

**TRIBUNALE REGIONALE
AMBURGO**
27 SETTEMBRE 2024
CAUSA 310 O 227/23
PRESIDENTE: HARTMANN
PARTI: KNESCKE LAION

Diritto d'autore • Estrazione di testo e dati • Direttiva (UE) 2019/790 • Ambito di applicazione materiale • Creazione di dataset di allenamento • Intelligenza Artificiale.

L'argomentazione che sostiene che il "AI-web scraping" riguarderebbe il contenuto espressivo delle opere utilizzate a fini di addestramento e che, "in ultima analisi", ciò condurrebbe alla creazione di prodotti concorrenti sostanzialmente identici o simili per contenuto non opera una distinzione sufficientemente rigorosa tra: da un lato, la creazione di un dataset utilizzabile — anche — per l'addestramento di un modello di Intelligenza Artificiale; dall'altro, il successivo addestramento della rete neurale artificiale mediante tale dataset e, terzo, la successiva utilizzazione dell'Intelli-

genza Artificiale addestrata per la generazione di nuovi contenuti. Al momento della compilazione del dataset di allenamento non è prevedibile quanto il secondo step avrà successo né quale contenuto in concreto possa essere generato mediante l'allenata Intelligenza Artificiale nel terzo step. Le concrete possibilità applicative di una tecnologia in rapido sviluppo come l'Intelligenza Artificiale non sono perciò prevedibili in modo definitivo al momento della creazione del dataset di allenamento e, di conseguenza, non possono essere stabilite con certezza giuridica. A causa di tale incertezza giuridica, la mera intenzione generica di ottenere contenuti generati con l'IA in futuro quando si costruisce un dataset di allenamento non rappresenta un criterio adeguato a stabilire normativamente l'ammissibilità della costruzione del dataset di allenamento in quanto tale.

In fatto. La convenuta è un'associazione formatasi con atto costitutivo in data 7 luglio 2021 (verbale in allegato B7, statuto in allegato B1). L'oggetto concreto dell'attività della convenuta è controverso tra le parti.

La convenuta mette gratuitamente a disposizione del pubblico, sotto la denominazione "...", un cosiddetto *dataset* di coppie immagine-testo. Si tratta di una sorta di documento tabellare che contiene collegamenti ipertestuali a immagini o file immagine accessibili pubblicamente su Internet, nonché ulteriori informazioni relative alle rispettive immagini, tra cui una descrizione testuale (detta anche "*alternative text*") che ne esplicita il contenuto. Il *dataset* comprende 5.85 miliardi di coppie immagine-testo. Può essere utilizzato per l'addestramento di cosiddette Intelligenze Artificiali Generative.

Il *dataset* è stato creato nella seconda metà del 2021, dopo la fondazione della convenuta. A tal fine, la convenuta ha attinto da un *dataset* preesistente proveniente da "..." negli Stati Uniti (www...org), che conteneva per ciascuna immagine l'URL e una descrizione testuale del contenuto dell'immagine stessa, rappresentando [così] un tipo di campione casuale dei contenuti reperibili su Internet. La convenuta ha quindi estratto da questo *dataset* gli URL delle immagini e ha scaricato i file immagine dai rispettivi siti di origine. Infine, mediante un *software*, le immagini sono state analizzate dalla convenuta per verificare se le descrizioni presenti nel *dataset* originale corrispondessero effettivamente al contenuto nelle immagini. Le immagini la cui descrizione non avesse

combaciato con il contenuto venivano escluse. Dalle fotografie rimanenti sono stati estratti i metadati, in particolare l'URL del sito di origine e la descrizione, i quali sono stati inseriti nel nuovo *dataset* denominato "...". Non è pacifico tra le parti se, successivamente, i file immagine scaricati siano stati cancellati — almeno per quanto riguarda la fotografia oggetto della controversia.

Nell'ambito di tale processo tecnico è stata acquisita, scaricata, analizzata e inclusa nel *dataset* "..." anche la fotografia oggetto della presente causa. In concreto, è stato scaricato un file immagine pubblicato sul sito dell'agenzia fotografica "..." (<https://www...com>), contrassegnato da una filigrana (*watermark*) della suddetta agenzia.

Sul sito dell'agenzia "..." era presente almeno dal 13 gennaio 2021, alla pagina <https://www...com/de/usage.html>, il seguente testo:

“RESTRICTIONS YOU MAY NOT: (...)

use automated programs, applets, bots or the like to access the ...com website or any content thereon for any purpose, including, by way of example only, downloading content, indexing, scraping or caching any content on the website.”

L'attore lamenta una violazione dei diritti d'autore sulla fotografia in questione nella forma di una riproduzione non autorizzata compiuta dalla convenuta durante il processo di analisi.

L'attore afferma di essere l'autore della fotografia. L'agenzia "..." aveva il diritto di offrire e anche pubblicare sul suo sito web l'immagine in questione, e di offrire contratti di licenza per il relativo utilizzo; "..." era in tal senso titolare di diritti d'uso non esclusivi e sub licenziabili.

La riproduzione — non contestata — avvenuta come parte del processo di analisi violerebbe i diritti dell'autore ai sensi del §16 *UrhG*, non essendo in particolare coperta dalle limitazioni previste ai §44a, 44b e 60d *UrhG*:

[Secondo l'attore:]

La limitazione posta dal §44a *UrhG* non sarebbe applicabile, poiché il *download una tantum* della fotografia non costituirebbe, in particolare, un atto di riproduzione temporanea ai sensi di tale norma.

La riproduzione non sarebbe neppure inclusa nell'ambito applicativo del §44b *UrhG*. L'aggregazione di dati per l'allenamento di modelli di Intelligenza Artificiale non consisterebbe in una estrazione di testo e dati ai sensi del §44b *UrhG*. Né il legislatore europeo né quello tedesco avrebbero avuto “in mente” tale utilizzo nell'introduzione dell'art. 4 DSM o del §44b *UrhG*. Nell'ambito dell'estrazione di testo e dati ex §44b *UrhG* dovrebbero “essere rese accessibili [solo] le informazioni contenute nei dati”, “ma il contenuto della creazione intellettuale non dovrebbe essere utilizzato”. Il caso del cd. “*AI web scraping*” in esame, invece, riguarderebbe precisamente il contenuto intellettuale delle opere utilizzate per finalità di allenamento “e in ultima istanza la creazione di prodotti competitivi identici o simili”. Inoltre, il *dataset* sarebbe, secondo lo stesso *disclaimer* della convenuta (presente a pag. 47), “non curato”. Infine, la raccolta e la conservazione per la creazione di un archivio

parallelo sarebbero escluse dall'ambito di applicazione del §44b *UrhG* secondo l'intenzione espressamente dichiarata del legislatore.

Inoltre, "l'assimilazione massiccia di opere protette dal diritto d'autore per finalità di allenamento nel contesto dell'IA generativa" pregiudicherebbe il normale sfruttamento delle opere, creando le condizioni per sostituire gli autori in molti casi o, quanto meno, complicando lo sfruttamento dell'opera a causa di un'offerta concorrenziale gratuita. Ai sensi dell'art. 7, par. 2 Direttiva DSM in combinato disposto con l'art. 5, par. 5 Direttiva *InfoSoc*, ciò escluderebbe l'applicazione della norma limitativa.

In ogni caso, la riproduzione sarebbe inammissibile ai sensi del §44b, par. 3 *UrhG* per via della riserva di diritti espressa sul sito web. La dichiarazione dell'agenzia fotografica sarebbe attribuibile all'attore, in quanto distribuisce per suo conto l'immagine in questione. Contrariamente all'opinione della convenuta, la riserva sarebbe altresì leggibile in via automatizzata ai sensi del §44b par. 3, secondo periodo *UrhG*. I requisiti a questo riguardo non sarebbero superiori a quelli della leggibilità automatizzata da parte dell'uomo. Sebbene la riserva sia scritta in stampatello, il testo sarebbe riconoscibile come riserva anche da parte di un programma per computer. Come il servizio "..." così gli strumenti specifici come *WebOpt-Out* sarebbero stati in grado di riconoscere la riserva espressa sul sito.

La convenuta non potrebbe neanche invocare la limitazione prevista al §60d *UrhG*. L'attore contesta che la convenuta soddisfi di fatto i requisiti del § 60d *UrhG*, ed in particolare:

— che fosse "iscritta nel registro delle associazioni" al momento dell'atto di riproduzione in questione;

— che il documento B1 presentato dalla convenuta rappresenti lo statuto dell'associazione della convenuta o la abbia rappresentata al momento dell'atto di riproduzione in questione;

— che i membri dell'associazione così come gli amministratori siano o siano stati volontari al momento dell'atto di riproduzione in questione;

— che la convenuta svolga esclusivamente attività di ricerca o fosse impegnata in attività di ricerca scientifica al momento dell'atto di riproduzione in questione, o che la convenuta svolga attività di ricerca scientifica, persegua scopi non lucrativi e reinvesta tutti i profitti nella ricerca scientifica od operi nell'ambito di un incarico riconosciuto di interesse pubblico dallo Stato; inoltre, secondo lo statuto presentato dalla convenuta, lo scopo dell'associazione sarebbe solo la "promozione della ricerca" e non la "ricerca" in quanto tale, e non sarebbe dunque chiaro quale dovrebbe essere la "ricerca" nella raccolta costruita dalla convenuta, che viene (indiscutibilmente) messa a disposizione di altre aziende;

— che la convenuta crei e testi propri modelli di Intelligenza Artificiale sulla base dei dati di addestramento per effettuare ulteriori ricerche sulle possibilità della tecnologia IA;

— che, rendendo il set di dati di addestramento disponibile al pubblico, offra ad altri ricercatori ed interessati la possibilità di addestrare i propri modelli di Intelligenza Artificiale; secondo le dichiarazioni della convenuta, il *dataset* in questione è stato utilizzato anche per addestrare i servizi "...", "...", "..." e "..." del fornitore "...", sebbene siano forniti da una società (puramente) commerciale; sebbene la convenuta neghi l'allena-

mento dei primi due servizi, l'utilizzo del *dataset* sarebbe stato comunque possibile per il fornitore.

La convenuta non potrebbe nemmeno beneficiare della limitazione posta dal §60d *UrhG* alla luce del §60d, par. 2, terzo periodo. La convenuta collaborerebbe infatti in modo intenso con fornitori commerciali di intelligenza artificiale:

— esisterebbe chiaramente una collaborazione con la società privata "...", la quale eserciterebbe un'influenza diretta attraverso il finanziamento del *dataset* in questione e la nomina di propri dipendenti a posizioni rilevanti presso la convenuta. Secondo un'intervista rilasciata dal suo fondatore e amministratore delegato, "...", avrebbe finanziato il *dataset* di "...";

— inoltre, alcuni membri del "team" della convenuta lavorerebbero "in diversi ambiti" nello stesso settore per grandi aziende tecnologiche, [in particolare] come dipendenti dell'azienda "...";

— in aggiunta, in una chat del 28 settembre 2021 sulla piattaforma "...", il co-fondatore della convenuta, il sig. "...", avrebbe sollecitato l'associazione a completare il "...", poiché si era ricevuto un finanziamento di 5.000 \$ da parte di "..." o dalla sua società; i dati avrebbero dovuto essere messi a disposizione di quest'ultimo anche nel caso in cui il *dataset* non fosse stato reso ancora accessibile al pubblico. Il menzionato "..." sarebbe un dipendente del fornitore commerciale di Intelligenza Artificiale "...".

Infine, la convenuta non potrebbe nemmeno avvalersi di una cosiddetta "semplice autorizzazione". Egli — l'attore — non avrebbe reso liberamente accessibile la fotografia in questione, bensì l'avrebbe messa a disposizione dall'agenzia "..." per la concessione di licenze a titolo oneroso.

L'attore aveva inizialmente richiesto, oltre all'inibizione dell'atto di riproduzione, anche informazioni sulla portata dell'utilizzo della fotografia. Tale richiesta di informazioni è stata dichiarata dalle parti consensualmente risolta nell'udienza dell'11.07.2024.

L'attore chiede ora:

che la convenuta sia condannata ad astenersi dalla riproduzione e/o dal consentire la riproduzione della fotografia in questione per la creazione di *dataset* di allenamento per l'Intelligenza Artificiale, come avvenuto nel contesto della creazione del *dataset* "...", pena, in caso di inosservanza, al pagamento di una sanzione amministrativa fino a euro 250.000 per ogni singolo caso di violazione o, in alternativa, alla reclusione fino a 6 mesi.

La convenuta chiede:

di respingere la domanda.

La parte convenuta nega che l'attore abbia creato personalmente la fotografia in questione o che sia altrimenti autorizzato a far valere le violazioni dei diritti in proprio nome in relazione all'immagine, nonché che l'attore fosse autorizzato a far valere le violazioni dei diritti riguardanti l'immagine in questione in proprio nome al momento dell'acquisizione dell'immagine da parte della convenuta.

Soprattutto, il *download* (*una tantum*) della fotografia in questione effettuato nell'ambito della creazione del *dataset* "..." costituirebbe, sì, una riproduzione rilevante ai fini del diritto d'autore, ma ricadrebbe nelle limitazioni previste dai §§ 44a, 44b e 60d *UrhG*, nonché autorizzata dall'attore:

La riproduzione effettuata rientrerebbe, da un lato, nell'ambito di applicazione del §44a *UrhG*. Non vi sarebbe stata alcuna memorizzazione permanente delle immagini da parte della convenuta; al contrario, le immagini sarebbero state utilizzate solo temporaneamente per l'analisi, per poi essere immediatamente cancellate in modo automatico e irreversibile. La riproduzione in sé non avrebbe avuto alcuna rilevanza economica autonoma.

Si applicherebbe, inoltre, la limitazione contenuta nel §44b *UrhG*. L'analisi del file immagine e l'estrazione di metadati per l'addestramento dell'Intelligenza Artificiale rappresenterebbero, secondo l'intenzione del legislatore, uno degli usi principali di estrazione di testo e dati. Non si configurerebbe nemmeno la creazione di un archivio digitale parallelo, poiché le immagini scaricate non sarebbero memorizzate permanentemente, bensì sarebbero registrati solo dei collegamenti ipertestuali. Non si applicherebbe l'eccezione prevista dal §44b, par. 3, *UrhG*:

— Secondo quanto affermato dall'attore, ad esprimere la riserva non sarebbe stato lui stesso in quanto titolare dei diritti, ma l'operatore del sito web *www...com* in qualità di terzo; l'attore stesso avrebbe chiarito nella sua e-mail del 13.02.2023 (allegato B5) di non avere né la qualifica né i mezzi economici per esprimere una riserva di utilizzo;

— inoltre, la riserva non sarebbe stata espressamente dichiarata, poiché la clausola presente sul sito web *www...com* era formulata in modo generale ed elencava diverse azioni non consentite. Non vi era alcun riferimento esplicito né all'estrazione di testo e dati né alle riproduzioni;

— mancherebbe, inoltre, il requisito della leggibilità automatizzata. Generalmente, una clausola redatta in linguaggio naturale non sarebbe leggibile in via automatizzata ai sensi del §44b, par. 3, *UrhG*. La condizione della leggibilità in via automatizzata consisterebbe, in tal senso, nell'elaborazione automatica della clausola da parte di un *software*, il che implicherebbe che l'informazione fosse codificata. In ogni caso, sarebbe necessario che nel testo fossero contenute almeno specifiche parole chiave come, ad esempio, "*Data Mining*";

— la dichiarazione non sarebbe intesa come riserva ai sensi §44b, par. 3, *UrhG*. Il fatto che, secondo quanto riferito dall'attore, la clausola fosse già presente sul sito web in quella forma sin dal 13.01.2021 dimostra che la clausola non poteva essere stata redatta "con riferimento alla normativa del §44b, par. 3 *UrhG*", dato che la disposizione non era ancora in vigore in quel momento. In aggiunta, "non è nemmeno credibile", che un fornitore statunitense si rifaccia ad una riserva di utilizzo derivante da una normativa tedesca.

In ogni caso, la convenuta potrebbe avvalersi della limitazione posta al §60d *UrhG*.

