Antonio Santangelo, Alberto Sissa,
Maurizio Borghi

Critica di ChatGPT

Prefazione di Juan Carlos De Martin
Postfazione di Marco Ricolfi

elèuthera

tutti i link segnalati nel testo sono stati consultati
entro il 31 dicembre 2024

progetto grafico di Riccardo Falcinelli

www.eleuthera.it
eleuthera@eleuthera.it

Indice

Prefazione

di *Juan Carlos De Martin*

Quando nel 2006 fu fondato il Centro Nexa su Internet e Società – al Politecnico di Torino, ma in stretta collaborazione con l'Università di quella città – l'idea di offrire non solo un luogo fisico, ma anche uno spazio online per ospitare chiunque fosse interessato a riflettere in maniera interdisciplinare sulla rivoluzione digitale era tra le principali priorità. Nexa, infatti, aveva l'ambizione di dare un contributo non solo locale, per quanto essenziale, ma anche a livello nazionale, tanto più che in quel periodo la classe dirigente italiana si distingueva, tranne rare eccezioni, per lo spiccato disinteresse che provava nei confronti delle tecnologie che già allora stavano cambiando il mondo. Come realizzare questo spazio online? Erano gli anni in cui si stavano affacciando su internet i primi «social media», ma a Nexa scegliemmo senza esitazioni un mezzo più tradizionale e più semplice, ovvero una lista di distri-

buzione di posta elettronica («mailing list», in inglese). Avevamo esperienza, infatti, della ricchezza rappresentata dalle discussioni e segnalazioni veicolate da molte «liste» di cui facevamo parte – a partire da quelle del Berkman Klein Center for Internet & Society della Harvard University, il centro di ricerca che servì da modello per la nascita del Centro Nexa – e non avevamo trovato alcun motivo per non usare anche noi lo stesso, ormai quasi venerabile strumento. La posta elettronica, infatti, la usavano tutte le persone interessate al digitale: non era quindi necessario installare nessuna nuova «app»; il mezzo testuale email consentiva riflessioni anche lunghe e strutturate (con citazioni, riferimenti, link, eccetera); i messaggi potevano essere facilmente archiviati, in modo da diventare un potenziale riferimento per il futuro; gli oneri di gestione per Nexa erano minimi.

Scegliemmo la strada della lista completamente libera, ovvero, chiunque fosse iscritto alla lista poteva inviare messaggi, senza moderazione[1]. Ci sembrava, infatti, importante che la discussione fosse spontanea, senza filtri a priori di nessun tipo. La lista debuttò il 2 maggio 2009, ovvero, pochi mesi dopo la presentazione pubblica del Centro Nexa, avvenuta il 22 gennaio di quell'anno nell'Aula Magna del Politecnico alla presenza, tra gli altri, di Stefano Rodotà (tra i Garanti fondatori di Nexa)[2].

Le speranze che avevamo riposto nell'idea di una lista pubblica si realizzarono pienamente: negli anni successivi, infatti, il numero degli iscritti aumentò sensibilmente, arrivando a sfiorare il migliaio (in questo momento sono 875), ma soprattutto la qualità delle segnalazioni e delle discussioni, fortemente interdisciplinari, si attestò a un livello

molto alto. Informatici, giuristi, umanisti, docenti, studenti, professionisti, e altri profili ancora impararono a confrontarsi in maniera reciprocamente comprensibile e costruttiva, permettendo ai lettori non solo di comprendere meglio alcune questioni particolarmente complesse, ma anche di venire tempestivamente informati sui principali sviluppi del settore.

Anche la scelta di lasciare la lista libera si rivelò corretta: in oltre quindici anni di esistenza, infatti, gli scambi si sono mantenuti sempre civili e corretti, e gli inviti alla moderazione, tipicamente per discussioni fuori tema della lista o per eccesso di messaggi da parte di determinate persone (eccesso che aveva l'effetto di scoraggiare il contributo di persone meno estroverse o con meno disponibilità di tempo) sono stati rarissimi.

Alla fine del 2022 la lista Nexa, animata in larga parte da persone non di Torino, fu investita, come quasi tutto il resto, dal «ciclone» ChatGPT. Lo tsunami di «hype» della cosiddetta «intelligenza artificiale generativa» trovò, però, una comunità ormai bene allenata ad analizzare in maniera oggettiva e distaccata l'ennesima novità partorita dalla costa ovest degli Stati Uniti. Nei mesi successivi, quindi, la lista fu caratterizzata da una serie di conversazioni e confronti particolarmente intensi, e non di rado illuminanti, come chiunque abbia voglia e tempo di consultare gli archivi potrà constatare di persona.

Questa volta, tuttavia, anche grazie allo stimolo dell'amico, nonché nexiano della prima ora, Giulio De Petra, è avvenuto qualcosa che avevamo spesso ventilato in passato, ma che non eravamo mai riusciti a concretizzare, ovvero, elaborare quanto passato in lista sotto forma di centinaia e

centinaia di messaggi individuali in qualcosa di sintetico e di sistematico. Il testo che segue è, quindi, frutto dell'intelligenza collettiva dei contributori alla lista Nexa, che colgo l'occasione per ringraziare, frutto a cui, però, gli autori hanno poi aggiunto non solo i loro preziosi interventi di sintesi e di organizzazione del materiale, ma anche importanti contributi personali legati ai loro specifici ambiti di studio, quello semiotico e quello giuridico. Il risultato complessivo è un'opera che aiuta a cogliere un fenomeno importante e attuale come ChatGPT e prodotti simili con l'ormai consolidato approccio nexiano, ovvero, rigoroso, ma accessibile a chiunque, e col contributo armonioso di tutti i saperi necessari per comprendere il fenomeno in questione. Buona lettura.

Note alla Prefazione

1. <https://server-nexa.polito.it/cgi-bin/mailman/listinfo/nexa>.
2. <https://nexa.polito.it/inaugurazione-del-centro-nexa-su-internet-societa/>.

Critica di ChatGPT

Nota degli autori

Questo libro è frutto di un progetto particolare. È stato immaginato dai tre autori, prendendo spunto dalle conversazioni tra una serie di studiosi ed esperti di IA generativa, all'interno della mailing list del Centro Nexa su Internet e Società del Politecnico di Torino, all'incirca dal febbraio del 2023 a oggi. Si tratta, dunque, di un lavoro che si basa sull'intelligenza collettiva di un gruppo di persone molto eterogeneo e interdisciplinare, che si occupa di intelligenza artificiale e desidera allo stesso tempo comprenderla e contribuire a realizzarla. Nello specifico, comunque, sistematizzando e rielaborando queste varie forme di sapere, Antonio Santangelo ha scritto l'introduzione, il capitolo 1, il capitolo 2 e le conclusioni, mentre Maurizio Borghi ha scritto il capitolo 3. Alberto Sissa ha collaborato alla stesura del capitolo 1 e del capitolo 2, raccogliendo molti dei materiali necessari e abbozzando i contenuti di quasi tutti i paragrafi, che sono stati poi ripresi e completati da Antonio Santangelo.

INTRODUZIONE

L'intelligenza artificiale e il mondo in cui vogliamo vivere

Se cerchiamo i titoli dei libri usciti in Italia per parlare di ChatGPT o dell'intelligenza artificiale generativa negli ultimi mesi, ci rendiamo conto che quasi tutti veicolano messaggi entusiastici, che spaziano da come imparare a fare soldi con questa tecnologia (la maggior parte), a come quest'ultima stia cambiando in meglio il nostro mondo. La chiave, secondo i vari autori, è il *prompting*, vale a dire l'arte di interrogare questo sistema secondo le logiche che lo fanno funzionare bene. L'idea, dunque, è che se impareremo a ragionare e operare nella maniera giusta, adattandoci ai suoi principi strutturali, ChatGPT e i suoi epigoni miglioreranno la nostra vita.

