuale,* in *AIDA,* 2014, 299 e ss., in cui l'A. critica la tesi secondo cui la proprietà intellettuale trova il suo fondamento nelle stesse ragioni che giustificano la tutela delle altre forme di proprietà privata. Queste tesi sono sostenute da una parte della dottrina, tra l'altro, facendo leva sull'art. 17, par. 2 della Carta dei diritti fondamentali dell'UE, dove si legge che "*la proprietà intellettuale è protetta*". V. sul punto OTTOLIA, *The Public Interest and Intellectual Property Models*, Giappichelli, 2010; BERTANI, *Diritto d'autore europeo,* cit., 148, in cui l'A. afferma, comunque, che la proprietà intellettuale può subire compressioni nelle ipotesi in cui l'interesse della collettività alla circolazione dell'opera abbia rango pari o

ampliamento della protezione autoriale deve quindi pur sempre essere sorretto da reali esigenze di sviluppo economico e non deve imporre sacrifici sproporzionati all'efficienza dinamica in altre direzioni. Si è detto che l'estensione dell'esclusiva alle copie "normalizzate" non si giustifica con esigenze di protezione dei mercati dell'opera, visto che l'autore resta comunque legittimato ad esercitare i propri diritti nei confronti della riproduzione "decodificata". Inoltre, questa lettura "espansiva" dell'esclusiva equivarrebbe sostanzialmente a bloccare nel mercato europeo lo sviluppo di una tecnologia altamente innovativa e suscettibile di innumerevoli applicazioni utili, molte delle quali nulla hanno a che vedere con i mercati dei prodotti culturali. Anche da questo punto di vista, la lettura appare dunque sproporzionata e, come tale, incompatibile con gli obiettivi generali del diritto d'autore.

In sintesi, assorbita la funzione della riproduzione normalizzata, tecnica e interna, nel processo volto all'addestramento dell'IA generativa, non pare violare il diritto di riproduzione la mera creazione di copie necessarie al procedimento destinate ad essere conservate in formati e linguaggi utilizzabili esclusivamente per l'analisi degli algoritmi.

Beninteso, questo non significa che il *dataset* sottoposto ad attività di normalizzazione "piena" metta completamente fuori gioco il diritto di riproduzione. Non è detto infatti che la normalizzazione avvenga già in fase di creazione del *dataset.* Può accadere, ad es., che il *dataset* contenga testi identici agli esemplari di base e che il *pre-processing* avvenga nella fase del *training* vero e proprio, cioè dopo la condivisione del *dataset* con i programmatori dell'IA. In questo caso (e in tutti i casi in cui il *dataset* si compone di copie "fruibili" da parte del pubblico) vale quanto detto nei paragrafi precedenti: la creazione del *dataset* comporta creazione di "copie" ed entra quindi in conflitto con il diritto di riproduzione, fatta salva l'eventuale applicazione delle eccezioni in tema di estrazione o di riproduzione temporanea.

---

superiore all'interesse del titolare. L'argomento relativo all'art. 17, par. 2 è stato ridimensionato dalla Corte di Giustizia. Si v. CORTE DI GIUSTIZIA, 29 luglio 2019, C-476/19, *Pelham,* par. 33: "*la Corte ha, in tal senso, già dichiarato che non risulta in alcun modo dall'articolo 17, paragrafo 2, della Carta né dalla giurisprudenza della Corte che il diritto di proprietà intellettuale sancito da tale disposizione sia intangibile e che la sua tutela debba essere garantita in modo assoluto*".

Nel senso che il fondamento "costituzionale" del diritto d'autore imponga un bilanciamento tra gli interessi degli autori e l'interesse generale all'innovazione, v., tra gli altri, SCHIUMA, *Diritto d'autore e normativa europea,* cit.; RICOLFI, *Il diritto d'autore,* cit., 461, in cui l'A. afferma che, nella disciplina sul diritto d'autore, l'incentivo proprietario degli autori deve essere coordinato con l'interesse generale alla diffusione dei contenuti protetti. V. anche con specifico riferimento all'IA l'analisi di HILTY, HOFFMANN, SCHEUERER, *Intellectual property justifications for artificial intelligence,* in *Artificial Intelligence and Intellectual Property,* edito da Lee, Hilty, Liu, 2021, 50 e ss.