— La convenuta sarebbe un'associazione no-profit costituita da ricercatori e dedicata alla ricerca, come indicato dallo statuto (allegato B1), e, soprattutto, si sarebbe posta l'obiettivo di sviluppare ulteriormente gli

algoritmi *self-learning* nell'ambito dell'Intelligenza Artificiale e di metterli a disposizione del vasto pubblico. A tal fine, fornirebbe gratuitamente i *dataset* e i modelli, e creerebbe e testerebbe i propri modelli di Intelligenza Artificiale sulla base dei dati di addestramento;

— la sua attività consisterebbe anche in “ricerca”. Anche solo rendendo pubbliche e trasparenti su Internet le modalità con cui i *dataset* di addestramento vengono creati, l'associazione contribuirebbe ad ottenere conoscenze sull'allenamento dell'Intelligenza Artificiale. In questo modo, altri ricercatori potrebbero comprendere le fasi di creazione dei *dataset* e basarsi su di essi. Inoltre, la convenuta ha pubblicato per la prima volta il 17.09.2022 un articolo scientifico sul *dataset* “...” in questione con il nome “...” (allegato B8). Fino al 5.04.2024, l'articolo sopra menzionato era stato citato complessivamente 1403 volte in altri contributi scientifici e aveva ricevuto ulteriori riconoscimenti;

— in aggiunta, la convenuta allenerrebbe anche propri modelli di Intelligenza Artificiale basati sui *dataset* da lei creati per ottenere conoscenze su come l'Intelligenza Artificiale possa essere migliorata attraverso un allenamento adeguato;

— anche le persone fisiche iscritte alla convenuta, ovvero gli amministratori e gli altri membri dell'associazione, sarebbero “ricercatori”. Il fatto che l'attore indichi che alcuni membri del gruppo lavorino nello stesso settore presso grandi aziende tecnologiche non sarebbe rilevante ai fini dell'accertamento sul perseguimento di finalità lucrative da parte della convenuta. Inoltre, le persone interessate lavorerebbero volontariamente per la convenuta e dovrebbero guadagnarsi il proprio sostentamento altrove;

— il fatto che i *dataset* resi pubblicamente disponibili dalla convenuta sarebbero stati utilizzati anche da fornitori commerciali sarebbe irrilevante per l'applicazione del §60d *UrhG*. Inoltre, i servizi “...” e “...” non sarebbero stati in concreto addestrati con il *dataset* della convenuta;

— anche l'eccezione regolata dal §60d, par. 2, terzo periodo, non si applicherebbe al presente caso. La società “...” avrebbe sì fornito risorse di calcolo alla convenuta nel periodo della sua fondazione, ma lo stesso sarebbe capitato anche con “...”. La convenuta non avrebbe ricevuto un sostegno finanziario in forma di denaro da “...”. Non ci sarebbero state neanche ulteriori collaborazioni con tale azienda e, in ogni caso, “...” non riceverebbe alcun accesso privilegiato ai risultati della ricerca. La società “...” non avrebbe un'influenza determinante. Né la società stessa né uno dei suoi rappresentanti legali sarebbero membri della convenuta.

Infine, ci sarebbe una cd. semplice autorizzazione a favore degli utilizzi della convenuta da parte dell'attore. La riproduzione in questione dovrebbe essere classificata quale atto di utilizzo normale.

Inoltre, si dovrebbe notare che l'attore stesso guadagna con immagini generate dall'Intelligenza Artificiale, mettendo con ciò in discussione il suo “bisogno di tutela giuridica”.

Per quanto attiene agli ulteriori dettagli delle argomentazioni delle parti, si fa riferimento agli scritti pervenuti agli atti, inclusi gli allegati, per quanto oggetto di udienza, nonché al verbale dell'udienza del 11.07.2024. [...]

Motivi

I. La domanda, sebbene ammissibile, non convince nel merito. La convenuta, attraverso la riproduzione della fotografia in questione, ha interferito con i diritti di sfruttamento economico dell'autore. Tale introduzione è tuttavia inclusa nell'ambito di applicazione della limitazione del §60d *UrhG*. In questo caso, non è necessario valutare in via definitiva se la convenuta possa avvalersi anche della disposizione di cui al §44b *UrhG*.

La fotografia in questione è comunque protetta quale opera fotografica ai sensi del §72, par. 1 *UrhG*. Dopo l'ispezione dei dati grezzi presenti sul laptop dell'autore, il Tribunale non ha alcun dubbio sulla qualità dello stesso quale autore dell'immagine fotografica ai sensi del §72, par. 2 *UrhG*. L'attore è anche legittimato ampiamente a far valere le proprie pretese ai sensi del §97 *UrhG*, compresa la pretesa di cessazione degli abusi ai sensi del par. 1 della norma; la convenuta non ha dimostrato che l'attore abbia concesso all'agenzia fotografica "..." diritti di sfruttamento più ampi di semplici diritti di utilizzo (sub licenziabili). L'agenzia fotografica "..." ha aggiunto una filigrana alla foto; si trattava di un'alterazione non libera ai sensi del §23, par. 1, primo periodo *UrhG*, sicché per il relativo utilizzo, in linea di principio, sarebbe stato necessario anche il consenso dell'attore in quanto autore. Nell'ambito del *download* effettuato, la convenuta ha riprodotto tale versione [della foto] ai sensi del §16 par. 1 *UrhG*, senza aver ottenuto il consenso dell'attore.

Tuttavia, la convenuta era legittimata a farlo in base ad un'autorizzazione normativa. La riproduzione non sarebbe però tutelata dalla limitazione prevista dal §44a *UrhG* (nel seguente punto 1.), e pare dubbio che la convenuta possa avvalersi dell'ambito applicativo del §44b *UrhG* (nel seguente punto 2.). Tuttavia, quest'ultimo punto non richiede un giudizio definitivo nel presente caso, in quanto l'atto di riproduzione comunque ricade nella limitazione [ai diritti d'autore] prevista dal §60d *UrhG* (di seguito 3.).

1. La riproduzione effettuata non ricade nella limitazione prevista dal §44a *UrhG*.

Secondo tale previsione, sono ammessi gli atti di riproduzione temporanea che sono transitori o accessori e parte integrante di un procedimento tecnologico, eseguiti all'unico scopo di consentire la trasmissione in rete tra terzi attraverso un intermediario o l'uso legittimo di un'opera o di altri materiali protetti e privi di rilievo economico proprio.

L'effettuata riproduzione in esame non è stata né transitoria né accessoria.

■) Una riproduzione è considerata transitoria ai sensi del §44a *UrhG* quando la sua durata è limitata a quanto necessario per il corretto funzionamento tecnico, ove tale procedimento deve essere automatizzato al punto che la riproduzione si cancelli automaticamente, cioè senza l'intervento di una persona fisica, non appena la sua funzione di consentire l'esecuzione sia stata compiuta (Corte giust. Sent. 16.07.2009, causa C-5/08 - *Infopaq/Danske Dagblades Forening*, punto 64 (juris) sull'art. 5, par. 1, Direttiva DSM).

L'argomentazione della convenuta secondo cui i file sono stati cancellati "automaticamente" durante il processo di analisi da lei condotto non è sufficiente a giustificare la transitorietà della riproduzione nella menzionata accezione. Oltre al fatto che la convenuta non ha fornito dettagli sulla durata concreta della memorizzazione, la cancellazione non è avvenuta in modo "indipendente dall'utente", bensì a fronte di una consapevole programmazione del processo di analisi da parte della convenuta.

b) Una riproduzione è considerata accessoria ai sensi del §44a *UrhG* quando non è autonoma rispetto al procedimento tecnico di cui fa parte e non serve ad uno scopo indipendente (Corte giust., Sent. 05.06.2014, causa C-360/13, punto 43 (juris)).

Nel seguente caso è stato effettuato un **download** mirato dei file immagine, per analizzarli mediante uno specifico *software*. Pertanto, il **download** non è un mero processo accessorio all'analisi effettuata, bensì un processo di acquisizione consapevole e attivamente controllato antecedente all'analisi.

2. Nel seguente caso, pare incerto se la convenuta possa avvalersi della limitazione di cui al §44b *UrhG*. Il **download** effettuato dalla convenuta rientra in linea di principio nell'ambito di applicazione del §44b, par. 2 *UrhG*, in particolare perché effettuato per finalità di estrazione di testo e dati ai sensi del §44b par. 1 *UrhG* (nel seguente punto a)). Tuttavia, vi sono alcune indicazioni che — senza che si necessiti di una decisione definitiva sul punto — suggeriscono che l'atto di riproduzione non potrebbe essere consentito ai sensi del §44b, par. 2 *UrhG* a causa di una riserva espressamente dichiarata ai sensi del §44b, par. 3 *UrhG* (nel seguente punto b)).

a) L'atto di riproduzione in questione rientra fundamentalmente nella limitazione posta dal §44b par. 2 *UrhG*.

(1) Il **download** in questione è stato effettuato per finalità di estrazione di testo e dati ai sensi del §44b par. 1 *UrhG*. Sulla base del dettato normativo, l'estrazione di testo e dati è l'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. Tale [definizione] è certamente applicabile al seguente atto di riproduzione (nel seguente (a)); una riduzione teleologica dell'ambito di applicazione della disposizione non può inoltre essere presa in considerazione (sotto (b)).

Nel presente caso non è necessario decidere ulteriormente sulla questione, ampiamente discussa in dottrina, se l'allenamento dell'Intelligenza Artificiale nel suo complesso rientri o meno nella limitazione del §44b *UrhG* (per una panoramica delle opinioni cfr. *BeckOK UrhR/Bomhard*, 42a ed. 15.2.2024, *UrhG* § 44b linee 11a-11b m.w.N.; cfr. anche lo studio dettagliato commissionato dall'Iniziativa per il Diritto d'Autore "*Urheberrecht & Training generativer KI-technologische und rechtliche Grundlagen*" presentato come allegato K11).

(a) La convenuta ha effettuato l'atto di riproduzione al fine di ottenere informazioni sulle "correlazioni" nel senso letterale del §44b, par. 1 *UrhG*. Essa ha scaricato l'immagine in questione dal sito di origine per confrontare il contenuto con la descrizione dell'immagine memoriz-

zata tramite un *software* già disponibile — apparentemente l'applicazione “...” di “...”. Questa analisi di confronto tra dati dell'immagine e una preesistente descrizione dell'immagine stessa costituisce senza dubbio un'analisi automatizzata finalizzata all'ottenimento di informazioni su “correlazioni” (ovvero riguardo alla questione della corrispondenza o meno tra immagini e descrizioni delle immagini). Non è stato contestato da parte dell'attore che la convenuta abbia analizzato le immagini contenute nel *dataset* “...” in questo modo.

L'applicabilità del §44b, par. 1 *UrhG* non è esclusa, contrariamente all'opinione dell'attore (replica pp. 13 ss.), dal fatto che, sulla base di un *disclaimer* a riguardo, la convenuta non abbia “curato” il *dataset* “...” da lei creato. Il *disclaimer* riportato dall'attore si riferisce esclusivamente ad un avviso volto a dichiarare che il *dataset* non è stato esaminato allo scopo di cercare “contenuti disturbanti” o simili. Tuttavia, un tale filtraggio aggiuntivo nel *dataset* in fase di creazione non integra un presupposto per l'applicazione del §44b, par. 1, *UrhG* e non osta alla constatazione che le immagini scaricate siano state analizzate, come spiegato, per la loro correlazione tra contenuto e relativa descrizione.

(b) L'atto di riproduzione in questione non deve essere escluso dall'ambito di applicazione del §44b *UrhG* mediante una riduzione teleologica.

È dubbio il fatto che l'esclusione della riproduzione dei dati per finalità di addestramento dell'Intelligenza Artificiale mediante una riduzione teleologica, occasionalmente sostenuta in dottrina sulla base del fatto che il §44b *UrhG* sia applicabile solo all'estrazione di “informazioni nascoste nei dati” ma non all'uso del “contenuto della forma espressiva” (*Schack*, NJW 2024, 113; in tal senso anche *Dormis/Strober*, *Urheberrecht und Training generativer KI-Modelle*, Allegato K11, pp. 67 ss., con una distinzione tra contenuto semantico e sintattico), sia convincente; ciò in quanto nel caso di opere digitalizzate non è sufficientemente chiaro quale sia la differenza tra “informazioni nascoste nei dati” e “contenuto della forma espressiva”.

Inoltre, l'argomentazione che sostiene che il “*AI-web scraping*” riguarderebbe il contenuto espressivo delle opere utilizzate a fini di addestramento e che, “in ultima analisi”, ciò condurrebbe alla creazione di prodotti concorrenti sostanzialmente identici o simili per contenuto (*Schack*, op. cit.), ad avviso del Collegio non opera una distinzione sufficientemente rigorosa tra:

- da un lato, la creazione (qui unico oggetto del contendere) di un *dataset* utilizzabile — anche — per l'addestramento di un modello di Intelligenza Artificiale;

- dall'altro, il successivo addestramento della rete neurale artificiale mediante tale *dataset* e

- terzo, la successiva utilizzazione dell'Intelligenza Artificiale addestrata per la generazione di nuovi contenuti.

Quest'ultima funzionalità potrebbe già essere perseguita mediante la creazione di un *dataset* di allenamento. Tuttavia, al momento della compilazione del *dataset* di allenamento non è prevedibile quanto il secondo *step* (l'allenamento) avrà successo né quale contenuto in concreto possa essere generato mediante l'allenata Intelligenza Artificiale nel terzo

step (nell'impiego dell'Intelligenza Artificiale). Le concrete possibilità applicative di una tecnologia in rapido sviluppo come l'Intelligenza Artificiale non sono perciò prevedibili in modo definitivo al momento della creazione del *dataset* di allenamento e, di conseguenza, non possono essere stabilite con certezza giuridica. A causa di tale incertezza giuridica, la mera intenzione generica di ottenere contenuti generati con l'IA in futuro quando si costruisce un *dataset* di allenamento non rappresenta un criterio adeguato a stabilire normativamente l'ammissibilità della costruzione del *dataset* di allenamento in quanto tale.

Infine, non è chiaramente sufficiente sostenere a favore di una riduzione teleologica della limitazione posta dal §44b *UrhG*, che il legislatore europeo "semplicemente non aveva ancora preso in considerazione il problema dell'IA" quando la disposizione della Direttiva sottostante (art. 4 DSM) è stata adottata nel 2019 (*Schack, op. ult. cit.*; analogamente per l'addestramento dei modelli di IA *Dormis/Stober, op. ult. cit.*, pagg. 71 ss., 87 ss.). In particolare, occorre considerare che lo sviluppo tecnico nel campo dell'Intelligenza Artificiale dal 2019 non riguarda tanto il tipo e la portata del (contestato) *data mining* per il reperimento dei dati di allenamento, quanto piuttosto le prestazioni delle reti neurali artificiali allenate con i dati (di conseguenza, *Dormis/Stober, cit.*, p. 95, ipotizzano anche che la mera creazione di serie di dati di allenamento "prima dell'allenamento vero e proprio" possa rientrare nella soglia del TDM). Si deve inoltre notare che il *database* richiamato dalla convenuta appartenente alla Common Crawl Foundation è costruito a partire dal 2008 (!), vedi <https://commoncrawl.org/overview>.

Sul punto, l'attuale legislatore europeo ha indubbiamente espresso nel Regolamento sull'Intelligenza Artificiale (Regolamento (UE) 2024/1689 del 13.06.2024, G.U. 12.07.2024, p. 1) che anche la creazione di *dataset* destinati all'allenamento di reti neurali artificiali rientri nella limitazione di cui all'art. 4 DSM. Secondo l'art. 53, par. 1, lett. c) del Regolamento sull'IA, i fornitori di modelli di Intelligenza Artificiale sono tenuti a prevedere una strategia volta soprattutto a identificare e rispettare le riserve dei diritti espresse ai sensi dell'art. 4, par. 3 DSM.

Il fatto che la creazione di *dataset* destinati all'allenamento di reti neurali artificiali rientri nella limitazione dell'art. 4 DSM corrisponde inoltre all'orientamento del legislatore tedesco nell'ambito dell'attuazione della norma nel 2021 (*Begr. RegE BT-Drs. 19/27426*, pag. 60).

(c) Anche il c.d. *three-step-test* ancorato nell'art. 5, par. 5 Dir. *InfoSoc* (in combinato disposto con l'art. 7, par. 2, primo periodo DSM) non giustifica infine una valutazione differente. Secondo tale test, le eccezioni normativamente previste possono essere applicate solo in determinati casi eccezionali che non siano in contrasto con lo sfruttamento normale dell'opera o degli altri materiali e non arrechino ingiustificato pregiudizio agli interessi legittimi del titolare. Tali requisiti non sono soddisfatti nel caso in esame. La riproduzione rilevante ai fini del diritto d'autore è limitata allo scopo di analizzare i file delle immagini per la loro corrispondenza con preesistenti descrizioni delle stesse e la successiva inclusione in un *dataset*. Non è evidente né è stato affermato dall'attore che questa utilizzazione possa compromettere le possibilità di sfruttamento delle opere interessate.

Il *dataset* creato in questo modo potrebbe essere successivamente utilizzato per allenare reti neurali artificiali e i contenuti generati dall'IA che ne derivano potrebbero entrare in concorrenza con le opere degli autori (umani). Tuttavia, ciò da solo non giustifica la qualificazione della creazione dei *dataset* di allenamento come una compromissione dei diritti di sfruttamento delle opere ai sensi dell'art. 5, par. 5 della Direttiva *InfoSoc*. Questo vale solo perché i considerati sviluppi tecnologici, attualmente non ancora totalmente prevedibili, non consentono una delimitazione giuridicamente sicura tra utilizzi consentiti e non consentiti (si veda sopra, punto (b)).

Poiché, sulla base dell'attuale sviluppo tecnologico, non si può escludere, nel dubbio, che le conoscenze ottenute attraverso l'estrazione di testo e dati siano utilizzate per addestrare reti neurali artificiali che possono poi entrare in concorrenza con gli autori, l'opinione contraria porterebbe, in ultima analisi, addirittura a dover vietare completamente l'estrazione di testo e dati ai sensi del § 44b *UrhG*; una tale abrogazione completa della limitazione sarebbe però chiaramente contraria all'intenzione del legislatore e non può quindi rappresentare un risultato interpretativo sostenibile.

(2) Il file della fotografia scaricato dalla convenuta era anche legittimamente accessibile ai sensi del § 44b, par. 2, primo periodo *UrhG* — cosa che, tra l'altro, l'attore non contesta.

“Legittimamente accessibile” in tal senso è, in particolare, un'opera quando è liberamente accessibile su Internet (*Begr. RegE BT-Drs. 19/27426*, p. 88).