Da questo punto di vista, Luciano Floridi sostiene da anni (Floridi, Cabitza, 2021; Floridi, 2022) che l'intelligenza artificiale è una tecnologia che ci richiede di adattarci a essa. Il suo ragionamento è il seguente: un po'

come un giardiniere che voglia utilizzare un tosaerba automatico deve fare in modo che sul suo prato non ci sia alcun oggetto che ne possa intralciare il passaggio, se organizzeremo il nostro mondo affinché esso sia in grado di accogliere e di favorire il buon funzionamento di questo strumento, potremo servircene godendo dei suoi servigi. Ma perché dovremmo farlo? Perché dovremmo sobbarcarci la fatica preliminare di ispezionare il nostro giardino e di sollevare da terra gli oggetti che noi stessi o i nostri animali domestici vi abbiamo lasciato cadere? Non sarebbe più comodo rivolgerci a qualcuno che svolga questo lavoro manualmente? Quali sono, fuor di metafora, i vantaggi di servirci della IA e di ChatGPT, adattandoci alle loro logiche?

I vari autori che provano seriamente a rispondere a questa domanda (Kaplan, 2024; Hoffmann, 2023; De Baggis e Puliafito, 2023; Da Empoli, 2023) sostengono non esserci ambito della nostra vita che non verrà investito dagli influssi positivi di queste tecnologie, che già ci forniscono e sempre più ci forniranno un supporto alla creatività: riassumeranno, scriveranno e tradurranno sempre meglio per noi, svolgeranno al posto nostro i compiti più ripetitivi, aumentando la nostra produttività senza però toglierci il lavoro, ci aiuteranno a organizzare nella maniera migliore i nostri pensieri e le nostre attività, ci assisteranno quando cercheremo informazioni, ci faranno entrare in contatto col sapere in una maniera sempre più esaustiva ed equilibrata, ci formeranno, interagiranno con noi quando ci sentiremo soli e avremo bisogno di un buon consiglio, ci forniranno consulti medici, ci assisteranno nelle controversie legali, eccetera. Tutto questo, tra l'altro, se lasceremo fare al libero mercato, verrà proposto a tutti, a prezzi abbordabili e con

livelli di qualità sicuramente competitivi. Dovremo affrontare piccoli e grandi cambiamenti, accettare magari qualche malfunzionamento di questi sistemi, imparando al limite a essere critici nei confronti dei loro risultati e vagliando questi ultimi sempre con estrema attenzione, ma i vantaggi che ne trarremo saranno di certo superiori agli svantaggi.

Al momento, però, forse proprio perché non abbiamo investito abbastanza nella lettura dei libri o nella frequentazione dei corsi a pagamento sul *prompting*, e dunque non ci possiamo definire maestri del *prompt engineering* (così viene chiamato questo sacro graal dei tempi in cui viviamo), ci rendiamo conto che ogni volta che chiediamo a strumenti come ChatGPT di riassumere un nostro articolo per confezionarne l'abstract, esso non lo fa nello stile che riteniamo più opportuno, oppure non coglie gli elementi che giudichiamo più interessanti. Dunque dobbiamo affinare la nostra richiesta e fargli svolgere il lavoro diverse volte, prima di essere almeno parzialmente soddisfatti, inducendoci a domandarci se non avremmo fatto più in fretta – e soprattutto meglio – a scrivere da soli anche questo breve testo. Ma ci rispondiamo fiduciosi che di sicuro, in futuro, miglioreremo entrambi: la tecnologia si evolverà e noi saremo più abituati a servircene con destrezza, standardizzando certi compiti noiosi e liberando tempo per occuparci di questioni più importanti.

È proprio nelle pieghe di questa fiducia in un futuro che non è ancora giunto e di un presente piuttosto disfunzionale come quello che abbiamo appena descritto, che si sviluppa questo libro di critica agli strumenti come ChatGPT. Cominciano a circolare calcoli inquietanti sui costi ambientali del funzionamento di questo genere di tecnologie (Bender et al., 2021; Crawford, 2021; Becker, 2023). Alcuni

studi dimostrano il livello di sfruttamento e di prostrazione psicologica a cui sono sottoposti i lavoratori di diversi paesi del sud del mondo per allenare questi sistemi ed evitare che si esprimano in maniera inappropriata, lasciando emergere i contenuti razzisti, violenti, in alcuni casi pedofili, con cui sono entrati inavvertitamente in contatto nella loro fase di addestramento[1]. Anche quando non lasciano trasparire nulla di scabroso, le posizioni che esprimono sono affette da diversi *bias*, che favoriscono l'affermazione di culture egemoniche a scapito delle minoranze (Bender et al., 2021; Numerico, 2021). Se utilizzati da individui fragili, questi strumenti possono contribuire alla loro sofferenza psicologica[2]. Quando vi fanno ricorso i giudici, per cercare informazioni utili e confezionare i testi delle loro sentenze, possono inventare di sana pianta i riferimenti a casi di giurisprudenza mai esistiti[3]. Le autorità statali come il Garante della privacy italiano avanzano diversi dubbi sul fatto che i loro progettisti tengano nel debito conto la tutela dei dati personali[4]. Gli scrittori e gli artisti che hanno creato i contenuti testuali o visivi impiegati dalle aziende per «addestrare» i loro prodotti di intelligenza artificiale generativa protestano e denunciano il plagio dei propri lavori[5]: essi rischiano di perdere il proprio impiego o di vedere diminuiti i propri introiti, senza nemmeno essere stati remunerati per il contributo che hanno fornito allo sviluppo di sistemi che fanno loro concorrenza.

Nella retorica di chi si dichiara favorevole all'avvento di ChatGPT e dei suoi epigoni, i problemi che abbiamo appena enumerato sarebbero piccoli e grandi costi che bisogna sostenere, in funzione dei ben più sostanziosi benefici del progresso verso cui tutto ciò ci condurrà[6]. Del resto, OpenAI, l'azienda che ha creato ChatGPT, dichiara di lavo-

rare allo sviluppo di una «intelligenza artificiale generale», che da anni incarna il sogno prometeico di molti ingegneri e informatici di tutto il mondo. Qualcuno, tra questi stessi scienziati, sostiene addirittura di temere l'avvento di una tecnologia del genere, poiché rappresenterebbe un rischio per la sopravvivenza dell'umanità e andrebbe dunque controllata e regolamentata con grande attenzione[7]. Ma molti osservatori sottolineano come affermazioni di questo tipo siano scarsamente suffragate dallo stato dell'arte di questi strumenti e che servano solo per creare un *hype*, un sistema di aspettative ingiustificatamente elevato, utile solo a far aumentare di valore gli ingenti investimenti dei soggetti che li stanno finanziando, sviluppando e vendendo[8].

Il nostro tentativo è di decostruire pezzo per pezzo le narrazioni troppo entusiastiche sul futuro che ci attende grazie a ChatGPT e all'intelligenza artificiale generativa nel suo complesso, mostrando quali sono le questioni più spinose che questi sistemi ci costringono ad affrontare oggi. A questo proposito, partiamo chiedendoci se essi siano davvero intelligenti e di che tipo di intelligenza siano dotati. Ne analizziamo il funzionamento e presentiamo le argomentazioni di diversi studiosi che si sono interrogati circa l'utilità di definirli, appunto, «intelligenti». Mostriamo i loro limiti, soprattutto la loro tendenza a produrre «allucinazioni», affermazioni sullo stato del mondo che ci sembrano vere ma che non lo sono. Ci interroghiamo, quindi, sull'utilità di farvi ricorso in ambiti della nostra vita in cui è fondamentale essere certi della veridicità di ciò che essi sostengono, come ad esempio quello giudiziario. Parliamo delle varie critiche che sono state mosse a chi li vuole vendere come sistemi dotati di intelligenza generale o addirittura, in alcuni casi, di

una «super intelligenza». Riportiamo tanti esempi concreti dei problemi che questi strumenti e chi li sviluppa stanno riversando nella nostra società, dal punto di vista ecologico, di sfruttamento del lavoro, di affermazione di una certa egemonia culturale, di incidenza negativa sulle fragilità psicologiche, eccetera. Infine, affrontiamo le controversie legali in cui sono coinvolte le aziende che li commercializzano e le persone che ne contestano l'utilizzo, dato che non è chiaro se e come essi rispettino la proprietà intellettuale di chi ha prodotto i contenuti su cui vengono allenati, né come tutelino la privacy dei miliardi di utenti di internet che forniscono i dati personali necessari per il loro sviluppo.