Se si adotta la lettura qui proposta, però, la creazione di un *dataset* che sia fin dall'inizio sottoposto a processi di forte normalizzazione non comporta la produzione di "copie". L'operazione cade dunque del tutto al di fuori del diritto di riproduzione ed è da considerare libera. In questo caso, il problema di applicare le eccezioni non si pone. Un *dataset* "normalizzato" potrebbe essere costruito anche andando al di là delle strette maglie dell'art. 5, par. 1: potrebbe, cioè, essere composto da *file* permanenti, suscettibili di trasferimento ad altri programmatori ed utilizzabili in una pluralità di processi di analisi computazionale. In questi casi, dovrebbe poi considerarsi libera anche la successiva circolazione del *dataset*. L'atto di trasferimento non ha infatti ad oggetto "copie" dell'opera e si colloca, pertanto, al di fuori del diritto di distribuzione. D'altra parte, l'operazione non pare neppure rientrare tra le forme di comunicazione al pubblico, visto che, come già detto, i *file* in questione sono incapaci di trasmettere direttamente l'opera agli utenti. Nel caso del *dataset* pienamente normalizzato, l'esclusiva potrebbe venire in rilievo, tutt'al più, per le copie iniziali create al momento del *download* che consistono in repliche pedisseque dei contenuti raccolti dal sito. Come già visto, però, se sottoposte a meccanismi di cancellazione automatica, queste copie possono rientrare nell'ambito di applicazione dell'art. 5, par. 1.

6. *Conclusioni.*

Bisogna, comunque, riconoscere che questa proposta interpretativa si pone in controtendenza rispetto agli orientamenti prevalenti in Europa e, soprattutto, rispetto alla tendenza "espansiva" generalmente adottata dalla Corte di Giustizia nell'interpretare la portata dell'esclusiva. Su un piano realistico, non si può dunque prescindere dall'affrontare il problema anche in una prospettiva *de iure condendo.*

D'altra parte, quest'analisi si rende necessaria anche per un'altra ragione. Analizzando il problema *de iure condito,* si è detto che consentire l'uso dei testi per l'addestramento dell'IA appare una soluzione preferibile rispetto all'opzione di ostacolare lo sviluppo di questa linea di innovazione. Non è detto però che lasciare del tutto libera l'IA di servirsi dei contenuti sia, in assoluto, la maniera più equilibrata di regolare il conflitto tra industria editoriale e innovazione tecnologica.

L'IA generativa può trovare utile applicazione in numerosi campi. Può essere adoperata, ad es., per scopi di carattere scientifico, per la soluzione di problemi tecnici o per automatizzare certi aspetti di un processo produttivo. Ancora, può svolgere compiti di ricerca statistica o assistere l'utente nella comprensione, nell'analisi o nella stesura di documenti. In tutti questi casi, il servizio svolto dall'IA non si pone in diretta competizione con le attività delle imprese editoriali.

Come già visto, però, ci possono anche essere sistemi di IA offerti al grande pubblico per rispondere alle domande più varie. In questo caso, l'utente potrebbe rivolgersi al *chatbot* per avere informazioni sulle notizie del giorno, su eventi di attualità, su temi storici, ecc. Allo stato questa funzionalità dell'IA non sembra essere arrivata al punto da produrre un concreto impatto negativo sull'industria editoriale. Con l'affinamento dei servizi, però, l'IA potrebbe effettivamente cominciare a sviluppare contenuti del tutto sostituibili ai prodotti editoriali. Il rischio si pone soprattutto per le forme più recenti di IA, nelle quali i meccanismi generativi sono combinati con la possibilità di ricercare su Internet in tempo reale le informazioni richieste dagli utenti. In circostanze del genere, l'IA sostanzialmente compete sul mercato dei contenuti editoriali senza sostenere gli investimenti necessari per la raccolta delle notizie e dei contenuti, per la selezione delle fonti e per la verifica delle informazioni. Qui si rischia allora effettivamente di disincentivare gli investimenti editoriali tradizionali, che sono però pur sempre necessari.

Il problema in esame non si riferisce tanto all'ipotesi in cui l'IA riproduca articoli o testi pubblicati sui siti informativi. Questa situazione – lo si è già detto – configura una violazione dei diritti d'autore ed è dunque già soggetta al controllo delle imprese editoriali. La questione si pone invece soprattutto nell'ipotesi in cui l'IA si serva delle informazioni raccolte per produrre testi completamente nuovi.

Per queste applicazioni dell'IA generativa una qualche protezione degli editori, specialmente nel campo giornalistico, appare necessaria. Qui si tratta, però, di cercare un compromesso tra protezione degli incentivi alla produzione culturale e protezione degli incentivi all'investimento nello sviluppo di innovazioni socialmente utili. Per le ragioni già espresse, il meccanismo dell'esclusiva non sembra in grado di realizzare questo bilanciamento. Una soluzione più convincente è probabilmente quella di attribuire, da un lato, agli operatori di IA il diritto di servirsi dei testi per l'addestramento e, dall'altro, alle imprese editoriali il diritto di ricevere un compenso ragionevole da parte delle IA che si rivolgono al pubblico degli utenti finali. Per un approfondimento di questi aspetti si deve però, a questo punto, rinviare ad un futuro sviluppo del lavoro.