Tale [definizione] si applica all'immagine scaricata dalla convenuta. Contrariamente a quanto sostenuto inizialmente dall'attore, la convenuta non ha scaricato la “fotografia originale” riprodotta nella domanda di ingiunzione inizialmente formulata nell'atto di citazione — che sarebbe stata resa disponibile dall'agenzia fotografica “...” solo previo acquisto di una licenza — ma ha scaricato una versione dell'immagine con una filigrana posta dall'agenzia. Si trattava evidentemente dell'immagine in anteprima inserita sul sito dell'agenzia a scopo pubblicitario. Tale immagine in anteprima con la filigrana è stata tuttavia resa liberamente accessibile su Internet da parte dell'agenzia.

b) Vi sono però alcune indicazioni — senza che sia necessaria una decisione definitiva sul punto — che fanno ritenere che la limitazione di cui al §44b par. 2 *UrhG* non si sarebbe applicata al caso di specie, in quanto è stata espressamente dichiarata una riserva di utilizzo ai sensi del par. 3 della disposizione; in particolare, la riserva di utilizzo indiscutibilmente dichiarata sul sito web “...com” è in grado di soddisfare i requisiti di leggibilità automatizzata ai sensi del §44b, par. 3, secondo periodo *UrhG*.

(1) Molti elementi inducono a ritenere che la riserva di utilizzo riportata sul sito web dell'agenzia sia stata emessa da una persona autorizzata a farlo e che l'attore possa fare affidamento su di essa per tutelare i propri diritti.

Secondo il testo del §44b, par. 3 *UrhG*, “il titolare dei diritti” può dichiarare la riserva di utilizzo. Ciò significa che devono essere prese in considerazione non solo le dichiarazioni di riserva dell'autore stesso, ma

anche quelle dei successivi titolari dei diritti, siano essi successori *ex lege* o titolari di diritti derivati dall'autore. Secondo la coerente dichiarazione dell'attore "...", egli aveva concesso all'agenzia fotografica "...", semplici diritti di utilizzo dell'immagine originale che potevano essere concessi in sublicenza. Di conseguenza, l'agenzia fotografica era diventata essa stessa titolare di diritti sulle immagini pubblicate sulla sua piattaforma ed era quindi pienamente autorizzata a rilasciare una riserva di utilizzo ai sensi del §44b, par. 3 *UrhG*. Non vi è alcuna prova o constatazione che ci fossero accordi nell'ambito del rapporto contrattuale tra l'attore e l'agenzia fotografica che lo avrebbero precluso.

L'attore è probabilmente anche autorizzato a fare affidamento sulla riserva di diritti dichiarata dal suo licenziatario. Da un punto di vista economico, lo sfruttamento della fotografia originale in questione è avvenuto tramite l'agenzia. In pratica, la decisione in concreto su quale terzo dovesse ricevere l'autorizzazione per quale utilizzo spettava all'agenzia. In una situazione di questo tipo, il Tribunale ritiene ragionevole che l'autore, nel far valere i diritti da lui conservati, possa fare affidamento sulla riserva dichiarata dal suo licenziatario ai sensi del §44b, par. 3 *UrhG*.

(2) È altresì irrilevante l'eccezione della convenuta secondo la quale il divieto di utilizzo per i *web crawler* indicato nelle condizioni generali di contratto dell'agenzia nei confronti dei suoi clienti non potrebbe essere formulato da un punto di vista temporale "in relazione al § 44b par. *UrhG*". Non è requisito per l'efficacia della dichiarazione che essa sia consapevolmente espressa alla luce di una specifica formulazione normativa.

(3) La riserva è anche formulata in modo sufficientemente chiaro. L'art. 4, par. 3 Direttiva DSM richiede una dichiarazione espressa della riserva di utilizzo. Questo requisito deve quindi essere considerato nell'interpretare il §44b, par. 3 *UrhG* conformemente alla Direttiva (così anche nella *Begr. RegE BT-Drs. 19/27426*, p. 89). La riserva deve quindi essere dichiarata sia esplicitamente (non in modo implicito) sia in modo così preciso e mirato (in modo concreto-individuale) da rendere chiaro un determinato contenuto e un determinato utilizzo (*Hamann, ZGE 16* (2024), p. 134). La riserva di utilizzo formulata sul sito web dell'agenzia fotografica "...", soddisfa senza problemi tali requisiti.

Non risulta convincente l'argomentazione aggiuntiva secondo la quale una riserva d'utilizzo espressa per tutte le opere pubblicate su un sito web contraddirebbe il requisito di esplicitazione richiesto dal § 44b, par. 3 *UrhG* (così, *Hamann, op. cit.*, p. 148). Infatti, anche la riserva espressa per tutte le opere pubblicate su un sito web è determinabile senza dubbio nella sua portata e nel suo contenuto e quindi risulta dichiarata esplicitamente.

(4) Infine, ci sono buone ragioni per ritenere che la riserva di utilizzo soddisfi i requisiti di leggibilità automatizzata ai sensi del § 44b, par. 3, secondo periodo *UrhG*.

Il concetto di leggibilità automatizzata, considerando la volontà del legislatore di consentire interrogazioni automatizzate da parte dei *web crawler* (cfr. *Begr. RegE BT-Drs. 19/27426*, p. 89), dovrebbe essere

interpretato nel senso di “comprensibilità meccanica” (per una panoramica delle opinioni cfr. *Hamann, op. cit.*, pp. 113, 128 ss.).

Il Collegio è incline a considerare “comprensibile meccanicamente” anche una riserva di utilizzo redatta esclusivamente in “linguaggio naturale” (a differenza dell’orientamento probabilmente prevalente in dottrina; vedi *Hamann, op. cit.*, pp. 131 ss., 146 ss., con ulteriori riferimenti all’orientamento, incluso un riferimento ad un articolo dei rappresentanti legali della convenuta in giudizio, ovvero *Akinci/Heidrich, IPRB 2023, 270, 272*, che apparentemente sostengono l’opinione del Collegio; tuttavia, il contributo non era direttamente accessibile al Collegio fino alla definizione del giudizio). Tuttavia, la questione se e sotto quali condizioni specifiche una riserva dichiarata in “linguaggio naturale” possa essere considerata “comprensibile meccanicamente” dovrà sempre essere risolta in base allo sviluppo tecnologico esistente al momento dello sfruttamento rilevante dell’opera.

Coerentemente, anche il legislatore europeo, nel contesto del Regolamento sull’IA, ha stabilito che i fornitori di modelli di IA debbano adottare una strategia, in particolare, per identificare e rispettare la riserva di diritti espressa ai sensi dell’art. 4, par. 3 Direttiva DSM, “anche attraverso tecnologie all’avanguardia” (Art. 53, par. 1, lettera c, del Regolamento sull’IA). Tra queste “tecnologie all’avanguardia” rientrano senza dubbio anche le applicazioni di IA in grado di comprendere il contenuto di un testo scritto in linguaggio naturale (così, apparentemente, anche i rappresentanti della convenuta *Akinci/Heidrich* nel contributo *IPRB 2023, 270, 272*, citato da *Hamann, op. cit.*, p. 148, che in linea di principio affermano questa possibilità dal punto di vista tecnico). In tale contesto, tutto suggerisce che il legislatore del Regolamento sull’IA avesse in mente proprio tali applicazioni intelligenti riferendosi a “tecnologie all’avanguardia”.

Contro tale orientamento, alcuni osservano che [tale interpretazione] porterebbe ad un circolo vizioso: se si richiede che nell’ambito dell’estrazione di testo e dati l’operatore debba verificare tramite applicazioni di Intelligenza Artificiale se sia stata espressa una riserva di utilizzo, allora tale ricerca supportata dall’IA richiederebbe a sua volta un’analisi che già soddisferebbe la definizione di “estrazione di testo e dati” ai sensi del §44b, par. 1 *UrhG*; in altre parole: l’applicazione della limitazione determinerebbe innanzitutto l’ammissibilità della sua stessa applicazione (così *Hamann, op. cit.*, p. 148). Il Collegio non condivide questo orientamento: l’atto di utilizzo rilevante ai fini del diritto d’autore, che necessita di legittimazione, non è, contrariamente alla sopracitata opinione, l’attività di un’“analisi” in quanto tale, bensì la riproduzione dell’opera protetta dal diritto d’autore ai sensi del §16 *UrhG*. Non convince l’argomentazione secondo la quale la precedente scoperta di tali opere su Internet e la verifica circa la sussistenza di riserve espresse ex §44b par. 3, secondo periodo *UrhG* richiedano necessariamente un’ulteriore analisi quasi preliminare di *data mining* ai sensi §44b, par. 1 *UrhG*, in quanto si potrebbe immaginare che l’elaborazione dei contenuti di un sito web avvenga tramite dei *web crawler*, i quali danno luogo solo a riproduzioni transitorie e accessorie già legittimate ai sensi del §44a *UrhG*.

In aggiunta, l'interpretazione più ampia del termine "leggibilità automatizzata" considerata dal Collegio potrebbe essere opinabile in quanto tale termine è inteso in modo più ristretto dal legislatore europeo in altri contesti. Si farebbe, con ciò, riferimento al Considerando 35 della Direttiva PSI (Direttiva (UE) 2019/1024), che richiede, tra l'altro, una riconoscibilità "semplice" per la "leggibilità" (vedi *BeckOK UrhR/Bomhard*, 42a ed. 15.2.2024, *UrhG* §44b, par. 31 con ulteriori riferimenti); tale [definizione] non potrebbe essere applicabile per una riserva formulata solo in linguaggio naturale. Ebbene, una simile argomentazione presupporrebbe che i termini utilizzati da entrambe le Direttive debbano essere intesi allo stesso modo. Il Collegio dubita che tale equiparazione delle terminologie sia convincente, perché le Direttive perseguono obiettivi differenti: mentre la Direttiva PSI riguarda l'accesso prettamente unilaterale del pubblico alle informazioni o l'obbligo in capo alle autorità pubbliche di pubblicare determinate informazioni, l'art. 4, par. 3, Direttiva DSM riguarda l'equilibrio tra gli interessi degli utenti all'estrazione di testo e dati (per poter operare nel modo più semplice e giuridicamente certo possibile) e gli interessi dei titolari dei diritti (per garantire i loro diritti nel modo più semplice ed efficace possibile). Secondo il Collegio, questo equilibrio di interessi non può essere risolto unilateralmente a favore degli utenti dell'estrazione di testo e dati considerando come sufficiente per l'efficacia della riserva di utilizzo espressa unicamente la soluzione tecnica più semplice possibile per gli utenti stessi. Una simile interpretazione sarebbe contraddetta anche dalla valutazione del legislatore della Direttiva DSM, che al Considerando 18 non richiede la dichiarazione della riserva "nel modo più semplice possibile", ma solo che sia compiuta "in maniera appropriata". Anche il legislatore tedesco richiede una dichiarazione "adeguata ai processi automatizzati di estrazione di testi e dati" (*Begr. RegE BT-Drs.* 19/27426, pag. 89).

Ad avviso del Collegio, vi sarebbe inoltre una certa contraddizione di principio nel consentire ai fornitori di modelli di Intelligenza Artificiale di sviluppare via via modelli di IA in grado di creare e comprendere i testi mediante la limitazione posta dal §44b, par. 2 *UrhG*, da un lato, e dall'altro non imporgli di utilizzare modelli di Intelligenza Artificiale già esistenti nell'ambito di applicazione della limitazione del §44b, par. 3, secondo periodo *UrhG*.

Se e in quale misura, al momento dell'atto di riproduzione nel 2021, fosse già disponibile una tecnologia sufficiente per la rilevazione automatizzata del contenuto della riserva di utilizzo in questione, non è stato finora dimostrato dalla parte attrice; l'attore ha fatto riferimento solo a servizi disponibili nel 2023 (replica, pag. 14 ss.). Tuttavia, ci sono indicazioni che la convenuta disponesse già di una tecnologia adeguata. Infatti, secondo l'argomentazione della convenuta, l'analisi condotta nell'ambito della creazione del *dataset* "..." sotto forma di confronto tra i contenuti delle immagini e le preesistenti descrizioni delle medesime richiedeva evidentemente anche e proprio una rilevazione del contenuto di queste descrizioni da parte del *software* utilizzato. In tale contesto, molto suggerisce che — in particolare anche per la convenuta — già nel 2021 fossero disponibili sistemi in grado di rilevare in modo automatizzato una riserva di utilizzo formulata in linguaggio naturale.

3. La convenuta può tuttavia avvalersi della limitazione riconosciuta al §60d *UrhG* riguardo alla riproduzione in questione.

Secondo tale disposizione, sono legittimi gli atti di riproduzione per l'estrazione di testo e dati a fini di ricerca scientifica compiuti da parte di organismi di ricerca.

a) La riproduzione è stata effettuata — come illustrato — per finalità di estrazione di testo e dati ai sensi del §44b, par. 1 *UrhG*. Essa è stata condotta per fini di ricerca scientifica ai sensi del §60d *UrhG*.

La ricerca scientifica consiste generalmente nel perseguimento metodico-sistematico di nuove conoscenze (*Spindler/Schuster/Anton*, 4a ed. 2019, *UrhG* § 60c linea 3; *BeckOK UrhR/Grübler*, 42a ed. 1.5.2024, *UrhG* § 60c linea 5; *Dreier/Schulze/Dreier*, 7a ed. 2022, *UrhG* § 60c linea 1). Il concetto di “ricerca scientifica”, ricomprendendo il metodico e sistematico “perseguimento” di nuova conoscenza, non dovrebbe essere interpretato in modo così restrittivo da includere solo quei passaggi direttamente associati con la generazione di nuove scoperte. Piuttosto, è sufficiente che il passaggio in questione sia diretto alla generazione di (futura) conoscenza, come spesso capita in numerose raccolte di dati che devono essere inizialmente effettuate per trarre le successive conclusioni empiriche. In particolare, il concetto di ricerca scientifica non presuppone alcun successivo successo nella ricerca.

Pertanto, contrariamente all'opinione dell'attore, anche la creazione di un *dataset* come il tipo in questione, il quale può fungere da base per l'allenamento di sistemi di Intelligenza Artificiale, può essere considerata quale ricerca scientifica nel senso sopracitato. Sebbene la creazione del *dataset* in quanto tale non sia ancora associata all'ottenimento di nuove conoscenze, essa rappresenta comunque un passaggio fondamentale nel suo impiego per il raggiungimento nuove conoscenze. Anche nel caso in esame può essere affermata l'esistenza di una tale finalità. È sufficiente che il *dataset* sia stato pubblicato gratuitamente e dunque messo a disposizione (anche) dei ricercatori nel campo delle reti neurali artificiali. È, in tal senso, irrilevante se il *dataset* sia stato utilizzato anche da aziende commerciali per allenare o sviluppare ulteriormente i loro sistemi di Intelligenza Artificiale, come sostenuto dall'attore riguardo ai servizi “...” e “...”, in quanto anche la ricerca condotta da aziende commerciali è ancora ricerca scientifica — anche se non privilegiata come tale ai sensi del §60c *UrhG* e ss.

In tale contesto non rileva la questione, controversa tra le parti, se la convenuta, oltre alla creazione di tali *dataset*, svolga anche ricerca scientifica sotto forma di sviluppo dei propri modelli di Intelligenza Artificiale.

b) La convenuta, inoltre, non persegue finalità lucrative ai sensi del §60d, par. 2, n. 1 *UrhG*.

Con riguardo alla questione se la ricerca non sia lucrativa, solo il tipo in concreto di attività scientifica perseguita assume rilevanza, mentre sono irrilevanti l'organizzazione e il finanziamento dell'ente nel quale la ricerca viene effettuata (Considerando 42 Direttiva *InfoSoc*).

Il perseguimento di finalità non lucrative da parte della convenuta riguardo alla creazione del *dataset* “...” in questione risulta già dal fatto che la convenuta lo mette a disposizione a titolo gratuito. Non è stato

sostenuto da parte attrice né risulta evidente che la creazione del *dataset* in questione serva allo sviluppo di un'offerta commerciale propria da parte della convenuta (cfr. su questo criterio *BeckOK IT-Recht/Paul*, 14a ed. 1.4.2024, *UrhG* § 60d linea 10). Ai fini della qualificazione dell'attività della convenuta, è irrilevante che il *dataset* possa essere utilizzato da aziende commerciali per addestrare i loro sistemi di Intelligenza Artificiale. La circostanza che singoli membri appartenenti alla convenuta, oltre alla loro attività per l'associazione, svolgano anche attività retribuite presso tali aziende non è sufficiente ad attribuire tali attività alla convenuta.

c) Alla convenuta non è inoltre precluso eccepire la limitazione di cui al §60d, par. 2, terzo periodo *UrhG*.

Secondo tale previsione, non possono beneficiare della limitazione di cui al §60d *UrhG* gli organismi di ricerca che collaborano con un'azienda privata che eserciti su di essa un'influenza determinante e abbiano un accesso privilegiato ai risultati della ricerca scientifica. L'onere della prova e l'argomentazione circa i presupposti fattuali dell'esclusione dal §60d, par. 2, terzo periodo *UrhG* sono, secondo il tenore letterale della norma, a carico dell'attore.

(1) L'argomentazione mossa dall'attore a supporto del fatto che la società “...” ha un'influenza diretta sulla convenuta attraverso il finanziamento del *dataset* in questione e la nomina di propri dipendenti a “posizioni rilevanti” presso la convenuta (replica, pag. 8) è priva di sostanza.

L'attore fa riferimento al mero fatto che uno dei co-fondatori della convenuta, il sig. “...”, sia impiegato presso “...” come “*Head of Machine Learning Operations*”, e che un altro membro della convenuta, il sig. “...”, lavori lì in qualità di “*Research Scientist*” (replica, pagg. 4-5). Tuttavia, il solo fatto che due membri dell'associazione lavorino per la società “...” non dimostra che questa abbia un'influenza determinante sul lavoro di ricerca della convenuta.