Ciò che possiamo attestare è che a fronte del grande entusiasmo mostrato da molti sostenitori di una visione neoliberista della nostra società – i quali come abbiamo anticipato vedono ChatGPT e le intelligenze artificiali generative come strumenti di *empowerment* individuale di chi se ne potrà permettere l'utilizzo, traendone svariati vantaggi competitivi – si sollevano diverse voci critiche da parte di chi ritiene invece che ci troviamo davanti all'ennesimo prodotto dell'apparato socio-tecnico dominante nel nostro tempo, che pubblicizza come innovazione ciò che, in realtà, non fa altro che acuire disuguaglianze e forme di sfruttamento che non hanno nulla di nuovo. Secondo quest'altra visione, forse minoritaria nella pubblicistica generalista, ma molto attestata – come mostreremo – in quella accademica (che viene spesso rilanciata anche dagli organi di stampa, accordando così largo credito a queste tecnologie e a chi le realizza), non ci stiamo spingendo verso un futuro radioso, ma stiamo bensì contribuendo a rafforzare quella sorta di tecnofeudalesimo nel quale già viviamo (Varoufakis, 2023).

In effetti, come mostriamo nelle Conclusioni, questo genere di riflessioni su come le tecnologie digitali del presente ci consentano di immaginare il futuro in cui vorremmo vivere è fondamentale oggi, dato che questi strumenti sono le infrastrutture delle nostre società e ce ne serviamo per svolgere la maggior parte delle nostre attività quotidiane, dalle più routinarie a quelle che determinano le scelte negli snodi più importanti delle nostre esistenze. Molti parlano, a questo proposito, del fatto che ci troviamo immersi in un'epoca di «rivoluzioni», e anche se di rivoluzione digitale si parla da molti decenni, facendo dubitare qualcuno che anche questo termine sia un prodotto del marketing delle imprese che operano in questo settore (Balbi, 2022), bisogna sottolineare che, come sostiene Hannah Arendt (1963), le rivoluzioni possono certamente essere un modo per cambiare il mondo, ma non sempre conducono verso qualcosa di nuovo che modifichi e migliori radicalmente il presente; anzi, spesso esse riportano indietro chi le attraversa, verso forme di vita che egli o i suoi predecessori avevano già sperimentato e da cui avrebbero voluto prendere le distanze.

Per questo ci auguriamo che una solida critica di ChatGPT e di tutto ciò che i sistemi di intelligenza artificiale generativa sono e rappresentano oggi possa contribuire a rendere più consapevoli i nostri lettori circa il modo in cui essi desiderano che una simile tecnologia venga realizzata e prenda piede: se proprio dobbiamo o vogliamo fare spazio a questo genere di macchine, adattandoci a esse, allora è meglio esercitare la nostra influenza affinché siano fatte come ci sembra più opportuno.

Note all'Introduzione

1. <https://time.com/6247678/openai-chatgpt-kenya-workers/?fbclid=PAAabiRhajL9uq0H8WJVr38iZdo0UmZ7DrRPjbWq4lu8Oyhvj9CNtu5mG9uJs>.
2. <https://www.law.kuleuven.be/ai-summer-school/open-brief/open-letter-manipulative-ai>.
3. <https://verfassungsblog.de/colombian-chatgpt/>.
4. <https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9978020>.
5. <https://all-in-giuridica.seac.it/document/327/5131425/0>.
6. <https://www.bloomberg.com/news/articles/2024-05-15/microsoft-s-ai-investment-imperils-climate-goal-as-emissions-jump-30>.
7. <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>.
8. <https://sts-news.medium.com/youre-doing-it-wrong-notes-on-criticism-and-technology-hype-18b08b4307e5>.

CAPITOLO PRIMO

Come funziona ChatGPT

Le macchine che funzionano secondo le logiche della cosiddetta «intelligenza artificiale generativa» sono dotate di una forma di intelligenza simile a quella degli esseri umani?

Il bisogno di trovare una risposta può essere facilmente compreso se si riflette sulle emozioni che tali tecnologie hanno suscitato, soprattutto in funzione della loro rappresentazione mediatica. In questo periodo, al centro dell'attenzione ci sono, appunto, i grandi modelli linguistici generativi (anche chiamati LLM, ovvero Large Language Models)[1], fra i quali spicca ChatGPT, la versione di OpenAI, un'azienda che, come si può leggere sul suo sito[2], si pone l'obiettivo di creare un'intelligenza artificiale generale, vale a dire un sistema intelligente come o addirittura più di noi: è chiaro che un risultato del genere può impressionare e giustificare riflessioni più o meno ottimistiche sul nostro futuro.

La somiglianza fra intelligenze artificiali ed esseri umani

è sempre stata vissuta come una questione delicata, perché la convinzione che una macchina possa pensare e agire come noi può comportare l'instaurarsi di un alto grado di fiducia nei confronti di quest'ultima, al costo di sacrificare il nostro spirito critico (Sadin, 2021), modificando per molti versi il nostro sviluppo cognitivo e quello delle nostre abilità logiche e analitiche (Floridi, Cabitza, 2021). Inoltre, può portarci a ritenere che un computer o un robot ci possano sostituire integralmente in molti ambiti, primo tra tutti quello lavorativo[3], diventando il perno su cui si reggono il sistema sanitario, quello scolastico, quello finanziario, eccetera (Kaplan, 2024). Infine, questa convinzione ci può indurre a ritenere che per consentire di sviluppare al meglio questo genere di tecnologie dobbiamo rinunciare ad alcuni nostri diritti, primo tra tutti quello alla privacy[4].

Un modo per cominciare a riflettere sul problema della somiglianza tra l'intelligenza umana e quella artificiale è pensare che la nostra specie ha una tendenza intrinseca all'animismo. A questo proposito, già Hume scriveva di quanto siamo inclini a concepire gli altri esseri viventi come simili a noi e a trasferire addirittura negli oggetti qualità che riteniamo presenti nella nostra coscienza (Hume, 1757). Questo avviene dalla notte dei tempi, per «addomesticare» ciò che ci appare diverso e spesso minaccioso, rendendolo famigliare (Beals, Hoijer, 1953). Ma la storia è piena di persone che, ritenendo che gli animali selvatici ragionino come noi, ne sono state sbranate. Più in generale, molti individui che hanno creduto che la natura e le cose che da essa derivano siano sempre benevole, ne hanno tratto nocumento. Dunque il rischio, secondo alcuni, è che questo accada anche con l'intelligenza artificiale.

Occorre quindi ritornare alla domanda da cui siamo partiti, per capire se quando interagiamo con strumenti come ChatGPT, ricavandone l'impressione di condurre una conversazione simile a quella che potremmo intrattenere con un altro essere umano, siamo davvero di fronte a una tecnologia destinata a confondersi definitivamente con noi, essendo effettivamente intelligente, oppure se dietro a questo modo di comunicare si nasconda qualcosa che di umano non ha proprio nulla. Una volta stabilito questo, potremo ragionare meglio su come relazionarci con questo tipo di strumenti e su come utilizzarli nella nostra vita di tutti i giorni.

L'IA è davvero intelligente?

Nonostante la sensazione di poter intrattenere con strumenti come ChatGPT interazioni che di solito portiamo avanti con gli esseri umani, molti esperti del settore della IA – generativa e non solo – evitano di parlare esplicitamente di *intelligenza* quando vi si riferiscono, poiché la nozione stessa di ciò che significa essere «intelligenti» è piuttosto indefinita. Ad esempio, in un documento redatto da un Gruppo di esperti di alto livello istituito della Commissione Europea, si specifica che «nonostante sia oggetto di numerosissimi studi di psicologi, biologi e neuroscienziati, l'intelligenza (nelle macchine come negli esseri umani) rimane un concetto vago, tanto che i ricercatori di IA preferiscono utilizzare la nozione di razionalità» (Gruppo di esperti di alto livello sull'intelligenza artificiale, 2019)[5].

In effetti, poiché come viene evidenziato da due scienziati informatici come Legg e Hutter (2007) si possono

reperire in letteratura più di settanta definizioni della nozione di intelligenza, Russell e Norvig, gli autori di uno dei manuali più utilizzati nel mondo dell'ingegneria per progettare l'IA, sostengono che un sistema di questo tipo deve essere considerato intelligente se persegue «il concetto ideale di intelligenza, che noi chiamiamo razionalità. Un sistema è razionale se, date le sue conoscenze, fa la cosa giusta» (Russell, Norvig 2010: 4). In sostanza, «per ogni possibile sequenza di percezioni, un agente razionale dovrebbe scegliere un'azione che massimizzi il valore atteso della sua misura di prestazione, date le informazioni fornite dalla sequenza percettiva e da ogni ulteriore conoscenza dell'agente» (Russell, Norvig, 2010: 47).

Rifacendosi alle teorie di uno psicologo come Gardner (1983) bisogna però ammettere che questa visione dell'intelligenza, sicuramente importante e significativa, non prende in considerazione molte altre competenze intellettive degli esseri umani che non possono essere ridotte al solo utilizzo della *logica di matrice utilitaristica* (massimizzare un risultato, in funzione dei dati a propria disposizione), ma che riguardano tutto uno spettro di sensibilità fondamentali come quella emotiva, relazionale, etico-morale, eccetera. Cosa vuol dire «fare la cosa giusta» se si è medici, si scopre che un proprio paziente ha una malattia cerebrale degenerativa incurabile e si capisce che starebbe meglio se non lo sapesse? Oppure quando si vive sotto le minacce delle mafie e si comprende che un proprio gesto eroico di sacrificio, al limite anche della vita, potrebbe forse dare una scossa alla propria gente e ispirarla a reagire alle angherie subite? Oppure, molto più banalmente, quando si è insegnanti e ci si rende conto che un proprio allievo fallisce nell'apprendi-

mento ma si impegna molto, e ci si domanda, quindi, come comunicargli i suoi errori senza demotivarlo?

Il fatto di riferirsi alla forma di razionalità di cui parlano Russell e Norvig, come abbiamo visto, consente di operare *misurazioni* precise, come in quei test del quoziente di intelligenza che studiosi come lo stesso Gardner tendono a criticare, dato che queste prove consentirebbero di verificare, al massimo, se un sistema o un essere umano sono «intelligenti» nello svolgimento di un determinato compito, a prescindere dal contesto in cui questo deve essere svolto, e non se lo siano in generale, sapendo magari anche scegliere tra criteri e «misure di prestazione» contrastanti.

Del resto, come ricorda Cordeschi (2002), quello dell'intelligenza artificiale è un insieme di discipline nato a partire dalla *cibernetica*, la quale a sua volta si è chiesta come poter creare delle macchine capaci di simulare i comportamenti complessi degli esseri viventi, operando delle scelte di campo molto precise. In questo contesto, infatti, secondo Lieto (2021), il termine «intelligenza» è stato associato per la prima volta alla nozione di «artificiale» in un workshop sperimentale tenutosi nel 1955 al Dartmouth College di Hanover, nel New Hampshire, poco prima dell'evento accademico seminale che, in quello stesso luogo, secondo molti ha segnato la nascita degli studi sulla IA (Numerico, 2021). In quella occasione, si affermò un paradigma chiamato *simbolico*, che si basa sull'idea che l'intelligenza stessa, sia naturale sia artificiale, consista in processi di memorizzazione e manipolazione di informazioni, intese come simboli. Esso dominò nel panorama della ricerca negli anni a seguire e determinò la nascita di quelle che successivamente vennero definite le «GOFAI» (Good Old Fashioned Artificial Intelligence). Da

quel momento in poi, l'ambito degli studi sull'intelligenza della IA si è diviso in due branche: la prima, che oggi appare dominante, è legata al modello *logicista* e matematico di cui abbiamo scritto, finalizzato a dare origine ad agenti sintetici dotati di capacità che ci sembrano razionali anche se essi funzionano in maniera diversa da noi (questo paradigma è anche detto «funzionalista»); la seconda, invece, è incentrata sul modello *cognitivista* (o «strutturalista»), interessato a riprodurre nelle macchine un funzionamento simile a quello del pensiero umano (Lieto, 2021).

Diversi studiosi sono tuttavia scettici circa la possibilità dei sistemi che chiamiamo di «intelligenza artificiale» di riprodurre i meccanismi della nostra mente. Ad esempio, Margaret Boden, professoressa di Scienze Cognitive alla University of Sussex, sostiene che nel tempo, nel mondo delle discipline neuroscientifiche e filosofiche, si sarebbe affermata la perniciosa convinzione che si possa utilizzare l'IA per ridiscutere i classici problemi della mente:

> L'IA non ha influenzato solo le scienze della vita, ma anche la filosofia. Molti filosofi oggigiorno basano le loro concezioni della mente su nozioni di IA, nozioni che vengono usate, ad esempio, per fare luce sul noto problema mente-corpo, sull'enigma del libero arbitrio o sui molti misteri della coscienza. Tuttavia, queste idee filosofiche sono enormemente controverse, così come vi sono profonde divergenze riguardo al fatto che un sistema di IA possa essere dotato di vera intelligenza, creatività, o vita (Boden, 2019: 22).

Nonostante ciò, non tutti sono così critici nei confronti dell'incontro fra scienze cognitive e intelligenza artificiale.

In effetti, alcuni ricercatori e ricercatrici sostengono che l'idea secondo cui la mente umana potrebbe essere rappresentata informaticamente abbia permesso numerose svolte teoriche interessanti e innovative (van Rooij et al., 2023). Ma la maggior parte di questi studiosi ritiene che si debba stare attenti a considerare la nostra mente come qualcosa di simile a quella di una macchina e che questa convinzione dovrebbe costituire più che altro una fonte di ispirazione per far progredire la ricerca, poiché la nostra cognizione non può essere riprodotta del tutto a livello computazionale.

Capire il significato delle cose e parlarne

Nonostante quanto abbiamo appena scritto, frutto delle riflessioni di chi studia altre forme di intelligenza artificiale, la maggior parte degli esperti del tipo di IA di cui ci occupiamo qui, vale a dire quella generativa basata sui Large Language Models e sull'utilizzo dei cosiddetti Transformer Neural Networks (un tipo particolare di reti neurali), concorda nel ritenere che quest'ultima non riproduca il modo di funzionare della nostra mente. Essa, piuttosto, ci dà l'impressione di essere intelligente come noi, e a volte più di noi, grazie al fatto che vi possiamo interagire utilizzando il nostro linguaggio naturale e che ne riceviamo risposte confezionate correttamente nella nostra lingua. Ma tutto ciò che avviene tra l'istante in cui inseriamo in input le nostre parole o frasi e otteniamo come output affermazioni e informazioni che ci appaiono dotate di senso, si verifica in una maniera molto particolare, frutto di una tanto sofisticata quanto «meccanica» logica ingegneristico-matematica.