A parte ciò, l'attore non ha nemmeno sostenuto che la convenuta avesse concesso alla società “...” accesso privilegiato ai risultati della sua ricerca scientifica, cioè al *dataset* in questione. Piuttosto, è stato solo presentato che “...” abbia allenato il suo servizio “...” utilizzando tale *dataset* (replica, pagg. 8-9).

(2) Neanche l'argomentazione dell'attore contenuta nell'atto del 7.04.2024 in riferimento ad una chat avvenuta nel 2021 sulla piattaforma “...”, secondo la quale il co-fondatore della convenuta, il sig. “...”, avrebbe dichiarato di essere disponibile a concedere alla società “...” un accesso anticipato al *dataset* (al tempo ancora più piccolo) a seguito di un contributo finanziario di 5.000 \$ soddisfa i requisiti per l'applicazione dell'eccezione prevista dal §60d, par. 2, terzo periodo *UrhG*.

In tale contesto, può rimanere irrisolta la questione se la conversazione nella chat — non contestata in quanto tale da parte della convenuta (atto del 9.07.2024, par. 3) — supporti effettivamente l'interpretazione fornita dall'attore. Allo stesso modo, può rimanere irrisolta la domanda se la dichiarazione di tale disponibilità a concedere un accesso anticipato — non è stato peraltro sostenuto dall'attore se tale accesso sia stato effettivamente concesso — possa considerarsi sufficiente ad integrare un accesso

privilegiato ai risultati di ricerca ai sensi del §60d, par. 2, secondo periodo *UrhG*.

In ogni caso, non è stato dimostrato dall'attore né è altrimenti evidente che la società “...” abbia un'influenza determinante sulla convenuta. Per quanto riguarda gli eventuali legami personali tra la convenuta e le aziende del settore dell'Intelligenza Artificiale, si tratta delle società “...” e “...” (replica, pag. 4 ss.).

P.Q.M., — 1. La domanda è rigettata.

2. L'attore è condannato alle spese legali.

3. La decisione è provvisoriamente esecutiva. L'attore può evitare l'esecuzione da parte della convenuta mediante prestazione di una cauzione pari al 110% dell'importo esigibile in forza della presente sentenza, salvo che la convenuta presti, prima dell'esecuzione, una cauzione pari al 110% dell'importo da corrispondere.

**L'ESTRAZIONE DI TESTO E
DATI NELLA COSTRUZIONE
DI DATASET DI
ALLENAMENTO PER
MODELLI GENERATIVI:
PROFILI CRITICI E NUOVE
PROSPETTIVE PER IL
DIRITTO D'AUTORE SULLE
OPERE DIGITALI**

1. IL CASO TEDESCO.

Con sentenza 27 settembre 2024¹ il Tribunale di Amburgo si interroga sull'applicabilità della definizione di *text and data mining* (di seguito "TDM") ex art. 2 della Direttiva (UE) 2019/790² all'estrazione di dati finalizzata alla creazione di un *dataset* di allenamento per modelli di intelligenza artificiale (di seguito "IA"), e analizza il riferibile quadro normativo. Il caso riguarda la riproduzione di una fotografia da parte di un ente no-profit avente l'obiettivo di costruire, con i dati estratti dall'opera, un *dataset* di allenamento per sistemi di IA generativa. I titolari dei diritti d'autore adiscono il Tribunale di Amburgo avverso la riproduzione, sostenendone l'illegittimità sul presupposto che l'attività condotta dall'ente, consistendo nell'estrazione della forma espressiva dell'opera e non di semplici informazioni, non sia ricompresa all'interno della deroga prevista dalla Direttiva (UE) 2019/790 (di seguito Direttiva Copyright o DSM). I giudici, tuttavia, confermano la liceità della riproduzione ritenendo che, da un lato, la distinzione tra forma espressiva (*die geistige Schöpfung*) e contenuto informativo (*verbogenen Informationen*) è labile nell'ambito dell'attività di costruzione di un *dataset*, e, dall'altro, che la raccolta di informazioni finalizzata all'allenamento di un modello di IA generativa si ponga in linea di continuità con la logica sottesa alle deroghe ai diritti autorali previste dalla Dir. (UE) 2019/790 e confermate dall'art. 53 del Regolamento (UE) 2024/1689 sull'Intelligenza Artificiale (di seguito "AI Act")³. Di conseguenza, l'attività compiuta dall'ente è qualificabile come TDM ed è pertanto riconducibile all'ambito di applicazione della Direttiva.

2. IL *TEXT AND DATA MINING* NELLA DIRETTIVA (UE) 2019/790 E NEL REGOLAMENTO (UE) 2024/1689.

La decisione del Tribunale di Amburgo si inserisce all'interno dell'ampio dibattito riguardante la protezione dei diritti autorali nell'ambito dello sviluppo di sistemi di IA, ponendo in evidenza un tema particolarmente rilevante. Si tratta della legittimità dello sfruttamento delle opere digitali tramite l'attività di estrazione di testo e dati (di seguito *data*

¹ *Landgericht Hamburg*, AZ.:310 O 227/23, 27 settembre 2024.

² Direttiva (UE) 2019/790 del Parlamento Europeo e del Consiglio del 17 aprile 2019 sul diritto d'autore e sui diritti connessi nel mercato unico digitale e che modifica le direttive 96/9/CE e 2001/29/CE.

³ Regolamento (UE) 2024/1689 del Parlamento Europeo e del Consiglio del 13 giu-

gno 2024 che stabilisce regole armonizzate sull'intelligenza artificiale e modifica i regolamenti (CE) n. 300/2008, (UE) n. 167/2013, (UE) n. 168/2013, (UE) n. 2018/858, (UE) n. 2018/1139 e (UE) n. 2019/2144 e le direttive 2014/90/UE, (UE) 2016/797 e (UE) 2020/1828 (regolamento sull'intelligenza artificiale).

mining o più semplicemente TDM) finalizzata all'addestramento di modelli di IA generativa ⁴.

Com'è noto, l'estrazione di testo e dati è definita dalla Direttiva (UE) 2019/790 come « 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 » ⁵. Sovente impiegate a fini scientifici, di indagini statistiche e di mercato, in tempi più recenti le tecniche di *data mining* sono utilizzate nell'ambito dell'addestramento di modelli di Intelligenza Artificiale, sollevando alcune perplessità in merito alla tenuta dell'impianto derogatorio previsto nella Direttiva Copyright ai diritti di riproduzione autorale ⁶. La Direttiva, infatti, interviene disciplinando il TDM alla presenza di due esigenze contrapposte: da un lato, l'interesse dei singoli autori a tutelare lo sfruttamento economico delle proprie opere dell'ingegno, inclusi *software* e banche di dati, nelle sue più ampie forme ⁷; dall'altro, lo sviluppo di soluzioni tecnologiche da parte di terzi in grado di facilitare le indagini commerciali, statistiche, industriali e scientifiche in presenza della disponibilità di ampie quantità di dati nell'ambiente digitale ⁸. La scelta del

⁴ Art. 3, par. 1), n. 63), Reg. (EU) 2024/1689: « « modello di IA per finalità generali »: un modello di IA, anche laddove tale modello di IA sia addestrato con grandi quantità di dati utilizzando l'auto supervisione su larga scala, che sia caratterizzato da una generalità significativa e sia in grado di svolgere con competenza un'ampia gamma di compiti distinti, indipendentemente dalle modalità con cui il modello è immesso sul mercato, e che può essere integrato in una varietà di sistemi o applicazioni a valle, ad eccezione dei modelli di IA utilizzati per attività di ricerca, sviluppo o prototipazione prima di essere immessi sul mercato ».

⁵ Art. 2, n. 2) Dir. (UE) 2019/790.

⁶ M. VON WELSER, *Generative AI und Urheberrechtsschranken*, in *GRUR-Prax.*, 2023, 516 ss.; T. W. DORNIS, S. STÖBER, *Urheberrecht und Training generativer KI-Modelle - Technologische und juristische Grundlagen*, Baden, 2024, 94 ss.; H. SCHACK, *Auslesen von Webseiten zu KI-Trainingszwecken als Urheberrechtsverletzung de lege lata et ferenda*, in *NJW*, 2024, n. 1, 113 ss.; P. HANDERSON, X. LI, D. JURAFSKY, T. HASHIMOTO, M. A. LEMLEY, P. LIANG, *Foundation Models and Fair Use*, in *Journal of Machine Learning Research*, 2023, n. 24, 10 ss.

⁷ Ampiamente è stato scritto sulla natura del diritto d'autore e la sua tutela nell'ordinamento europeo ed italiano. Per un'ampia disamina della natura e delle tematiche sottese al diritto d'autore nella società digitale, G. FINOCCHIARO, *L'equilibrio titolare/users nel diritto d'autore dell'Unione Europea*, in questa *Rivista*, 2016, n. 3, 499 ss.; G. SPEDICATO, *Principi di di-*

ritto d'autore, Bologna, 2020; S. ERCOLANI, *Il diritto d'autore e i diritti connessi - La legge n. 633/1941 dopo l'attuazione della direttiva n. 2001/29/CE*, Torino, 2004; G. SPEDICATO, *Interesse pubblico e bilanciamento nel diritto d'autore*, Bologna, 2013; R. MONCILLLO, *Proprietà intellettuale ed opera fotografica*, Napoli, 2018; G. COLANGELO, *La proprietà delle idee: le privative intellettuali tra comparazione e analisi economica*, Bologna, 2015; M. L. Montagnani, M. Borghi (a cura di), *Proprietà digitale: diritti d'autore, nuove tecnologie e digital rights management*, Milano, 2006; O. POLICINO, *Tutela del diritto d'autore e protezione della libertà di espressione in chiave comparata. Quale equilibrio su Internet?*, in F. Pizzetti (a cura di), *I diritti "nella rete" della rete - Il caso del diritto d'autore*, Torino, 2013, 60 ss.; P. AUTERI, *Il diritto di autore*, in G. Olivieri, M. Ricolfi (a cura di), *Diritto industriale, proprietà intellettuale e concorrenza*, Torino, 2020; L. MOSCATI, *Diritti d'autore - Storia e comparazione nei sistemi di civil law e di common law*, Milano, 2020, 57 ss.; T. BRUNO, *Diritti d'autore*, in L. Lucchini (diretto da), *Digesto italiano*, IX, parte II, 1899, Torino, 548 ss.; F. BENATTI, *Diritto d'autore e supporto dell'opera*, Torino, 2020; M. FABIANI, *Il diritto d'autore*, in P. Rescigno (diretto da), *Trattato dir. priv.*, Milano, 1983, 131 ss.; L. CHIMIENTI, *Lineamenti del nuovo diritto di autore*, VI ed., Milano, 2006; E. PIOLA CASSELLA, *Codice del diritto di autore*, Torino, 1943.

⁸ A. MUSSO, *Eccezioni e limitazioni ai diritti d'autore nella Direttiva UE n. 790/2019*, in questa *Rivista*, 2020, 411 ss.; L.

legislatore europeo è ricaduta nell'individuazione di due discipline differenti, il cui ambito di applicazione dipende dalla finalità in concreto perseguita tramite le tecniche di *data mining*. Con riferimento all'art. 3 DSM, il legislatore ha demandato agli Stati Membri di porre limitazioni ai diritti di riproduzione dell'opera⁹ e di estrazione dalle banche di dati¹⁰ unicamente quando il *data mining* è svolto nell'ambito di ricerche scientifiche prive di scopi lucrativi condotte da organismi di ricerca e istituti di tutela del patrimonio culturale¹¹. Per tutte le finalità diverse dalla ricerca scientifica priva di scopo di lucro, le limitazioni ai diritti patrimoniali trovano applicazione a condizione che il titolare dei diritti non abbia espressamente riservato l'utilizzo dell'opera « in modo appropriato »¹². Entrambe le eccezioni trovano come contraltare l'esigenza di mantenere

MANSANI, *Le eccezioni per estrazione di testo e dati, didattica e conservazione del patrimonio culturale*, in AIDA, 2019, 3 ss.; M. GRANIERI, *Right of revocation of authors and performers in the European online copyright directive*, in AIDA, 2019, 128 ss.; M. GRANIERI, *Il data mining nella disciplina del diritto d'autore e la strategia europea dei dati*, in AIDA, 2022, 20 ss.; G. ROSSI, *Opere dell'ingegno come dati: il text and data mining nella direttiva 2019/790*, in AIDA, 2019, 235 ss.; A. OTTOLIA, *L'utilizzo computazionale dell'opera dell'ingegno in internet*, in AIDA, 2014, 386 ss.; A. OTTOLIA, *L'opt out commons nella nuova disciplina del data mining*, in *Giurisprudenza italiana*, 2022, 1253 ss.; S. ORLANDO, *Il diritto al text and data mining (TDM) non esiste*, in *Rivista italiana di informatica e di diritto*, n. 1, 2023, 67 ss.; R. SERVANZI, *Le estrazioni di testo e dati*, in *Le nuove leggi civili commentate*, 2022, n. 5, 1151 ss.; T. MARCONI, M. KRETSCHMER, *A deeper look into the EU Text and Data Mining exceptions: Harmonisation, data ownership, and the future of technology*, in *CREATE Working Paper*, 2021, n. 7, 10 ss.; W. CARRARO, S. LONGHINI, F. UNNIA (a cura di), *Internet e diritto d'autore: la Direttiva Copyright e la sua applicazione in Italia*, Milano, 2022; S. LAVAGNINI, *Il diritto d'autore nel mercato unico digitale: Direttiva (UE) 2019/790 e D.lgs. n. 177/2021 di recepimento*, Torino, 2022; P. GITTO, *New York Times vs. OPE-NAI, Microsoft et al.: conflitti attuali fra intelligenza artificiale e diritto d'autore*, in *Giustizia civile*, 2024, n. 1, 185 ss.; A. AMIDEI, *La tutela del diritto d'autore online tra responsabilità e libertà di espressione*, in *Giurisprudenza Italiana*, 2024, n. 2, 483 ss.; A. G. PARISI, *Intelligenza artificiale e diritto d'autore*, in *Comparazione e diritto civile*, 2023, n. 3, 1195 ss.; L. NIVARRA, *Intelligenza artificiale e lesione della proprietà intellettuale*, in *La Nuova Giurisprudenza Civile Commentata*, 2024, n. 1, 153 ss.; nella dottrina tedesca, con riferimento

alla Direttiva 2019/790 e alle norme di recepimento nel *Urheberrecht Gesetz*, P. HACKER, *Ein Rechtsrahmen für KI-Trainingsdaten*, in ZGE, 2020, n. 12, 239 ss.; T. DREIER, *Die Schlacht ist geschlagen - ein Überblick zum Ergebnis des Copyright Package des EU-Kommission*, in GRUR, 2019, n. 1, 771 ss.; G. SPINDLER, *Text- und Datamining im neuen Urheberrecht un in der europäischen Diskussion*, in ZGE, 2018, n. 2, 273 ss.; H. HAMANN, *Nutzungsvorbereitung für KI-Training in der Rechtsgelehrtschaftslehre der Maschinenkommunikation*, in ZGE, 2024, n. 1, 113 ss.

⁹ Art. 5, lett. a) Dir. 96/9/CE; art. 2, Dir. 2001/29/CE; art. 15 par. 1 Dir. (UE) 2019/790.

¹⁰ Art. 7, par. 1, Dir. 96/9/CE.

¹¹ Art. 3, par. 1), Dir. (UE) 2019/790: « Gli Stati membri dispongono un'eccezione ai diritti di cui all'articolo 5, lettera a), e all'articolo 7, paragrafo 1, della direttiva 96/9/CE, all'articolo 2 della direttiva 2001/29/CE, e all'articolo 15, paragrafo 1, della presente direttiva per le riproduzioni e le estrazioni effettuate da organismi di ricerca e istituti di tutela del patrimonio culturale ai fini dell'estrazione, per scopi di ricerca scientifica, di testo e di dati da opere o altri materiali cui essi hanno legalmente accesso ». L'art. 70-ter, l. 633/1941, norma di recepimento nell'ordinamento interno, ha confermato tra i soggetti beneficiari dell'eccezione gli organismi senza scopo di lucro che perseguono finalità di interesse pubblico non soggetti ad influenze dominanti da parte di imprese commerciali. Sul punto, M. GRANIERI, *Il data mining nella disciplina del diritto d'autore e la strategia europea dei dati*, cit., 32.

¹² Art. 4, Dir. (UE) 2019/790: « Gli Stati membri dispongono un'eccezione o una limitazione ai diritti di cui all'articolo 5, lettera a), e all'articolo 7, paragrafo 1, della direttiva 96/9/CE, all'articolo 2 della direttiva 2001/29/CE, all'articolo 4, paragrafo 1, lettere a) e b), della direttiva 2009/24/CE e

protetti i diritti di sfruttamento insistenti sull'opera da parte dei suoi titolari. Per tale ragione, l'art. 7 par. 2 DSM prevede che le opere dell'ingegno e altri materiali possono essere riprodotti per finalità di estrazione di testo e dati ove la riproduzione o l'estrazione a tali scopi non si ponga in contrasto con lo sfruttamento normale dell'opera e non arrechi ingiustificato pregiudizio agli interessi legittimi dei suoi titolari ¹³.