Per fornire una risposta alla domanda circa il tipo di «intelligenza» proprio della IA generativa, è dunque necessario comprendere come funziona. Un buon punto di partenza per impostare questo genere di riflessioni ci pare quello proposto dal famoso linguista americano Noam Chomsky. In un articolo comparso sul «New York Times» l'8 marzo 2023 afferma quanto segue:

> ChatGPT di OpenAI, Bard di Google e Sydney di Microsoft sono capolavori di apprendimento automatico. In parole povere, essi elaborano enormi quantità di dati, ne determinano gli schemi e diventano sempre più abili nel generare risultati statisticamente probabili – che prendono la forma di un linguaggio e di un pensiero apparentemente simili a quelli umani. Questi programmi sono stati acclamati come i primi segni di un avvicinamento all'intelligenza artificiale generale, quel momento a lungo profetizzato in cui le menti meccaniche supereranno i cervelli umani non solo quantitativamente, in termini di velocità di elaborazione e dimensioni della memoria, ma anche qualitativamente, in termini di intuizione intellettuale, creatività artistica e ogni altra facoltà distintiva dell'uomo […]. Ma la mente umana non è, come ChatGPT e i suoi simili, un macchinoso e dispendioso motore statistico per la ricerca di pattern, che si ingozza di centinaia di terabyte di dati ed estrapola la risposta più probabile a una conversazione o a una domanda scientifica. […] non cerca di inferire correlazioni brutali tra i dati, ma di creare spiegazioni[6].

Chomsky, in sostanza, si concentra sulle differenze che si possono riscontrare fra esseri umani e agenti artificiali, concludendo che non esistono motivazioni valide per ritenere che le macchine siano intelligenti come o

più di noi. La sua argomentazione si basa sull'idea che i «ragionamenti» degli strumenti dotati di intelligenza artificiale siano strutturalmente diversi dai nostri, per due ragioni: la prima riguarda il modo di fare «esperienza» del mondo da parte della IA, legato all'utilizzo di una grande mole di dati; la seconda è collegata con il significato che questi stessi dati assumono nel momento in cui le medesime macchine esibiscono le loro competenze, che sono di natura sintattica poiché discendono dall'utilizzo della statistica e del calcolo di ciò che di solito viene prima o dopo, in un testo in cui compaiono certe parole[7].

Il primo problema sollevato dal linguista americano è facilmente illustrabile con l'esempio da lui stesso proposto, ovvero quello di un bambino che inferisce le regole grammaticali della propria lingua partendo dalle poche informazioni che può cogliere nei primi anni della sua vita. L'essere umano sarebbe dotato di un sistema biologico per l'immagazzinamento e il processamento dei dati che gli giungono dall'esterno, molto più economico in termini energetici rispetto a quello di una IA, che lo predisporrebbe all'apprendimento senza bisogno di fare ricorso ai cosiddetti *big data*[8] (usiamo il condizionale perché comunque ci vuole molto tempo prima di imparare a parlare): a una creatura umana basterebbe molto meno di ciò che serve a una macchina per comprendere il senso delle cose e poi per capire come riferirvisi. In questo, i sistemi dotati di intelligenza artificiale sarebbero molto diversi da noi e, almeno da questo punto di vista, dovrebbero essere giudicati come meno intelligenti.

La seconda questione, invece, è più complessa e ricca di implicazioni. Innanzitutto, quando ci serviamo della lingua per parlare del mondo, quest'ultima, oltre a necessi-

tare della sintassi, si appoggia anche sulla *semantica* e sulla *pragmatica* (Morris, 1938). Non basta sapere come devono essere messe in sequenza le parole, ma bisogna conoscere anche il loro significato (la loro natura semantica, per l'appunto) e ciò che si fa quando le si pronuncia (la loro consistenza pragmatica). Ma mentre le macchine dotate di intelligenza artificiale sembrano saper gestire molto bene la dimensione sintattica del linguaggio, sarebbe molto dubbia la loro capacità di maneggiare le altre due, per una serie di ragioni che approfondiremo meglio tra poco.

Inoltre, per costruire i loro discorsi, come abbiamo scritto, le IA si basano su una grande mole di dati che, però, non sono *tutti i dati* che possediamo sulla realtà e, anche se lo fossero, potrebbero non essere quelli giusti, per dire come quest'ultima funziona effettivamente. Anche noi umani potremmo trovarci in una condizione simile (non essere certi che ciò che sappiamo coincida con il modo in cui le cose funzionano veramente), ma ne saremmo consapevoli, mentre i nostri strumenti dotati di intelligenza artificiale sarebbero condannati a effettuare le loro elaborazioni solo sulle informazioni che fanno parte dei loro database, senza «sapere» che potrebbero essere sbagliate o desuete.

Infine, la loro natura statistica impedirebbe a questo genere di sistemi di concepire *spiegazioni improbabili* rispetto ai dati da cui partono. Chomsky in proposito porta l'esempio della teoria della gravitazione universale, per rimarcare come lo statuto di verità di un'ipotesi circa il funzionamento del mondo sia indipendente dal grado di probabilità che essa ha di essere verificata. Citando Popper, questo può essere riassunto in una massima che andrebbe

sempre ripetuta, secondo il linguista americano, come una sorta di mantra del pensiero scientifico moderno: «Non cerchiamo teorie altamente probabili, ma spiegazioni, cioè teorie solide e altamente improbabili»[9] (Popper, 1962: 58). L'intelligenza umana e quella della IA, in sostanza, anche a fronte di questi ragionamenti sarebbero decisamente diverse e quella delle macchine, ancora una volta, sarebbe più limitata della nostra.

Per spiegare meglio le argomentazioni dello stesso Chomsky, in particolare la sua critica alla natura puramente sintattica dell'intelligenza della IA, si può ricorrere all'esperimento mentale della stanza cinese, proposto da John Searle in un famoso articolo intitolato *Minds, brains, and programs* (Searle, 1980). Qui, il filosofo statunitense tenta di fornire un'argomentazione efficace riguardo l'impossibilità di parlare di una forma di pensiero «umano», quand'anche ci si trovasse davanti a un'intelligenza artificiale forte[10] (come, lo ricordiamo, secondo alcuni potrebbe essere ChatGPT). Egli immagina di essere rinchiuso in una stanza con un prontuario di regole procedurali per poter utilizzare il cinese, nonostante non lo parli e non lo comprenda. Immagina, inoltre, che un cinese all'esterno della stanza interagisca con lui nel suo idioma, ignaro del fatto che egli non lo conosca. Ma il libro delle regole, scritto evidentemente in inglese, è talmente efficace da permettergli di capire quali simboli deve utilizzare per rispondere correttamente alle sollecitazioni che gli giungono dal suo interlocutore. Come avviene nel test di Turing[11], il cinese all'esterno della stanza avrebbe l'impressione di trovarsi di fronte a un suo simile, ma questo sarebbe problematico, perché Searle, dall'interno, non avrebbe la minima

idea di cosa gli venga detto e di cosa stia rispondendo: egli applicherebbe semplicemente alcune regole scritte in un linguaggio che conosce, proprio come succede nel caso delle macchine dotate di intelligenza artificiale.

In pratica, saper combinare sintatticamente i simboli che compongono una lingua, senza conoscerne il significato, vale a dire la semantica, e senza comprendere che effetti pragmatici si producono componendo certe frasi, non vorrebbe dire «ragionare» come un umano ed essere «intelligente» come lui, anche se ciò può fornire *l'impressione* che sia così. Del resto, come hanno sostenuto Natale (2021) e Numerico (2021), l'IA è stata concepita fin da subito – da alcuni dei pensatori che come il già citato Turing si sono interrogati su ciò che dovesse e potesse essere – come uno strumento diverso da noi nel suo modo di funzionare, ma capace di *dissimulare* questa sua diversità, con tutte le conseguenze che questo può comportare.

Allucinazioni

La riflessione sulla differenza tra come noi e i sistemi tipo ChatGPT ci diciamo (per mezzo del linguaggio) come funziona il mondo e lo comunichiamo potrebbe andare avanti ancora a lungo. Qualcuno, ad esempio, potrebbe sostenere che senza sperimentare col proprio corpo il significato delle cose non lo si possa comprendere appieno (Horchak et al., 2014) e che siccome i chatbot che usano i Large Language Models non hanno sensori, ma accedono al loro sapere solo attraverso dataset di testi che parlano delle cose stesse, la loro «intelligenza» sarebbe, ancora una volta, inferiore alla

nostra. Ma qualcun altro, facendo riferimento alle semantiche non referenzialiste, non legate cioè al rapporto diretto tra le parole e ciò a cui esse si riferiscono (ad esempio Saussure, 1916), potrebbe invece argomentare che quando diciamo a qualcuno di avere un gatto, costui, non avendo visto il nostro, comprenderà solo il valore differenziale di quel vocabolo, opposto ad esempio a quello del lessema /cane/ nella nostra lingua. Dunque, allo stesso modo, una macchina, con i suoi strumenti statistici, potrebbe avere chiara questa medesima differenza, non distanziandosi poi troppo da come noi ragioniamo e assegniamo un significato al mondo.

Comunque, pur non essendo ancora del tutto chiaro quanta differenza ci sia tra l'intelligenza umana e quella artificiale, dopo ciò che abbiamo riportato appare evidente – e magari anche un po' scontato – che le macchine non funzionino *esattamente* come noi. Tuttavia, se proprio si dovesse trovare qualcosa in cui ci assomigliamo, forse si potrebbe citare l'innata capacità di sbagliare e di dire inavvertitamente il falso. L'unico problema è che, come abbiamo anticipato nei paragrafi precedenti, gli strumenti dotati della cosiddetta IA non sembrerebbero consapevoli di farlo e quindi non potrebbero accorgersene per poi chiedere ammenda.

Tra i vari errori in cui possono incorrere le tecnologie che si servono dell'intelligenza artificiale, ve ne sono alcuni che gli esperti hanno deciso di chiamare *allucinazioni*, per tradurre in termini comprensibili qualcosa che, altrimenti, sarebbe complesso da spiegare tecnicamente. In effetti, il significato della parola «allucinazione» viene comunemente inteso come quel «fenomeno psichico, provocato da cause

diverse, per cui un individuo percepisce come reale ciò che è solo immaginario»[12]. Questa definizione, se applicata a strumenti come ChatGPT che si basano sugli LLM, riesce a cogliere efficacemente il fatto che questi ultimi, non potendo relazionarsi direttamente col mondo, ma dovendo sottostare alla mediazione di programmi, linguaggi informatici e, soprattutto, testi dai quali «apprendono», possono non essere in grado di distinguere correttamente la realtà – o forse, meglio, ciò che è vero da ciò che non lo è. Tutto dipende da come sono fatti i contenuti su cui essi si appoggiano e dal funzionamento degli algoritmi che stabiliscono come processarli. Per questo motivo, tali tecnologie vengono viste spesso come una sorta di *stampanti digitali* o di *pappagalli stocastici* (Bender et al., 2021), per sottolineare che, di fronte alla richiesta di descrivere o di parlare di un certo fenomeno, non farebbero altro che «stampare», per l'appunto, pezzi di diversi testi che lo hanno già fatto. Questo presuppone un'idea molto precisa di cosa sia il mondo fenomenico, che viene così concepito come un contesto in cui si trovano oggetti e si verificano situazioni simili tra di loro, in modo tale che alcune descrizioni o discorsi ricorrenti su di essi possano essere utilizzati tutte le volte che ve ne sia la necessità. Una scelta che inevitabilmente sacrifica il particolare per il generale, l'inaspettato, ciò che è diverso, in favore di ciò che è già codificato. Cosicché, ogni tanto, può accadere che una macchina basata su questi principi non si «renda conto» che le cose che è chiamata a «interpretare» non sono come essa le «vede».

Ancora una volta, il lettore più attento si sarà accorto che questi meccanismi non appaiono così diversi da quelli attraverso cui la nostra intelligenza ci consente di rappor-

tarci con la realtà e di interpretarla correttamente. Anche noi, infatti, facciamo ricorso all'esperienza, mediata da linguaggi e da codici che spesso apprendiamo a partire da testi che dicono cose simili, dandoci l'impressione che la realtà stessa sia come viene definita al loro interno (Violi, 1997). Il nostro cervello opera continuamente interpolazioni con le informazioni che sono già in suo possesso in modo da offrirci la sensazione di essere in un mondo «stabile» e «continuo» (Ferraro, 2012; Santangelo, 2013), ad esempio quando gli organi sensoriali gli fanno mancare qualche dato necessario per completare il quadro di ciò che esso si trova davanti (Lindzey, Thompson, Spring, 1975). Tuttavia, come si vede bene se si riflette sul significato delle parole che vengono utilizzate per parlare degli LLM – «stampanti» e «pappagalli» – è chiaro che gli strumenti che se ne servono vengono spesso giudicati negativamente, come se ci fossero inferiori, perché compiono degli errori che noi non commetteremmo.

Di questo tema si è discusso molto. Ad esempio, come riporta il podcast *Data Knightmare* di Walter Vannini[13], Stefano Quintarelli, un ex parlamentare della Repubblica Italiana, secondo ChatGPT sarebbe stato membro di un certo insieme di istituzioni a cui non aveva in realtà mai preso parte e, dopo che gli era stato fatto notare più volte che forniva informazioni sbagliate, lo stesso chatbot si è scusato, sostenendo che quella persona fosse morta, mentre in realtà era viva e vegeta. Questo consente di capire qual è il tipo di «deviazione» dalla realtà che gli output degli LLM producono quando prendono un abbaglio, ma soprattutto consente di intuire le conseguenze che ciò può comportare. In questo caso, il succitato Vannini era infatti

a conoscenza del fatto che Quintarelli fosse ancora in vita, ma qualsiasi utente ignaro della verità, senza un dovuto controllo, sarebbe potuto cadere nella trappola. Dunque viene da chiedersi: come ha fatto ChatGPT a inventare una notizia falsa di questo genere? Inoltre, appurato che non sia progettato per mentire, cosa è successo durante la generazione delle sue affermazioni?

Per rispondere a questi quesiti, si può partire da un articolo del «New Yorker»[14]. Al suo interno, si parla di uno strano errore commesso da una fotocopiatrice Xerox, di cui furono testimoni i lavoratori di un'azienda edile tedesca:

> Nel 2013, alcuni lavoratori di un'impresa edile tedesca hanno notato uno strano fenomeno riguardo le loro fotocopiatrici Xerox: quando facevano una copia della pianta di una casa, la copia differiva dall'originale in modo sottile ma significativo. Nella pianta originale, ciascuna delle tre stanze della casa era accompagnata da un rettangolo che ne specificava la superficie: le stanze erano rispettivamente di 14,13 mq, 21,11 mq e 17,42 mq. Tuttavia, nella fotocopia, tutte e tre le stanze erano indicate come di 14,13 metri quadrati […]. Le fotocopiatrici Xerox utilizzano un formato di compressione con perdita noto come JBIG2, progettato per l'uso di immagini in bianco e nero. Per risparmiare spazio, la fotocopiatrice identifica regioni simili nell'immagine e memorizza una singola copia per tutte; quando il file viene decompresso, utilizza quella copia ripetutamente per ricostruire l'immagine. Si è scoperto che la fotocopiatrice aveva giudicato le etichette che specificavano l'area delle stanze abbastanza simili da doverne memorizzare solo una – 14,13 – e l'ha riutilizzata per tutte e tre le stanze quando ha stampato la planimetria (Chiang, 2023)[15].