Calate tali considerazioni sul piano dello sviluppo di sistemi di Intelligenza Artificiale, il legislatore europeo è intervenuto da ultimo con l'art. 53 AI Act prevedendo che i fornitori ¹⁴ di modelli di IA per finalità generali si adoperino per individuare e rispettare le riserve di diritti espresse dagli autori ai sensi dell'art. 4 DSM ¹⁵ durante tutto l'arco di sviluppo e messa in commercio dei modelli in questione ¹⁶. Il quadro normativo così delineato parrebbe dunque propendere per un'inclusione del regime previsto dalla Direttiva nell'ambito di tutto il ciclo di vita del modello di Intelligenza Artificiale a finalità generali, ponendosi in linea di continuità con l'esigenza, già avvertita nel 2019, di promuovere lo sviluppo delle tecnologie innovative senza limitare *in toto* lo sfruttamento economico delle opere nell'ambiente digitale da parte dei loro titolari ¹⁷.

all'articolo 15, paragrafo 1, della presente direttiva per le riproduzioni e le estrazioni effettuate da opere o altri materiali cui si abbia legalmente accesso ai fini dell'estrazione di testo e di dati. Le riproduzioni e le estrazioni effettuate a norma del paragrafo 1 possono essere conservate per il tempo necessario ai fini dell'estrazione di testo e di dati. L'eccezione o la limitazione di cui al paragrafo 1 si applica a condizione che l'utilizzo delle opere e di altri materiali di cui a tale paragrafo non sia stato espressamente riservato dai titolari dei diritti in modo appropriato, ad esempio attraverso strumenti che consentano lettura automatizzata in caso di contenuti resi pubblicamente disponibili online». Attenta dottrina ha peraltro rilevato come il legislatore europeo non abbia inteso affrontare la relazione che lega le opere dell'ingegno alla loro trasformazione in meri dati. Cfr. G. Rossi, *Opere dell'ingegno come dati: il text and data mining nella direttiva 2019/90*, in *AIDA*, 2019, 252 ss.

¹³ Art. 7 par. 2, Dir. (UE) 2019/790: «L'articolo 5, paragrafo 5, della direttiva 2001/29/CE si applica alle eccezioni e alle limitazioni di cui al presente titolo. L'articolo 6, paragrafo 4, primo, terzo e quinto comma, della direttiva 2001/29/CE si applica agli articoli da 3 a 6 della presente direttiva».

¹⁴ Art. 3, par. 1, n. 3), Reg. (EU) 2024/1689: ««fornitore»: una persona fisica o giuridica, un'autorità pubblica, un'agenzia o un altro organismo che sviluppa un sistema di IA o un modello di IA

per finalità generali o che fa sviluppare un sistema di IA o un modello di IA per finalità generali e immette tale sistema o modello sul mercato o mette in servizio il sistema di IA con il proprio nome o marchio, a titolo oneroso o gratuito».

¹⁵ Art. 53, par. 1, lett. c), Reg. (UE) 2024/1689: «I fornitori di modelli di IA per finalità generali: [...] attuano una politica volta ad adempiere al diritto dell'Unione in materia di diritto d'autore e diritti ad esso collegati e, in particolare, a individuare e rispettare, anche attraverso tecnologie all'avanguardia, una riserva di diritti espressa a norma dell'articolo 4, paragrafo 3, della direttiva (UE) 2019/790».

¹⁶ Gli adempimenti previsti dall'art. 53 Reg. (UE) 2024/1689 in capo ai fornitori di modelli di IA a finalità generali entrano in vigore dal 2 agosto 2025.

¹⁷ Solleva perplessità sull'impostazione scelta dal legislatore europeo, per esempio, EUROPEAN COPYRIGHT SOCIETY, *Copyright and Generative AI: Opinion of the European Copyright Society*, January 2025, consultabile al seguente link ecs_opinion_genai_january2025.pdf (consultato il 18 febbraio 2025), la quale ha sollevato dubbi sulla mancanza di una chiara definizione dei limiti applicativi degli artt. 3-4 Dir. (UE) 2019/790 con riferimento all'intero ciclo di vita dei modelli di IA a finalità generali. In particolare, non è chiaro se tutte le riproduzioni ed estrazioni di opere protette dal diritto d'autore coinvolte nel processo di sviluppo e allenamento dell'IA possano qualificarsi come TDM. In dottrina, sono avan-

3. ALCUNI PROFILI CRITICI SULL'INTERPRETAZIONE ESTENSIVA DELLA DEFINIZIONE DI TDM DEL TRIBUNALE DI AMBURGO: L'ESTENSIONE DEL REGIME DI ECCEZIONE ALLA CREAZIONE DI DATASET DI ALLENAMENTO PER MODELLI DI INTELLIGENZA ARTIFICIALE.

Nell'ambito della sentenza del Tribunale di Amburgo, il delineato regime di eccezione è stato considerato applicabile al caso in cui la riproduzione dell'opera fotografica¹⁸ fosse intesa alla costruzione di un *dataset* di allenamento tramite l'utilizzo di tecniche di analisi automatizzata. Eppure, la decisione solleva alcune perplessità alla luce dei confini definitivi dell'estrazione di testo e dati previsti dalla Direttiva. Come è stato fatto notare da autorevole dottrina, l'attività di *data mining* è tesa ad individuare correlazioni e informazioni sottoponendo i contenuti digitali a strumenti di analisi automatizzata in grado di utilizzare il relativo « valore non espressivo »¹⁹. In altre parole, il TDM consente di sfruttare le opere digitali solo in quanto fonti di dati, non interessando tanto il carattere creativo delle stesse quanto i rapporti che possono emergere dalla combinazione dei dati contenuti al loro interno.

Sul punto parrebbe che, quando tesa alla costruzione di *dataset* di allenamento per modelli di IA generativa, l'attività in commento si ponga in via eccezionale rispetto alla citata funzionalità di generazione di correlazioni²⁰. Ciò sia in considerazione della particolare finalità per la quale tali tecniche sono utilizzate — la creazione di un insieme di dati

zate alcune considerazioni critiche in relazione all'applicabilità della Dir. (UE) 2019/790 all'estrazione di testo e dati finalizzata all'addestramento di sistemi di IA generativa. Cfr. M. VON WELSER, *Generative KI und Urheberrechtsschranken*, in *GRUR-Prax*, 2023, 516 ss., il quale ritiene che la deroga ai diritti autoriali non debba trovare applicazione con riferimento allo sviluppo di IA generativa in quanto la Direttiva sarebbe tesa a disciplinare il TDM quale strumento di mera analisi dei dati e non, invece, quale attività tesa alla generazione di nuovi contenuti. Dello stesso avviso, T. W. DORNIS, S. STÖBER, *Urheberrecht und Training generativer KI-Modelle - Technologische und juristische Grundlagen*, cit., 94 ss.

¹⁸ §72 *UrhG*: « *Lichtbilder und Erzeugnisse, die ähnlich wie Lichtbilder hergestellt werden, werden in entsprechender Anwendung der für Lichtbildwerke geltenden Vorschriften des Teils I geschützt* ». La fotografia è tutelata dal diritto d'autore tedesco ai sensi del §72 *UrhG*, il quale equipara le fotografie e le produzioni espresse con procedimento analogo alle opere fotografiche sul piano dei diritti morali e patrimoniali riconosciuti nella *Teil I* della *Urheberrechtsgesetz*. Le fotografie semplici godono di tutela fino a cinquant'anni dal momento di prima pubblicazione o, in assenza di pubblicazione, dal momento della sua produzione. Nel diritto d'autore italiano, è man-

tenuta la distinzione tra opera fotografica e fotografia semplice sul piano della tutela del diritto d'autore. Mentre l'opera fotografica è stata riconosciuta quale vera e propria opera dell'ingegno a partire dalla riforma intervenuta con d.p.r. 3 gennaio 1979 n. 19 e inclusa nel novero dell'art. 2 l.d.a., la tutela per le fotografie prive di carattere creativo è limitata ai diritti connessi, e si protrae per vent'anni dal momento di prima riproduzione. Cfr. R. MONCILLO, *Proprietà intellettuale ed opera fotografica*, in *Quaderni della Rivista di Diritto dell'Impresa*, Napoli, 2018, 49 ss.

¹⁹ Cit. A. OTTOLIA, *L'utilizzo computazionale dell'opera dell'ingegno in internet*, cit., 408. Il TDM può presentare numerose applicazioni. Si annoverano a titolo esemplificativo il cd. *text classification*, spesso utilizzato nell'ambito delle ricerche per categorie nelle banche di dati; l'*information extraction*, tipicamente impiegato nell'ambito finanziario per la valutazione creditizia di una persona; l'*opinion mining*, che, permettendo la classificazione del testo sulla base di informazioni soggettive quali opinioni o atteggiamenti, è spesso utilizzato dalle imprese al lancio di un nuovo prodotto per valutarne l'accoglienza. Cfr. C. ZONG, R. XIA, J. ZHANG, *Text Data Mining*, Beijing, 2021, 3 ss.

²⁰ T. W. DORNIS, S. STÖBER, *Urheberrecht und Training generativer KI-*

referiti ai contenuti digitali da cui sono estratti ²¹ — sia, più precipuamente, per la tipologia di contenuti spesso riprodotti a tale scopo da parte degli sviluppatori. Nello scenario delineato, la riproduzione delle opere è volta, dunque, alla costruzione di *dataset* a loro volta impiegati nell'addestramento di modelli generativi di contenuti multimediali ²². Pare dunque opportuno interrogarsi se tale scopo giustifichi l'applicazione del regime di eccezione di cui alla Direttiva oppure se le caratteristiche tecniche dell'attività in commento possano considerarsi peculiari al punto da ricondurne l'applicazione entro le tutele previste dalla normativa generale ²³.

4. LA DIFFICILE ADERENZA DELL'ANALISI AUTOMATIZZATA PER LA CREAZIONE DI DATASET DI ALLENAMENTO ALLA DEFINIZIONE DI TEXT AND DATA MINING DELLA DIRETTIVA COPYRIGHT.

Le riflessioni circa l'applicabilità delle deroghe della Direttiva 2019/790 alla creazione di *dataset* di allenamento parrebbero non poter prescindere da un'analisi in concreto riguardante le modalità tecniche utilizzate a tal

Modelle, Berlin, 103 ss. Gli a. valutano il diverso impatto dell'estrazione a fini di allenamento sui diritti d'autore sulla base della distinzione tra contenuto "sintattico" e contenuto "semantico" dell'opera. Mentre il primo consisterebbe nella forma espressiva dell'idea ricadendo quindi nell'ambito applicativo del diritto d'autore, il secondo riguarderebbe le informazioni meramente fattuali del contenuto estratto e ricadrebbe pertanto nell'ambito di applicazione delle eccezioni previste dalla Direttiva Copyright.

²¹ È ancora ampiamente discussa la configurabilità di una tutela autorale insistente sulle opere generate da sistemi di Intelligenza Artificiale. Da un lato, vi è chi sostiene che i contenuti *AI-Generated* non possano trovare protezione nell'ambito del diritto d'autore proprio in mancanza dell'apporto minimo creativo da parte di un umano nelle fasi di generazione dei contenuti. In giurisprudenza, le prime pronunce delle corti americane ed europee propendono dunque ad escludere tutelabilità all'opera automaticamente generata da sistemi intelligenti. Cfr. *United States Court of Appeals for the district of Columbia Circuit, Thaler v. Perlmutter*, decision no. 23-5233, 18 marzo 2025; Tribunale di Praga, *č. j. 10 C 13/2023-16*, 11 ottobre 2023. Dall'altro lato vi è chi sostiene una forma di autorialità in seno allo sviluppatore oppure all'utente, sull'assunto che la creatività, seppure non manifestata nell'immagine generata, trova esplicazione nel momento della descrizione del contenuto che si intende creare. Sul punto, C. DI BENEDETTO, *Creatività artificiale e diritto d'autore*, in *AIDA*,

2022, 388 ss.; D. BOMIARD, N. A. GAJECK, *Softwareentwicklung durch Künstliche Intelligenz* in *RDi*, 2021, 472 ss.

²² Cfr. T.W. DORNIS, S. STROBER, *Urheberrecht und Training generativer KI-Modelle*, Berlin, 2024, 103 ss. Sul punto pare utile precisare che i modelli di Intelligenza Artificiale generativa non solo possono imitare lo stile degli autori, ma sono in grado di replicare esattamente i dati con i quali sono stati addestrati. Cfr. P. HANDERSON, X. LI, D. JURAESKY, T. HASHIMOTO, M. A. LEMLEY, P. LIANG, *Foundation Models and Fair Use*, cit., 10 ss., i quali hanno verificato che i cd. *foundation models*, particolari modelli di apprendimento automatico pre-addestrati su cui si basano piattaforme quali ChatGPT o Midjourney, sono in grado, all'inserimento di determinati *prompt* testuali, di riprodurre porzioni integrali di opere dell'ingegno tutelate dal diritto d'autore, i cui dati sono stati utilizzati per il relativo allenamento.

²³ Si anticipa che l'approccio del legislatore nazionale sul punto potrebbe consistere nell'estendere il quadro normativo previsto alla Direttiva (UE) 2019/790 anche all'allenamento di sistemi di IA, non precisando, tuttavia, se il caso di costruzione di *dataset* di allenamento vi rientri effettivamente. L'art. 24 del disegno di legge n. 1146 "Disposizioni e delega al Governo in materia di intelligenza artificiale" presentato dalla Presidenza del Consiglio dei ministri include la riproduzione e l'estrazione di opere o altri materiali attraverso modelli e sistemi di intelligenza artificiale anche generativa nell'ambito di applicazione degli articoli 70ter e 70-quater l. 633/1940.

fine. Riprendendo la descrizione compiuta dal giudice di Amburgo, la contestata attività è legata all'utilizzo di specifici *software* in grado di analizzare le opere riprodotte, valutare se le descrizioni ad esse correlate siano corrette, ed estrapolare i *link* di provenienza dalle opere per inserirli nel *dataset* di allenamento. In altre parole, nell'ambito dell'attività automatizzata in esame, generalmente chiamata *scraping*²⁴, il *software* è in grado di valutare automaticamente la correttezza delle descrizioni già associate alle immagini e di estrarre una serie di dati, ciascuna composta dal *link* dell'immagine e dalla descrizione stessa, da inserire nel *dataset* di allenamento²⁵. La serie di dati così creata non si limita a riferirsi direttamente al contenuto dell'opera ad essa collegata, ma è generata in modo da rendere funzionale il suo riutilizzo nell'ambito dell'allenamento²⁶.

Ripercorrendo le caratteristiche dell'attività di estrazione di testo e dati²⁷, la riconduzione dell'attività in esame all'interno dei suoi confini definitivi pare risultare dubbia²⁸. Come correttamente da taluno osser-

²⁴ Sul *web scraping* in rapporto alla tutela del diritto *sui generis* delle banche dati, ampiamente in M. W. MONTEROSI, *Estrazione e (ri)utilizzo di informazioni digitali all'interno della rete Internet. Il fenomeno del c.d. web scraping*, in questa *Rivista*, 2020, 327 ss.; P. SAMMARCO, *L'attività di "web scraping" nelle banche dati ed il riuso delle informazioni*, in nota a Trib. Roma, 4 settembre 2019, in questa *Rivista*, 2020, n. 2, 219 ss.

²⁵ Sul contenuto della descrizione si veda, C. SCHUHMAN ET AL., *LAION-5B: An open large-scale dataset for training next generation image-text models*, in *36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks*, 2022, 4 ss.: « *The alt-text is an HTML attribute of IMG tags containing alternative text for situations where the corresponding image cannot be rendered. For instance, screen reader software for a visually impaired person may read the alt-text in place of an image, or a search engine may use the alt-text to better index a web page without analyzing the actual image content* ».

²⁶ Tali coppie di testo-immagini, in ampissime quantità, sono infatti spesso impiegate nell'allenamento dei modelli generativi in modo che, alla presentazione di determinati *input* di immagini o di testo, tali modelli riescano ad associare con un alto livello probabilistico quell'*input* con il contenuto dei dati con il quale è stato allenato. Una delle tecniche con le quali i modelli *general purpose* possono essere allenati mediante tali *dataset* è spiegata in A. DOSOVITSKIY ET AL., *An image is worth 16x16 words: transformers for image recognition at scale*, in *ICRL 2021 Papers*, 2021, 1 ss. e in C. SCHUHMAN ET AL., *LAION-5B: An open large-*

scale dataset for training next generation image-text models, cit., 34 ss., ove la descrizione associata alle immagini raccolte per la costruzione del *dataset* si riferisce ad elementi tutelabili dal diritto d'autore.

²⁷ Art. 2, n. 2), Dir. (UE) 2019/790: « 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 ». Generalmente, con *data mining* ci si riferisce al cd. TDM computazionale, consistente nell'analisi automatizzata finalizzata a creare correlazioni automatiche tra i dati (*pattern*) per giungere ad un nuovo e diverso risultato. Cfr. G. CERUTTI, R. TISCI, *L'impatto dell'intelligenza artificiale sui diritti di privacy da copyright*, in W. Carraro, S. Longhini, F. Unnia (a cura di), *Internet e diritto d'autore: la Direttiva Copyright e la sua applicazione in Italia*, Milano, 2022, 203 ss.

²⁸ Anche sul piano prettamente tecnico, tali tecniche di estrazione e raccolta di dati a partire da contenuti presenti in Rete sono generalmente denominate di *data scraping*, e svolgono una funzione differente rispetto a quelle di *data mining*. Le prime si limitano all'estrazione e raccolta di dati, senza procedere ad una successiva elaborazione dei dati raccolti; le seconde prevedono l'uso di *software* in grado di analizzare ampissime quantità di informazioni con lo scopo di individuare correlazioni tra i dati stessi. Spesso è possibile che le tecniche di *scraping* siano ricomprese tra i passaggi tecnici delle tecniche di *data mining*, in quanto servono lo scopo di estrarre o riprodurre i dati successivamente analizzati. Gli standard ISO/IEC, *Information technology*

vato ²⁹, gli elementi che consentono di identificare una determinata attività tecnica quale « estrazione di testo e dati » ai sensi dell'art. 2 DSM consistono, nell'ordine, nella presenza di un'analisi computazionale automatizzata di materiale digitale, nella generazione di informazioni, nell'ottenimento di nuove conoscenze o nella scoperta di nuove correlazioni. Tali elementi, che rendono necessaria la presenza di una elaborazione dei dati raccolti, non paiono essere presenti nell'attività volta alla costruzione di un *dataset* di allenamento. In questo caso, l'analisi automatizzata sull'opera conduce unicamente all'estrazione di dati già esistenti, con il principale scopo di raccogliarli e renderli disponibili per lo sviluppo di modelli generativi. In altre parole, mancando, in seno all'utilizzo del *software*, una rielaborazione dei dati raccolti, tale analisi tecnica non è improntata all'« acquisizione di nuove conoscenze » o alla « rilevazione di nuove tendenze » ³⁰. Un conto è, infatti, l'analisi automatizzata di dati che, comprendendo un'elaborazione degli stessi, si dimostra essere finalizzata alla generazione di correlazioni spesso nascoste agli occhi di chi analizza tali informazioni tramite strumenti tradizionali; un altro è, invece, « fermarsi » all'analisi di un'opera con l'unico scopo di trasformarla in dati strutturati senza aggiungere alcun elemento che non fosse già precedentemente noto ³¹ capace di mantenere con essa ³² un'associazione diretta ³³.

Il fatto che il risultato di tale analisi, dunque, consista in una raccolta di dati già esistenti nell'opera riprodotta rende l'attività di *scraping* in concreto teleologicamente differente dal *text and data mining* di cui alla

— *Artificial intelligence — Artificial intelligence concepts and terminology*, 2022, definiscono il *data mining* come « *computational process that extracts patterns by analysing quantitative data from different perspectives and dimensions, categorizing them, and summarizing potential relationships and impacts* », mentre non sussiste al momento una definizione standardizzata di *data scraping*.

²⁹ C. GALLESE, *Web scraping ed addestramento dei Modelli Generativi nella Direttiva 790/19*, in *i-lex - Rivista di Scienze Giuridiche, Scienze Cognitive ed Intelligenza Artificiale*, 2023, n. 16-2, 10 ss.; L. MANSANI, *Le eccezioni per estrazione di testo e dati, didattica e conservazione del patrimonio culturale*, cit., 3;

³⁰ Cons. 8), Dir. (UE) 2019/790.

³¹ Non convince pertanto il giudice di Amburgo nel considerare quale « correlazione » generata dal software la conformità/non conformità dell'immagine alla descrizione ad essa associata: « *Diese Analyse der Bilddatei zum Abgleich mit einer vorbestehenden Bildbeschreibung stellt ohne Weiteres eine Analyse zum Zwecke der Gewinnung von Informationen über „Korrelationen“ (nämlich der Frage der Nicht-/Übereinstimmung von Bildern und Bildbeschreibungen) dar* ». Da un lato, perché la correlazione è già presente al momento dell'analisi dell'immagine, talché il *software* si limita a confermarla estrapolando dalla

stessa l'URL di provenienza; dall'altro lato, perché il risultato ultimo dell'attività in commento non consiste nella raccolta degli esiti, bensì nel riutilizzo di dati preesistenti che hanno superato un « vaglio » di qualità.

³² L'originalità è uno degli elementi dell'opera che ne consente la sua protezione. L'opera originale è definita nella Direttiva 96/9/CE del Parlamento Europeo e del Consiglio, dell'11 marzo 1996, relativa alla tutela giuridica delle banche di dati, nella Direttiva 91/250/CE del Consiglio, del 14 maggio 1991, relativa alla protezione giuridica dei programmi per elaboratore, seguita dalla Direttiva 2009/24/CE del Parlamento europeo e del Consiglio, del 23 aprile 2009, relativa alla protezione giuridica dei programmi per elaboratore, e nella Direttiva 2006/115/CE del Parlamento europeo e del Consiglio, del 12 dicembre 2006, concernente il diritto di noleggio, il diritto di prestito e taluni diritti connessi al diritto di autore in materia di proprietà intellettuale come « creazione intellettuale dell'autore », facendo in tal modo riferimento sia all'elemento personalistico tipico del diritto d'autore sviluppatosi negli ordinamenti di *civil law* sia al requisito dell'abilità intellettuale ripreso dalla tradizione del *copyright* elaborata dagli ordinamenti di *common law*. Cfr. G. M. RICCIO, *Diritto d'autore, creatività e requisiti dell'opera: la Cassazione si affaccia (timidamente) nel dibattito europeo*, in

Direttiva, e consente di sollevare alcuni dubbi sulla scelta del giudice anseatico di ricondurre la riproduzione dell'opera all'interno delle deroghe di cui al Capo II della Direttiva³⁴. Tale considerazione emerge soprattutto alla luce della sostanziale differenza tra la finalità di creazione di un *dataset* e la finalità di allenamento di modelli generativi³⁵. Mentre la prima attività consiste nella costruzione di una raccolta di dati a partire dall'analisi automatizzata di grandi quantità di contenuti digitali³⁶, la seconda è volta all'elaborazione di contenuti e informazioni al fine di permettere al modello in allenamento di apprendere schemi (cd. *pattern*)

Le Nuove Leggi Civili Commentate, 2021, n. 1, 100 ss.

³³ Dello stesso avviso ma con argomentazioni differenti, M. VON WELSER, *Generative KI und Urheberrechtsschranken*, in *GRUR-Prax.*, 2023, 516 ss.; C. GALLESE, *Web scraping ed addestramento dei Modelli Generativi nella Direttiva 790/19*, cit.

³⁴ L'estrapolazione, l'inserimento dell'URL dell'opera nel *dataset* e la successiva messa a disposizione parrebbero, piuttosto, essere riconducibili al *framing* o *linking*, attività tipiche in grado di consentire collegamenti ipertestuali a contenuti presenti in altre pagine della Rete. In entrambi i casi, infatti, è consentito all'internauta, tramite *click*, di risalire facilmente all'opera di origine e, conseguentemente, di poterla sfruttare. Sul punto, la Corte di Giustizia Europea ha avuto modo di individuare un bilanciamento tra l'esigenza di tutelare i diritti d'autore e i diritti degli utenti alla libertà di espressione e di informazione, stabilendo che ove l'autore metta a disposizione l'opera online senza misure che ne impediscano la riproduzione e la diffusione, si presume che abbia acconsentito a che terzi compissero atti di comunicazione tramite *linking*. Cfr. Corte giust. UE, C-392/19, sent. 9 marzo 2021, n. 392, con nota di L. BOCCIO, *Diritto d'autore - Dalla legittimità (condizionata) del framing ad uno statuto legale (in fieri) per gli ISP*, in *Giurisprudenza Italiana*, n. 11, 2385 ss. Nel caso in esame, l'autore aveva adottato specifiche misure tecniche in grado di rendere palese l'intenzione di proteggere la propria opera da attività di tal genere. Dalla sentenza *Landgericht Hamburg*, AZ.:310 ● 227/23, cit., 8: « Auf der Webseite der Bildagentur ... befand sich mindestens seit dem 13.01.2021 auf der Unterseite <https://www...com/de/usage.html> folgender Text: „RESTRICTIONS YOU MAY NOT: (...) Use automated programs, ap-plets, bots or the like to access the ...com website or any content thereon for any purpose, including, by way of example only, downloading Content, indexing, Scraping or caching any content on the website” ».

³⁵ Sottolineato anche dal Giudice della sentenza in commento per evidenziare il

differente carattere lesivo delle due attività con riguardo ai diritti di riproduzione autorali. Dalla sentenza *Landgericht Hamburg*, AZ.:310 O 227/23, cit., 77-81: « Soweit ergänzend angeführt wird, dass es beim „KI-Web scraping“ um den geistigen Inhalt der zu Trainingszwecken genutzten Werke und „letztlich“ um die Schaffung inhaltsgleicher oder ähnlicher Konkurrenzsergebnisse gehe (Schack, a.a.O.), unterscheidet diese Argumentation nach Auffassung der Kammer nicht streng genug zwischen: — zum einen der (hier allein streitgegenständlichen) Erstellung eines — auch — für KI-Training nutzbaren Datensatzes, — zum anderem dem nachfolgenden Training des künstlichen neuronalen Netzes mit diesem Datensatz und — zum dritten der nachfolgenden Nutzung der trainierten KI zum Zwecke der Erstellung neuer Bildinhalte. Diese letztere Funktionalität mag zwar bereits bei der Erstellung des Trainingsdatensatzes angestrebt sein. Jedoch ist im Zeitpunkt der Zusammenstellung des Trainingsdatensatzes weder absehbar, in welcher Weise der zweite Schritt (das Training) erfolgreich sein wird, noch, welche konkreten Inhalte im dritten Schritt (bei der KI-Anwendung) durch die trainierte KI werden generiert werden können ».

³⁶ Sulla configurabilità del *dataset* quale banca di dati tutelabile dal diritto sui generis, vi è chi ritiene che l'impiego di *software* in grado di ricercare, indicizzare ed organizzare i contenuti raccolti in autonomia si ponga al di fuori dell'ambito di applicazione dell'art. 7 Dir. (UE) 96/9/CE alla luce della dicotomia esistente tra l'attività di "raccolta" e l'attività di "generazione" di dati, compiuta in prima istanza dalla Corte di Giustizia Europea nell'ambito delle decisioni 9 novembre 2004 *British Horseracing Board v William Hill Organization*, C-203/02, in ECR, 2004, I, 10415 ss. e *Fixtures Marketing v Svenska Spel*, C-338/02, in ECR, 2004, I, 10497. Si veda sul punto M. LEISTNER, *Big Data and the EU Database Directive 96/9/EC: Current Law and Potential for Reform*, in S. Lohse, R. Schulze, D. Staudenmayer (a cura di) *Trading Data in the Digital Economy: Legal*

utili a generare autonomamente nuovi contenuti simili a quelli presenti nel *dataset* originario³⁷. Mentre, dunque, l'allenamento di modelli di IA pare possa essere ricondotto all'interno della definizione di « estrazione di testo e dati » di cui alla Direttiva proprio alla luce della rielaborazione di ampie quantità di contenuti digitali per la generazione di nuove conoscenze³⁸, così non parrebbe per la creazione di *dataset*, la quale è limitata dall'estrazione (senza elaborazione) di dati già presenti nei materiali riprodotti³⁹.

5. L'ATTIVITÀ DI COSTRUZIONE DEL *DATASET* DI ALLENAMENTO COME ATTO DI RIPRODUZIONE DELL'OPERA DELL'INGEGNO.

La difficile riconducibilità della costruzione di un *dataset* di allenamento alla definizione normativa di *text and data mining* impone valutare se l'attività in esame integri una forma di sfruttamento delle opere dell'ingegno nell'ambiente digitale, come tale riconducibile ai diritti di privativa autorale. Sul punto, pare non potersi prescindere dal considerare l'aspetto teleologico sotteso a tali *dataset*. Come osservato anche dal giudice di Amburgo⁴⁰, la creazione di questi insiemi di dati è strettamente finalizzata all'allenamento di modelli generativi⁴¹, al punto che la loro struttura tipica consente il riutilizzo dei dati ivi contenuti proprio per

Concepts and Tools, Nomos, Baden, 2017, 25 ss. In senso contrario, si ritiene che l'investimento consistente nella scelta e nell'acquisto di una licenza *software* in grado di verificare e presentare i contenuti sia elemento sufficiente per il riconoscimento di tutela del diritto *sui generis* ai *dataset*. Cfr. G. NOTO LA DIEGA, *Artificial Intelligence and databases in the age of big machine data*, in *AIDA*, 2018, 93 ss.

³⁷ Sulle modalità di allenamento di un modello di Intelligenza Artificiale, ampiamente in A. BURKOV, *Machine Learning Engineering*, Londra, 2020, 13 ss.

³⁸ L'art. 53 Reg. (UE) 2024/1689 è limitato ai fornitori di modelli di IA generativa, prevedendo che i medesimi attuino una politica volta a rilevare le misure tecniche poste dai titolari di diritti d'autore a riserva dell'utilizzo delle proprie opere nelle fasi di sviluppo del modello intelligente. La medesima previsione non compare, invece, nell'ambito dell'art. 10 « Dati e governance di dati », il quale purtuttavia impone una serie di pratiche di *governance* in riferimento alla costruzione dei set di dati di addestramento.

³⁹ Il criterio di distinzione tra *text and data mining* e *data scraping* consistente nella presenza dell'elaborazione successiva dei dati raccolti potrebbe trovare conforto alla luce dell'interpretazione spesso restrittiva che la Corte di Giustizia Europea compie nell'ambito di applicazione delle eccezioni al diritto d'autore previste dall'ordi-

namento europeo. Cfr. Corte giust., *Commissione/Spagna*, C-36/05, 26 ottobre 2006, I-10313, pt. 31.

⁴⁰ Il quale, purtuttavia, giunge alla conclusione che proprio alla luce dell'interdipendenza tra costruzione di *dataset* e allenamento di reti neurali artificiali, escludere la prima attività dalla definizione di *text and data mining* condurrebbe a rendere inefficace l'ambito di applicazione delle eccezioni previste dalla Direttiva. *Landesgericht Hamburg*, AZ.:310 O 227/23, cit., 13: « Da eine Verwendung von im Wege des Text und Data Mining gewonnenen Erkenntnissen zum Trainieren Künstlicher neuronaler Netze, die dann in Konkurrenz zu Urhebern treten können, auf Grundlage der aktuellen technologischen Entwicklung im Zweifel nie ausgeschlossen werden kann, würde die Gegenauffassung in letzter Konsequenz sogar dazu nötigen, das Text und Data Mining i.S.d. § 44b UrhG letztlich in seiner Gänze zu untersagen ».

⁴¹ Sebbene infatti la branca scientifica dell'Intelligenza Artificiale esista dagli anni '50, l'ampia disponibilità su Internet di modelli *general purpose* è fenomeno recente, al punto da essere stato considerato in sede europea solo nelle ultime versioni della Proposta di Regolamento. Cfr. G. OLIVATO, *Spianando il percorso verso una regolamentazione dei sistemi di IA di scopo generale nella legge sull'IA: un'analisi delle proposte del Parlamento e del Consiglio*, in *Media-Laws*, 2023, n. 3, 50 ss. Nel recente caso

l'allenamento, la validazione o il *testing* di modelli di Intelligenza Artificiale ⁴².

Il profilo teleologico sotteso all'attività di *scraping* in esame suggerisce dunque di provare a ripensare i confini applicativi del diritto d'autore e dei diritti connessi, spesso considerati limitati al contenuto creativo dell'opera nella sua interezza e alle singole parti di esso fruibili dagli utenti ⁴³. Se, infatti, per il tramite dell'attività in commento le opere dell'ingegno sono "veicolate" con il precipuo scopo di rendere utilizzabile la forma espressiva, pare utile valutare se i diritti di privativa autorale possano estendersi anche alle fasi successive all'atto di riproduzione classicamente inteso, quando, cioè l'opera è sottoposta all'analisi automatizzata per l'estrazione di dati ad essa riferibili. Il tema riguarderebbe, in estrema sintesi, la persistenza o meno del carattere espressivo delle opere nei dati risultato dell'analisi automatizzata. Qualora si riconosca la stessa forma espressiva nel formato "datificato", infatti, l'attività di generazione della serie di dati si concretizzerebbe in un ulteriore atto di riproduzione dell'opera dell'ingegno, come tale passibile di essere soggetto a diritti di privativa autorale ⁴⁴.

Prima facie, pare tuttavia complesso immaginare che i singoli dati componenti l'opera siano suscettibili di ricadere nell'alveo dei diritti

Thomson Reuters enterprise Centre GmbH and West Publishing Corp. vs. Ross Intelligence Inc. affrontato dalla Corte del Delaware, l'esclusione del *fair use* all'utilizzo di opere dell'ingegno per allenamento di modelli intelligenti è stato valutato anche alla luce del probabile effetto dell'uso dei dati, senza il consenso del titolare, sul mercato dell'opera originaria. Sul punto, il giudice ha valutato non solo l'impatto sul mercato attuale, ma anche su quelli potenziali in cui lo sfruttamento dell'opera dell'ingegno potrebbe inserirsi in futuro.

⁴² È altresì noto che i modelli maggiormente performativi sul piano della generazione di contenuti multimediali poggiano su *dataset* costituiti a partire da dati associati ad opere creative. Sul punto, A. BUICK, *Copyright and AI training data—transparency to the rescue?*, in *Journal of Intellectual Property Law & Practice*, 2025, n. 20-3, 184 ss.: « The case of GPT-3, the breakthrough Large Language Model (LLM) announced by OpenAI in 2020, provides an illustration. In the initial paper describing the development and capabilities of GPT-3, the authors revealed that the data the model had been trained upon consisted of a refined version of the Common Crawl dataset, a smaller dataset of higher-quality web-based text called WebText2, two datasets of books and the English-language version of Wikipedia ». Secondo il recente rapporto del Centre for Research on Foundation Models dell'Università di Stanford, le maggiori imprese sviluppatrici di modelli generativi ten-

dono a mantenere opachi l'origine e la legittimità dei dati impiegati nell'allenamento. Cfr. BOMMASSANI ET AL., *The Foundation Model Transparency Index v1.1*, in *Transaction on Machine Learning Research*, 2024, 13 ss.