Il problema, dunque, era dovuto al fatto che, nel mondo delle fotocopiatrici digitali, alcuni algoritmi diminuiscono il numero di bit che dovrebbero altrimenti essere immagazzinati nei file che contengono le informazioni necessarie per riprodurre ciò che deve essere copiato. Questi algoritmi si definiscono *lossy* – che significa, grossomodo, «perdenti informazione» – perché in cambio di una maggiore economicità sacrificano la precisione e la completezza. Essi si contrappongono agli algoritmi di tipo *lossless* – «che non perdono informazione» – i quali sono meno efficienti, ma riproducono pedissequamente tutto ciò che le fotocopiatrici si trovano davanti. Le Xerox in questione utilizzavano i primi e, per fare economia, interpretavano male i dati che giungevano loro dagli scanner di cui erano dotate, giudicando erroneamente come simili alcune componenti della realtà che dovevano riprodurre. In pratica, realizzando fotocopie in cui tre riquadri che avrebbero dovuto contenere cifre diverse riportavano sempre la stessa, producevano un'allucinazione.

Qualcosa di analogo accade quando si utilizzano gli LLM su cui si fonda ChatGPT:

> Questa analogia con la compressione *lossy* non è solo un modo per comprendere la capacità di ChatGPT di riconfezionare le informazioni trovate sul web utilizzando parole diverse. È anche un modo per comprendere le «allucinazioni», o le risposte senza senso a domande concrete, a cui i modelli linguistici di grandi dimensioni come ChatGPT sono decisamente inclini. Queste allucinazioni sono artefatti di compressione, ma, come le etichette errate generate dalla fotocopiatrice Xerox, sono abbastanza plausibili da richiedere un confronto con gli originali, che

in questo caso sono il web o la nostra conoscenza del mondo. Se ci pensiamo bene, queste allucinazioni sono tutt'altro che sorprendenti; se un algoritmo di compressione è progettato per ricostruire un testo dopo che il 99% dell'originale è stato scartato, dobbiamo aspettarci che una parte significativa di ciò che genera sia interamente inventata[16].

Come le fotocopiatrici *lossy*, ChatGPT e le intelligenze artificiali generative che usano gli LLM si trovano di fronte a una realtà – la grande mole di dati da cui attingono per parlare del mondo – che non riproducono per intero, effettuando delle scelte di tipo statistico per generare documenti il più possibile simili a quelli che utilizzano. Queste scelte possono però rivelarsi sbagliate e dare origine ad affermazioni che non corrispondono alla verità. Inoltre, proprio come accadde ai lavoratori della Xerox che in fondo videro un'immagine quasi del tutto identica a quella fotocopiata, chi usa questo genere di strumenti potrebbe non accorgersi dei loro errori, a meno di non effettuare dei controlli minuziosi: come spesso succede nel mondo dei chatbot «intelligenti», una forma espositiva convincente può nascondere molte insidie.

Prendere sul serio ChatGPT e i suoi simili

Una volta compreso, in termini generali, come si producono le allucinazioni di ChatGPT e delle IA basate sugli LLM, ci possiamo inoltrare in alcune ulteriori riflessioni. Un tema importante è sicuramente quello della *qualità delle informazioni* contenute nei dataset su cui si fonda la conoscenza della realtà da parte degli strumenti dotati di

intelligenza artificiale. Riprendendo l'esempio da cui siamo partiti in questo paragrafo, di Stefano Quintarelli che ha scoperto di essere considerato morto proprio da ChatGPT, può accadere che, per qualche fortuito caso di omonimia, tra i testi in possesso del famoso chatbot di OpenAI ve ne fossero di più legati a un altro Stefano Quintarelli deceduto, rispetto a quello vivo. Da qui l'errore «allucinatorio» della macchina.

Questo ci riporta alla definizione di «*stochastic parrots*» («pappagalli stocastici»). Se non c'è possibilità che gli strumenti basati sugli LLM accedano al piano della semantica (si veda il paragrafo precedente), i loro output saranno sempre un sofisticato *collage* di parole scritte da altri, di cui essi non conoscono il significato e di cui, per di più, non possono controllare l'effettiva corrispondenza con la realtà. Che riescano quindi a fornirci una fonte di conoscenza utile e non dannosa dipende solamente da meccanismi statistici molto impersonali e dalla buona «qualità» dei dati a cui attingono. Eppure, questo non scoraggia chi, con entusiasmo, sostiene che dovremmo prendere molto sul serio le affermazioni prodotte da questi sistemi.

Ad esempio, qualcuno riprende l'obiezione del «sistema» all'esperimento mentale della stanza cinese di Searle (che, lo ricordiamo, è volto a dimostrare che le macchine dotate di intelligenza artificiale non sono intelligenti come noi)[17]. Essa consiste nel ritenere che anche se un computer che utilizza degli algoritmi per interagire con i parlanti di una lingua che non conosce può riuscire in questo intento senza comprendere il significato di ciò che «dice» o che gli viene detto, questo non è rilevante, perché tale significato è stato compreso da qualcun altro, per l'appunto all'interno

del *sistema* che ha prodotto tali algoritmi (ad esempio, rimanendo nel settore dei chatbot basati sugli LLM, da coloro che hanno scritto i testi che li alimentano e su cui si allenano). In quest'ottica, il significato si troverebbe comunque, in qualche modo, nella macchina e nei suoi output. Searle e altri pensatori rispondono che nondimeno l'IA non si possa definire intelligente, dato che semplicemente, in questa prospettiva, incorporerebbe l'intelligenza altrui, continuando a non capire nulla del cinese[18], ma evidentemente questo non basta a convincere tutti quelli che ritengono che ciò sia sufficiente a poterla ritenere tale.

Una serie di altri argomenti utilizzati per mettere in discussione la metafora dei pappagalli stocastici, e sostenere che con ChatGPT e i suoi simili ci troviamo di fronte a qualcosa di diverso, si può leggere in questo post scritto da Giuseppe Attardi, professore ordinario di informatica presso l'Università di Pisa, in una discussione dentro la mailing list del Centro Nexa del Politecnico di Torino che riportiamo qui, perché riprende in maniera chiara e sintetica posizioni che si possono reperire anche altrove[19]:

Gli LLM sono tutt'altro che «stochastic parrots» […]. Non attaccano affatto insieme «sequenze di forme linguistiche». Questo riprodurrebbe appunto pezzi di frasi a pappagallo. Invece il loro meccanismo di base è quello del calcolo della distribuzione di probabilità della prossima parola a seguire in una sequenza.

Come spiega Giorgio Parisi nel suo libro *In un volo di storni*, il funzionamento di un sistema complesso è il risultato dell'applicazione su larga scala di semplici leggi probabilistiche. I Large Language Models come GPT-3[20] fanno appunto questo: sono stati allenati a stimare una distribuzione di probabilità.

Da questa capacità usata su larga scala (miliardi di parametri a rappresentare connessioni tra neuroni) emergono altre capacità, apparentemente scollegate, come quella di rispondere a domande, riassumere testi, tradurre, scrivere codici, generare immagini, compiere semplici ragionamenti, eccetera.

Il risultato più che altro esula dalla nostra capacità di «comprensione» e tendiamo a concludere che non hanno capacità di «comprensione», anche se sono in grado di superare molti dei test che sono stati sviluppati proprio per valutare la capacità di «comprensione» umana.