⁴³ Sui nuovi bilanciamenti tra diritti autorali e "diritti" degli utenti nell'ambiente digitale, ampiamente in G. FINOCCHIARO, *L'equilibrio titolare/users nel diritto d'autore dell'Unione Europea*, cit.

⁴⁴ Art. 2, Dir. 2001/29/CE del Parlamento europeo e del Consiglio, del 22 maggio 2001, sull'armonizzazione di taluni aspetti del diritto d'autore e dei diritti connessi nella società dell'informazione: « Gli Stati membri riconoscono ai soggetti sotto elencati 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: a) agli autori, per quanto riguarda le loro opere; b) agli artisti interpreti o esecutori, per quanto riguarda le fissazioni delle loro prestazioni artistiche; c) ai produttori di fonogrammi per quanto riguarda le loro riproduzioni fonografiche; d) ai produttori delle prime fissazioni di una pellicola, per quanto riguarda l'originale e le copie delle loro pellicole; e) agli organismi di diffusione radio-televisiva, per quanto riguarda le fissazioni delle loro trasmissioni, siano esse effettuate su filo o via etere, comprese le trasmissioni via cavo o via satellite ».

Ove le opere siano direttamente impiegate nell'allenamento di modelli generativi,

d'autore⁴⁵. Tale assunto, peraltro accolto anche dal giudice di Amburgo⁴⁶, si fonda sul presupposto che il dato, consistendo in un mero elemento informativo e, pertanto, fattuale, non possa beneficiare della tutela accordata invece ai contenuti che presentano elementi espressivi ed originali di un'idea e sono percepibili come tali all'occhio umano⁴⁷. Eppure, con riferimento all'attività in commento, l'apprezzamento umano sulle coppie di testo-immagine risultato dell'analisi automatizzata rappresenta un momento rilevante ai fini della corretta costruzione dell'insieme di dati⁴⁸. Lo sviluppatore, infatti, interviene nell'ambito dell'aderenza delle serie alle relative immagini analizzate, ricorrendo a specifici *software* in grado di inserire all'interno del *dataset* solo i dati che correttamente individuano il contenuto creativo dell'immagine analizzata⁴⁹. Tale intervento è fondamentale per la costruzione di un insieme di dati utile a svolgere la sua funzione: ove il *software* non confermi che l'opera è stata

è già prassi acquisita dalle maggiori aziende nel settore concludere veri e propri contratti di licenza per il relativo sfruttamento. Si richiamano gli accordi conclusi tra OpenAI e Financial Times per l'utilizzo dei dati per l'allenamento del noto ChatGPT, ovvero quelli conclusi tra Google e il social network Reddit. Cfr. OECD, *Intellectual property issues in AI trained on scraped data*, Febbraio 2025, 35 ss.

⁴⁵ Oltre ai profili critici emergenti dalla tutelabilità del diritto d'autore, la configurabilità di un "diritto a disporre" dei dati contenuti nell'opera in capo all'autore condurrebbe a considerare tali informazioni quali beni giuridici suscettibili *ex se* di tutela "proprietary", un tema attualmente ampiamente dibattuto in dottrina. Cfr. P. PERLINGERI, *L'informazione come bene giuridico*, in *Rassegna di diritto civile*, 1990, n. 2, 326 ss., il quale considera l'informazione quale un bene in senso giuridico, suscettibile quantomeno di una "disponibilità qualificata"; G. FINOCCHIARO, *Consenso al trattamento e libertà*, in *Persona e Mercato*, 2024, n. 1, 21 ss. cala invece le considerazioni sulla natura patrimoniale dell'informazione al caso specifico del dato personale, osservando sul punto che il confine tra patrimonialità e non patrimonialità del dato riguardante l'interessato si è fatto attualmente più labile. *Contra* D. MESSINETTI, *Oggettività giuridica delle cose incorporali*, Milano, 1970, 197 ss.; R. PARDOLESI, C. MORTI, « L'idea è mia! »: *lusinghe e misfatti dell'Economics of Information*, in *Diritto dell'informatica*, 1990, n. 2, 345 ss., i quali sostengono che l'informazione, avente funzioni prismatiche, non possa risultare suscettibile di privativa in quanto mancherebbe lo *ius excludendi alios* e A. ALTIERI, *La "dimensione organizzativa" dei sistemi di intelligenza artificiale*, in *AIDA*, 2023, 509 ss. il quale cala il dibattito dottrinale

sulla difficile configurabilità dell'AI come "bene informativo".

⁴⁶ *Landsgericht Hamburg, AZ.:310 O 227/23*, cit., 13: « Die vorliegend urheberrechtlich relevante Vervielfältigung ist auf den Zweck der Analyse der Bilddateien auf ihre Übereinstimmung mit einer vorbestehenden Bildbeschreibung nebst anschließender Einstellung in einen Datensatz beschränkt. Dass durch diese Nutzung die Verwertungsmöglichkeiten der jeweils betroffenen Werke beeinträchtigt werden würden, ist nicht ersichtlich und wird auch klägerseits nicht behauptet ».

⁴⁷ Sul bilanciamento tra la circolazione dei dati per lo sviluppo dell'economia digitale e l'esigenza di tutelare le prerogative autoriali ampiamente in C. SCANGA, *Limiti e limitazioni al diritto d'autore nel nuovo EU Data Package: alla ricerca del sistema perduto*, in *AIDA*, 2023, 338 ss.

⁴⁸ Calate le considerazioni sul punto della Direttiva 2019/790, il legislatore europeo ha in effetti mancato di prevedere una norma *ad hoc* in relazione alla titolarità dei risultati ottenuti tramite l'attività estrattiva. Solleva il tema M. GRANIERI, *Il data mining nella disciplina del diritto d'autore e la strategia europea dei dati*, cit., 40 ss., il quale constata che, qualora i risultati ottenuti tramite TDM manchino dei requisiti minimi per l'accesso alla tutela derivante dalla proprietà intellettuale, tali informazioni dovrebbero considerarsi di pubblico dominio, al netto di attività realizzate dal *miner* in grado di impedirne la diffusione.

⁴⁹ Nell'allenamento, il modello si riferirà quindi all'opera — o a più riproduzioni della medesima opera — quando quella serie di dati ad essa associata è presente nell'*input*. Questo, si suppone, in assenza di ulteriori limiti posti alla macchina durante l'allenamento. È ormai prassi dei maggiori

correttamente individuata nei suoi tratti identificativi, la serie di dati sarebbe sostanzialmente inutilizzabile nell'eventuale allenamento del modello generativo⁵⁰. Al contrario, l'individuazione del preciso contenuto creativo da parte del *software* per il tramite delle informazioni associate consente di massimizzare lo sfruttamento dell'opera dell'ingegno nell'allenamento⁵¹. La serie di dati generata sarebbe pertanto in grado di garantire che l'opera di riferimento rimanga « identificabile con sufficiente precisione ed obiettività »⁵² conservando la sua « funzione *latu sensu* espressiva » anche da parte dei fornitori di modelli generativi nel momento della scelta dei *set* di dati di addestramento. Più la descrizione è vicina al contenuto creativo dell'immagine analizzata, maggiori sono infatti le probabilità che i modelli intelligenti allenati con quel determinato insieme di dati siano in grado di generare *output* facenti riferimento a quel determinato contenuto creativo.

Calate tali considerazioni tecniche sul punto della tutelabilità autoriale sui dati derivanti dalle opere dell'ingegno, si potrebbe dunque prospettare che ove consenta la generazione di dati in concreto veicolanti la forma espressiva dell'opera, l'attività di *scraping* consista nella riproduzione in un formato diverso del medesimo carattere creativo e rientri, perciò, nell'alveo di tutela dei diritti connessi⁵³. In effetti, ripercorrendo la giurisprudenza europea nell'ambito dell'attività di riproduzione di

fornitori di modelli di IA generativa bloccare determinati *input* che possano generare *output* riferibili a opere dell'ingegno.

⁵⁰ Sempre C. SCHÜHMANN ET AL., *LAION-5B: An open large-scale dataset for training next generation image-text models*, cit., 9: « Birhane et al. described the shortcomings of alt-text and noted that alt-text is not necessarily a good description of the corresponding image. For instance, the alt-text may be search engine optimization (SEO) spam, an incoherent list of keywords, or overly corrupted otherwise. In such cases, the language in the text annotations may become less informative or entirely useless for training ». Pare ovvio sottolineare che la correttezza degli alt-text rispetto alle opere analizzate non sia l'unico parametro impiegato nell'allenamento dei modelli intelligenti; purtroppo, la sua aderenza al contenuto creativo di origine consente di « facilitare » la generazione di contenuti da esso derivati da parte del modello in allenamento.

⁵¹ Rare sono le decisioni giurisprudenziali sul punto della legittimità dell'attività di *scraping* per costruzione di insiemi di dati. Si ricorda Trib. Roma, Sez. Spec. Impresa, sent. 5 maggio 2019, in *AIDA*, 2020, 794 ss., ove i giudici hanno stabilito la legittimità dell'attività di *screen scraping* e successivo reimpiego di informazioni da una banca dati pubblicamente disponibile, a condizione che le informazioni raccolte non rappresentassero una parte sostanziale e

purché non vi fosse stato pregiudizio ingiustificato al suo costituente. Nel caso affrontato dal Tribunale di Amburgo, si argomenta a supporto dell'assenza di pregiudizi agli interessi dell'autore la presenza su Internet del noto CommonCrawl.org, attivo dal 2008, la cui costruzione è stata basata ampiamente su materiale creativo presente online. Più recentemente, *United States District Court for the District of Delaware*, sent. 11 febbraio 2025, *Thomson Reuters enterprise Centre GmbH and West Publishing Corp. vs. Ross Intelligence Inc.*, ha tuttavia escluso l'applicazione del *fair use* per l'attività di riproduzione dei dati contenuti in una banca dati giuridica quando tale riproduzione sia teleologicamente correlata all'allenamento di sistemi intelligenti venduti sullo stesso mercato.

⁵² Corte EU, C-310/17, *Levola Hengelo*, cit.

⁵³ La nozione di « opera dell'ingegno » non trova una sua definizione nell'ordinamento italiano, il quale si limita, piuttosto, ad individuare i confini applicativi della legge sul diritto d'autore. Ai sensi dell'art. 1 l.a., infatti, possono godere di tutela « le opere dell'ingegno di carattere creativo che appartengono alla letteratura, alla musica, alle arti figurative, all'architettura, al teatro ed alla cinematografia, qualunque ne sia il modo o la forma di espressione. Sono altresì protetti i programmi per elaboratore come opere letterarie ai sensi della Convenzione di Berna sulla protezione delle opere

un'opera dell'ingegno, pare utile evidenziare che la stessa Corte abbia riconosciuto un'ampia portata definitoria al diritto di riproduzione, anche parziale, di un'opera dell'ingegno, essendo sufficiente per il suo riconoscimento che, da un lato, gli elementi ripresi nell'atto di riproduzione siano « l'espressione della creazione intellettuale del loro autore »⁵⁴, e, dall'altro, che ove sia condotta nell'ambito di un procedimento tecnico, la generazione di copie sia funzionale per l'accesso alle opere protette da parte degli utenti⁵⁵. Sul punto, pare che la serie di dati risultato di analisi si presti ad essere riconosciuta quale atto di riproduzione non temporanea⁵⁶ nel momento in cui consente di replicare l'opera individuata tramite *link* e descrizione per avere accesso alla sua forma

letterarie ed artistiche ratificata e resa esecutiva con legge 20 giugno 1978, n. 399, nonché le banche di dati che per la scelta o la disposizione del materiale costituiscono una creazione intellettuale dell'autore ». Pur in assenza di una compiuta definizione, si tende in ogni caso a considerare che la nozione di opera non costituisca una clausola aperta, ma trovi nell'esistenza di una forma espressiva il suo principale confine. La dottrina, sul punto, si è divisa nella configurabilità di ulteriori criteri che contribuiscono a circoscrivere la definizione di "opera dell'ingegno". Cfr. T. ASCARELLI, *Teoria della concorrenza e dei beni immateriali*, Milano, 1960, 698 ss.; G. SPEDICATO, *Interesse pubblico e bilanciamento nel diritto d'autore*, cit., 157 i quali propendono per individuare quale oggetto di protezione del diritto d'autore non la forma espressiva *ex se*, ma solo l'uso di tale forma per esprimere un determinato contenuto; di visione differente P. GRECO, P. VERCELLONE, *I diritti sulle opere dell'ingegno*, in *Trattato Vassalli*, Torino, 1974, 42 ss., i quali considerano tutelata dal diritto d'autore unicamente la sua forma espressiva. Alla luce di tale vuoto definitorio, la giurisprudenza europea è intervenuta in più occasioni ad individuare le caratteristiche comuni a tutti i beni immateriali qualificabili come "opere dell'ingegno". Cfr. Corte giust., C-5/08, *Infopaq v. DDF*, 16 luglio 2009, in *AIDA*, 2009, 428 ss., e in *Diritto d'autore*, 2009, 640 ss., con nota di G. MARI, *Rassegne stampa digitali e 11 parole d'autore*, ove è stato riconosciuto che anche una minima serie di parole in sequenza può integrare la forma espressiva identificante la personalità dell'autore; Corte giust., C-310/17, *Levola Hengelo*, 13 novembre 2018, in *AIDA*, 2019, 1879 ss., ove la Corte ha inteso individuare quale elemento caratterizzante l'opera dell'ingegno il requisito dell'espressione, intesa come capacità del contenuto creativo di risultare identificabile con sufficiente obiettività, poi ulteriormente ripreso in Corte giust., C-476/17, *Pelham GmbH v. Hütter*,

29 luglio 2019; Corte giust., C-833/18, *SI, Brompton Bicycle Ltd. c. Chedech/Get2Get*, 11 giugno 2020; Corte giust., C-683/17, *Cofemel - Sociedade de Vestuário SA c. G-Star Raw CV*, 12 settembre 2019.

⁵⁴ Questa da valutarsi caso per caso. Cfr. Corte giust., C-5/08, *Infopaq v. DDF*, cit.

⁵⁵ Corte giust. C-403/08, *Football Association c. QC Leisure*, 4 ottobre 2011. Anche nel noto caso Corte giust. C-5/08, *Infopaq v. DDF*, cit., è stato considerato quale riproduzione la conversione di un file in una serie di dati unicamente leggibile da un computer. In entrambi i casi affrontati dalla Corte l'attività di riproduzione rientrava nell'eccezione di cui all'art. 5 par. 1 Dir. 2001/29/CE in quanto considerata di carattere accessorio e temporaneo. In dottrina, cfr. A. OTTOLIA, *L'utilizzo computazionale dell'opera dell'ingegno in internet*, cit., 392: « Seguendo una metodologia interpretativa che valorizzi l'effetto complessivo realizzato attraverso un processo tecnologico rispetto alla modalità attuativa adottata sul piano tecnico, non pare si dovrebbe poter discriminare tra atti di riproduzione unitaria e contestuale ed atti di riproduzione progressiva realizzati su singole porzioni non tutelabili dell'opera dell'ingegno; così ad esempio la riproduzione di un testo attraverso la riproduzione progressiva di singole parole di per sé non tutelabili dovrebbe ritenersi in conflitto con il diritto del titolare ».

⁵⁶ Ai sensi dell'art. 5, par. 1 Dir. 2001/29/CE, sono esentati dal diritto di riproduzione di cui all'articolo 2 gli atti di riproduzione temporanea privi di rilievo economico proprio che sono transitori o accessori, e parte integrante ed essenziale di un procedimento tecnologico, eseguiti all'unico scopo di consentire la trasmissione in rete tra terzi con l'intervento di un intermediario, ovvero un utilizzo legittimo di un'opera o di altri materiali. Nel caso in cui lo *scraping* insista su un'opera dell'ingegno, pare dubbio che tale atto possa considerarsi temporaneo od accessorio alla luce dell'obiettivo di costru-

espressiva tramite l'utilizzo di modelli generativi⁵⁷. Se dunque a mutare è solo il “supporto” — in senso atecnico — con cui le singole opere si presentano nel *dataset*, non pare potersi escludere *a priori* la possibilità di configurare tale attività quale atto di riproduzione in forma “datificata” dell'opera dell'ingegno.

6. I DATI COME RIPRODUZIONI DELL'OPERA DELL'INGEGNO: SPUNTI DAL REGOLAMENTO (UE) 2022/868 SULLA GOVERNANCE DEI DATI.

La qualificazione dell'attività di *scraping* per la costruzione di *dataset* di allenamento quale atto di riproduzione di opere dell'ingegno pare trovare conferma alla luce delle recenti novità previste dal legislatore europeo nell'ambito della circolazione dei dati personali e non personali nell'ambiente digitale. Ci si riferisce, in particolare, al Regolamento (UE) 2022/868 del Parlamento europeo e del Consiglio del 30 maggio 2022 relativo alla governance europea dei dati (Regolamento sulla governance dei dati o *Data Governance Act*)⁵⁸, il quale definisce un quadro normativo di *governance* nella circolazione dei dati personali e non personali tra gli enti pubblici che hanno il diritto o l'obbligo di concedere l'accesso a tali dati e i soggetti interessati al relativo riutilizzo anche a scopi commerciali⁵⁹.

Il *Data Governance Act* si inserisce nell'ambito della più ampia *European Strategy for Data*⁶⁰ con l'obiettivo di rendere disponibili i dati detenuti dagli enti pubblici pur mantenendo tutelati i diritti dei terzi su essi insistenti. Calato nell'ambito del diritto d'autore, il Reg. 2022/868

zione di un *dataset* che mantiene conservati i dati ad essa riferibili e di frequente messi altresì a disposizione.

⁵⁷ G. SPEDICATO, *Interesse pubblico e bilanciamento nel diritto d'autore*, Milano, 2013, 148: « Nella mera riproduzione non si dà, infatti, creazione in senso stretto, bensì replica dell'esistente, e per conseguenza chi riproduce *sic et simpliciter* non è autore ».

⁵⁸ Regolamento (UE) 2022/868 del Parlamento Europeo e del Consiglio del 30 maggio 2022 relativo alla governance europea dei dati e che modifica il regolamento (UE) 2018/1724, «Regolamento sulla governance dei dati».

⁵⁹ Sul *Data Governance Act*, più ampiamente in F. BRAVO, *Il principio di solidarietà tra “data protection” e “data governance”*, in questa Rivista, 2023, 481 ss.; E. CREMONA, *Quando i dati diventano beni comuni: modelli di “data sharing” e prospettive di riuso*, in *Rivista italiana di informatica e diritto*, 2023, n. 2, 111 ss.; D. POLETTI, *Il controllo dell'interessato e la strategia europea sui dati*, in *Osservatorio sulle fonti*, 2023, n. 2, 367 ss.; A. MORACE PINELLI, *Dalla “data protection” alla “data governance”*: il regolamento UE 2022/868, in *La Nuova Giurisprudenza Civile Commentata*, 2024,

n. 2, 486 ss.; F. BRAVO, *Intermediazione di dati personali e servizi di “data sharing” dal GDPR al “Data Governance Act”*, in *Contratto e Impresa Europa*, 2021, n. 1, 199 ss.; A. VICORITO, *Sul crinale tra “data altruism” e “social scoring”: esperienze applicative della sequenza dati-algoritmi nel nuovo contesto regolatorio europeo*, in *MediaLaws*, 2023, n. 1, 104 ss.

⁶⁰ La Strategia Europea per i Dati presentata dalla Commissione Europea a Parlamento europeo e Consiglio nel febbraio 2020 si è articolata in un piano quinquennale volto a creare spazi europei comuni sui dati, proponendosi di superare l'esistente frammentazione normativa a livello nazionale mediante l'adozione di una serie di Regolamenti tesi ad incentivare la condivisione dei dati. Tra le diverse misure adottate oltre al *Data Governance Act*, si annovera il Reg. (UE) 2023/2854 del 13 dicembre 2023 riguardante norme armonizzate sull'accesso equo ai dati e sul loro utilizzo, avente lo scopo di regolare l'accesso ai dati, spesso non personali, generati dall'uso dei cd. « prodotti connessi » e dai servizi digitali ad essi correlati e il Reg. (EU) 2025/327 dell'11 febbraio 2025, sullo Spazio Europeo dei Dati Sanitari, «*European Health Data Space*».

prevede una serie di indicazioni volte ad ampliare il novero di dati ⁶¹ e di soggetti titolati ad averne accesso ⁶², al contempo riconoscendo la possibilità che sui medesimi insistano diritti di proprietà intellettuale. Il nuovo approccio europeo pare emergere in proposito a partire dalla definizione di « dato », il quale passa da essere considerato sinonimo di “documento” ⁶³ od “informazione” ⁶⁴ ad avere dignità propria quale « rappresentazione digitale di atti, fatti o informazioni e qualsiasi raccolta di tali atti, fatti o informazioni, anche sotto forma di registrazione sonora, visiva o audiovisiva » ⁶⁵. La nuova definizione, sicuramente di ampissima portata, riconosce quale potenziale oggetto di circolazione nel mercato dei dati tutti i contenuti sulla Rete, inclusi i contenuti creativi ⁶⁶. Più specificamente, apre alla possibilità di riconoscere al dato non solo il suo noto valore informativo, ma anche di attribuire alla rappresentazione digitale lo stesso valore volta per volta riconducibile all’atto, fatto o informazione ad essa connessa.

Tale assunto pare essere confermato nel momento in cui il DGA mette in relazione il dato con i diritti di proprietà intellettuale all’art. 5, par. 3). La disposizione, finalizzata a regolare i rapporti di circolazione di dati detenuti a vario titolo dalla pubblica amministrazione, indica chiaramente che i dati sono protetti da diritti di proprietà intellettuale ove i contenuti digitali da cui traggono origine ricadano nella medesima tutela. Il par. 3) obbliga infatti gli enti pubblici a dare accesso agli utenti solo ove i dati protetti da diritti di proprietà intellettuale siano stati previamente modificati, aggregati o trattati mediante metodi di controllo della divulgazione ⁶⁷, mentre il par. 7) autorizza la circolazione “in chiaro” solo ove

⁶¹ Art. 3, par. 1, Reg. (UE) 2022/868: « Il presente capo si applica ai dati detenuti da enti pubblici, che sono protetti per motivi di: a) riservatezza commerciale, compresi i segreti commerciali, professionali o d’impresa; b) riservatezza statistica; c) protezione dei diritti di proprietà intellettuale di terzi; o d) protezione dei dati personali, nella misura in cui tali dati non rientrano nell’ambito di applicazione della direttiva (UE) 2019/1024 ».

⁶² Art. 2, n. 9), Reg. (UE) 2022/868: « « utente dei dati »: una persona fisica o giuridica che ha accesso legittimo a determinati dati personali o non personali e che ha diritto, anche a norma del regolamento (UE) 2016/679 in caso di dati personali, a utilizzare tali dati a fini commerciali o non commerciali ».

⁶³ Nell’ambito delle prime normative europee riferibili al promovimento della circolazione dei dati, la Direttiva (UE) 2019/1024 del Parlamento europeo e del Consiglio, del 20 giugno 2019 relativa all’apertura dei dati e al riutilizzo dell’informazione del settore pubblico riconduce il « dato » al « documento in formato digitale ».

⁶⁴ Il noto Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, del

27 aprile 2016 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE, « Regolamento generale sulla protezione dei dati » definisce all’art. 4, n. 1) il « dato personale » quale « qualsiasi informazione riguardante una persona fisica identificata o identificabile ».

⁶⁵ Sulle luci ed ombre derivanti dal pacchetto di riforme europeo riguardante la circolazione dei dati e il relativo rapporto con il diritto d’autore, ampiamente in C. SCANGA, *Limiti e limitazioni al diritto d’autore nel nuovo EU Data Package: alla ricerca del sistema perduto*, cit., 392.

⁶⁶ Con ciò discostandosi dagli orientamenti più tradizionali che consideravano i dati avulsi dall’applicazione della normativa sul diritto d’autore. Cfr. R. PARDOLESI, C. MOTTI, « *L’idea è mia!* »: *lusinghe e misfatti dell’Economics of Information*, cit.; R. DUCATO, A. M. STROWEL, *Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to Machine Legibility*, in *CRIDES Working Paper Series*, 2018.

⁶⁷ Art. 5 par. 3), Reg. (UE) 2022/868: « Gli enti pubblici garantiscono, conformemente al diritto dell’Unione e nazionale, la

il riutilizzatore si impegni contrattualmente a rispettare i diritti di proprietà intellettuale gravanti sui medesimi⁶⁸. Pur dunque aprendo alla disponibilità di riutilizzare un novero più ampio di dati ai fini di promuovere l'innovazione tecnologica, il dato rappresentativo di una o di una parte dell'opera dell'ingegno sembra essere riconosciuto nel suo potenziale di "veicolo" della forma espressiva ad esso sottesa. Il *Data Governance Act* richiede infatti l'utilizzo di tecniche di aggregazione volte a ridurre il dato direttamente a mero elemento informativo prima della sua circolazione nel mercato⁶⁹, in modo da impedire al potenziale riutilizzatore di sfruttare il contenuto creativo ad esso collegato.

Ancorché le norme esaminate attengano unicamente ai rapporti di circolazione dei dati tra settore pubblico e privato, dalla loro analisi emergono alcune brevi considerazioni sull'attività di *scraping* in commento. Poiché il legislatore europeo parrebbe abbracciare una definizione di « dato » onnicomprensiva⁷⁰, anche la serie di dati risultato di *scraping* è riconducibile a tale definizione e, a determinate condizioni, è suscettibile di essere ricondotta all'interno dei diritti di proprietà intellettuale. ●ve, infatti, la serie di dati mantenga la funzione di « rappresentare » l'opera originaria da cui trae origine, pare complesso immaginare la sua esclusione dal novero dei « dati protetti da diritti di proprietà intellettuale »⁷¹. Mentre, dunque, il suo riutilizzo potrebbe considerarsi legittimo da parte del riutilizzatore solo ove questi si obblighi contrattualmente a rispettare i diritti di proprietà intellettuale insistenti sul medesimo, la serie di dati potrà dirsi esclusa dal diritto d'autore solo ove non consenta di risalire all'opera dell'ingegno da cui trae origine. Dall'inclusione nel novero dei

tutela della natura protetta dei dati. Essi garantiscono il rispetto dei requisiti seguenti: a) concedere l'accesso per il riutilizzo dei dati soltanto qualora l'ente pubblico o l'organismo competente abbia garantito, in seguito alla richiesta di riutilizzo, che i dati sono stati: i) anonimizzati, nel caso di dati personali; e ii) modificati, aggregati o trattati mediante qualsiasi altro metodo di controllo della divulgazione, nel caso di informazioni commerciali riservate, compresi i segreti commerciali o i contenuti protetti da diritti di proprietà intellettuale; b) accedere ai dati e riutilizzare gli stessi da remoto all'interno di un ambiente di trattamento sicuro, fornito o controllato dall'ente pubblico; c) accedere ai dati e riutilizzare gli stessi all'interno dei locali fisici in cui si trova l'ambiente di trattamento sicuro, rispettando rigorose norme di sicurezza, a condizione che l'accesso remoto non possa essere consentito senza compromettere i diritti e gli interessi di terzi ».

⁶⁸ Art. 5, par. 7), Reg. (UE) 2022/868: « Il riutilizzo dei dati è consentito solo nel rispetto dei diritti di proprietà intellettuale. Il diritto del costituente di una banca di dati di cui all'articolo 7, paragrafo 1, della direttiva 96/9/CE non è esercitato dagli enti pubblici al fine di impedire il riutilizzo dei

dati o di limitarlo oltre i limiti stabiliti dal presente regolamento ». Sui limiti e le opportunità degli obblighi contrattuali quali rimedi alternativi o aggiuntivi per la tutela della proprietà intellettuale cfr. M. W. MONTEROSSO, *Estrazione e (ri)utilizzo di informazioni digitali all'interno della rete internet. Il fenomeno del c.d. web scraping*, in questa Rivista, 2020, 327 ss.

⁶⁹ Pare utile evidenziare che l'esigenza di tutela dei diritti di proprietà intellettuale emerge nonostante il *Data Governance Act* si proponga di superare le barriere precedentemente poste dalla Direttiva (UE) 2019/1024, la quale esclude dall'ambito di applicazione i contenuti protetti dal diritto d'autore. L'approccio europeo rimane dunque caratterizzato da una forte protezione dei diritti di proprietà intellettuale e, più in generale, dalla particolare attenzione a non creare effetti distorsivi nel mercato digitale. Sul punto, E. CREMONA, *Quando i dati diventano beni comuni: modelli di data sharing e prospettive di riuso*, in Rivista italiana di informatica e diritto, 2024, 2, 112 ss.

⁷⁰ La stessa definizione è presente all'art. 2, n. 1) *Data Act*, ed è richiamata dall'art. 2, lett. c), *EHDS*.

⁷¹ Cons. 18), Reg. (UE) 2022/868; Art. 5, par. 10), Reg. (UE) 2022/868.

« dati protetti da diritti di proprietà intellettuale » emerge che lo *scraping* volto alla loro creazione, consistendo in un'attività ove il contenuto dell'opera dell'ingegno è rappresentato tramite il *link* e *alt-text*, consista in uno sfruttamento economico assimilabile all'atto di riproduzione. Non potendo infatti escludersi *in nuce* che, ove compiuta per la creazione di *dataset* di allenamento, lo *scraping* non raccolga dati direttamente riferiti ai contenuti creativi, ne deriva che l'attività in commento si configura quale riproduzione delle opere dell'ingegno in un formato differente a fini di sfruttamento del loro valore creativo. Conseguentemente, la sua legittimità dipenderà dalla sussistenza di condizioni specifiche in grado di farlo ricadere nell'alveo delle eccezioni previste dalla normativa unionale, ovvero dall'eventuale consenso espresso dell'autore. La qualificazione dello *scraping* quale attività di riproduzione parrebbe, in ultima istanza, trovare terreno fertile nell'ambito della circolazione dei dati medesimi, ben potendo il risultato dell'attività essere suscettibile di privativa autorale ove mantenga un'associazione con i contenuti creativi di provenienza tale da rendere sfruttabile la forma espressiva ad essa sottesa.

7. CONSIDERAZIONI CONCLUSIVE.

La decisione del Tribunale di Amburgo ha consentito di muovere alcune critiche e di provare a valutare nuove prospettive con riguardo ai confini applicativi del diritto d'autore alla luce dell'evoluzione tecnologica attualmente in corso. Sul punto, l'inclusione nel regime del TDM dell'attività di *scraping* per la costruzione di *dataset* di allenamento consiste, ad avviso di chi scrive, in un'interpretazione particolarmente estensiva dell'ambito di applicazione della Direttiva 2019/790.

Accanto alle anzidette considerazioni sul diverso fine che contraddistingue le attività di *data scraping* e di *data mining*, pare utile evidenziare come l'estensione della prima al regime delle eccezioni previste per la seconda finirebbe per ribaltare il rapporto tra diritto d'autore e interesse degli utenti al godimento delle opere nell'ambiente digitale. Da un lato, l'autore si troverebbe nella posizione di dover "giustificare" l'interesse di tutela mediante l'adozione di misure tecniche in grado di impedire sia la messa a disposizione (come peraltro già avviene nell'ambito delle attività di *linking* e *framing*) sia gli atti di riproduzione della propria opera da parte di terzi nell'ambiente digitale. Dall'altro lato, l'inclusione dell'attività di *scraping* nell'eccezione di *text and data mining* consentirebbe ai fornitori di modelli di IA di beneficiare dell'eccezione concessa ai costruttori "no-profit" di *dataset* di allenamento per il riutilizzo delle opere dell'ingegno.

Considerato che nell'ambito dell'allenamento è complesso risalire al singolo dato che consente al modello di generare un *output* nello stile dell'autore, ben potendo quest'ultimo essere anche il risultato di diverse correlazioni tra dati, il fornitore di modelli generativi potrà accertarsi della presenza di una riserva espressa da parte dell'autore ai sensi dell'art. 53 AI Act solo riferendosi al *dataset* di allenamento. Purtuttavia, se il set di allenamento è reperito sulla Rete, l'inclusione della sua creazione nell'ambito della Direttiva implicherebbe che la raccolta di dati a finalità di ricerca scientifica senza scopi di lucro possa essere condotta a prescindere dal diritto di riserva dell'autore. Aggiunto alla ancora discussa qualificazione del dato quale oggetto di diritti di proprietà

intellettuale deriva la considerazione che al momento dell'impiego di set di dati di addestramento da parte dei fornitori di modelli di Intelligenza Artificiale nulla potrà essere eccepito da parte dell'autore per lo sfruttamento della sua opera. In altre parole, la riserva espressa dall'autore non consentirà di tutelare i propri diritti né al momento della raccolta dei dati a partire dalla propria opera in virtù dell'eccezione ex art. 3 DSM, né al momento di riutilizzo del *dataset* per finalità di allenamento, alla luce del complesso riconoscimento autorale sui dati ivi raccolti. Vi è da precisare, quindi, che l'unico caso in cui l'autore potrà esercitare la propria riserva in modo efficace sarà ove il ruolo di costituente di *dataset* e di fornitore di modelli generativi coincidano nella stessa persona che agisce per finalità commerciali.

D'altro canto, alla luce delle complessità tecniche dell'allenamento di modelli generativi spesso ricadenti nel fenomeno della *black box*, delle difficoltà inerenti all'applicazione del cd. *reverse engineering* su modelli allenati con dati riferiti alle opere dell'ingegno e della complessa configurazione di una violazione dei diritti autorali nell'ambito degli *output* generati dai modelli generativi (che ne riproducono lo stile ma non la forma espressiva), è chiaro che l'intervento per la tutela del diritto d'autore debba farsi a monte, al momento, cioè, in cui l'autore sia in grado ancora di esercitare un controllo sulla circolazione della propria opera. In una prospettiva di bilanciamento tra innovazione e tutela dei diritti nella nuova economia dell'Intelligenza Artificiale, dunque, la riconduzione dell'attività di *scraping* nelle logiche del classico diritto di riproduzione previsto dall'art. 5 Dir. 2001/29/CE consentirebbe di imputare in capo all'autore un adeguato riconoscimento, morale ed economico, laddove l'attività sia funzionale a veicolare la forma espressiva per un successivo riutilizzo. Una qualificazione più restrittiva dell'attività di *scraping* — intesa come forma di riproduzione in dati dell'opera — consentirebbe, in ultima istanza, di ristabilire un punto di equilibrio tra l'interesse al progresso tecnologico e la tutela della creatività individuale nell'ambito della circolazione dei dati nell'ambiente digitale.

FRANCESCA CONTE