Se dicessero sempre sciocchezze, dovremmo solo farci una risata: in realtà le risposte sono quasi sempre plausibili, e quindi coerenti col «meaning», anche se a volte inesatte.

Come si vede, Attardi non afferma che le macchine dotate di intelligenza artificiale siano intelligenti come noi o che comprendano la nostra lingua, ma prende seriamente in considerazione il fatto che esse producano frasi e testi dotati di significato. Lo fanno probabilmente in maniera diversa rispetto agli esseri umani, ma raggiungendo a modo loro questo risultato possono diventare delle nostre interlocutrici.

Tutto questo genera, a cascata, una serie di problemi, di cui ci occuperemo più a fondo nel prosieguo di questo libro. Ad esempio, come vedremo, Bender, Gebru e le altre autrici dell'articolo sui pappagalli stocastici di cui stiamo discutendo, sostengono che l'approccio *lossy* alla riproduzione del nostro sapere sul mondo, su cui si fonda l'«intelligenza» di strumenti come ChatGPT, può portare a trascurare quello prodotto da minoranze o gruppi che non hanno accesso a internet e che non producono i contenuti che entrano a far parte dei *dataset* sui quali si appoggiano

gli LLM. Inoltre, come sostiene Attardi, dobbiamo considerare il fatto che l'IA, oggi, è in grado di superare diversi test di intelligenza che siamo soliti utilizzare per valutare quella umana. Questo non può non avere conseguenze, soprattutto sulla credibilità di questa tecnologia come mezzo di cui servirci nella vita quotidiana per svolgere attività che reputiamo «intelligenti». Di questo tema ci occuperemo nel prossimo paragrafo.

Le macchine ci stanno superando?

Come abbiamo ricordato più volte, il dibattito è ancora molto aperto a proposito del fatto che le macchine che utilizzano l'IA siano dotate di una qualche forma di intelligenza più o meno simile alla nostra. Eppure è chiaro che esse sono in grado di svolgere compiti che fino a oggi riconoscevamo come appannaggio degli esseri umani intelligenti. A questo proposito, il ragionamento di Giuseppe Attardi che abbiamo riportato nel paragrafo precedente contiene un'interessante riflessione, ovvero che per giudicare l'operato degli LLM bisognerebbe concentrarsi sui loro risultati, evitando di addentrarsi in discussioni inconcludenti sul loro reale grado di competenza semantica. Sulla scia della logica che sta alla base del test di Turing, l'idea sarebbe quella che se uno strumento come ChatGPT fa delle cose che ci appaiono intelligenti, allora lo dobbiamo considerare tale.

Se la si pensa così, però, ci sono delle conseguenze che si chiariscono leggendo i contenuti di un articolo di Katz, Bommarito e altri (2023), in cui si afferma che il grande

pacchetto di aggiornamenti che OpenAI, la società produttrice di ChatGPT, ha apportato alla versione numero quattro del suo noto chatbot, ha riscosso un inaspettato successo all'esame di abilitazione statunitense alla professione di avvocato (chiamato «UBE – Uniform Bar Examination»), superando ogni più rosea aspettativa:

> Valutato su tutti i componenti dello UBE, come accadrebbe per qualsiasi esaminato umano, GPT-4 ha ottenuto circa 297 punti, superando in modo significativo la soglia di sufficienza per tutte le giurisdizioni UBE. Questi risultati documentano non solo il rapido e notevole progresso delle prestazioni dei modelli linguistici di grandi dimensioni in generale, ma anche il potenziale di tali modelli per supportare l'erogazione di servizi legali nella società[21].

Si tratta di un risultato sicuramente d'impatto, soprattutto a livello sociale: il solo fatto che un'intelligenza artificiale sia in grado di superare con successo un esame di questo tipo suggerisce inevitabilmente l'idea che il mestiere dell'avvocato, un giorno, potrebbe essere sostituito, così come molti altri. A latere di questa inquietante prospettiva, esisterebbero comunque degli evidenti risvolti positivi che, secondo gli autori dell'articolo, corrisponderebbero al supporto che questi nuovi mezzi tecnologici potrebbero garantire a chi già lavora in questo settore, sgravandolo dei compiti più noiosi e ripetitivi e aumentando la sua produttività. Inoltre, come sostiene Kaplan (2024), questi strumenti – che sarebbero capaci di scrivere memorie difensive del tutto efficaci, contratti e quant'altro – potrebbero essere utilizzati per fornire, a prezzi contenuti, un'assistenza legale di ottimo livello a chi oggi non se la può permettere.

Viene da chiedersi, allora, se sia vero che il successo nel test UBE possa essere un indizio di un reale superamento delle capacità intellettive delle persone da parte delle macchine. Non sono di questo avviso Kapoor e Narayanan[22], i quali sostengono che l'UBE non può essere un terreno di prova adeguato a comprendere la capacità di ragionamento di GPT-4, perché questa versione di ChatGPT è stata allenata con il materiale da cui si è attinto per creare l'esame in questione. Inoltre, i due studiosi sottolineano che situazioni simili possono interessare anche altri test di abilitazione professionale ed è per questo motivo che non possono essere considerati dei banchi di prova adeguati per verificare le prestazioni dei modelli linguistici. In sostanza, il successo ottenuto da GPT-4 racconterebbe solamente che esso è in grado di memorizzare correttamente le risposte giuste per ottenere la sufficienza.

Stando così le cose, sembrerebbe che, almeno per ora, non ci sia da preoccuparsi che nel prossimo futuro un'intelligenza artificiale possa prendere il posto degli esseri umani e difenderci in un'aula di tribunale. Purtroppo, però, non tutti sono della stessa idea e sono altresì ben noti gli episodi in cui un'eccessiva fiducia in questa tecnologia ha causato problemi. A questo proposito, può essere utile riportare le parole di un articolo di Juan David Gutiérrez[23], in cui viene analizzato l'operato di alcuni giudici colombiani, che sono stati al centro di scandali giudiziari causati da un utilizzo improprio di ChatGPT per il fatto di essersene serviti per scrivere le loro sentenze. Secondo l'autore, un professore di diritto colombiano, questi giudici avrebbero seguito pedissequamente i responsi di ChatGPT dopo aver posto domande essenziali per prendere le proprie decisioni – ad

esempio, in un caso sul pagamento delle spese mediche da parte delle assicurazioni per un individuo autistico, hanno chiesto se vi fosse giurisprudenza sull'argomento – fidandosi della bontà delle stesse come se provenissero da un oracolo:

I giudici non hanno usato ChatGPT in modo informato o responsabile […]. La risposta breve è che hanno usato ChatGPT come se fosse un oracolo: una fonte di conoscenza affidabile che non richiedeva alcun tipo di verifica. Sebbene i giudici siano stati trasparenti sul fatto di aver utilizzato lo strumento e abbiano inserito le virgolette per distinguere il contenuto prodotto da ChatGPT, il loro uso non è stato informato né responsabile.

Ci sono tre ragioni principali per cui il modo in cui ChatGPT è stato usato dalla magistratura in questi casi è molto preoccupante per i colombiani e non solo. In primo luogo, la posta in gioco nelle sentenze giudiziarie è troppo alta – soprattutto quando sono coinvolti i diritti umani – per giustificare l'uso di tecnologie inaffidabili e non sufficientemente testate. A causa del modo in cui i LLM come ChatGPT sono sviluppati e operano, questi strumenti tendono a produrre risposte errate e imprecise e a confondere la realtà con la finzione. Persino il CEO di OpenAI ha riconosciuto, nel dicembre 2022, che «ChatGPT è incredibilmente limitato […] è un errore fare affidamento su di esso per qualsiasi cosa importante in questo momento». Inoltre, per motivi strutturali, è improbabile che questi problemi di LLM vengano risolti a breve.

In secondo luogo, le risposte di ChatGPT non avrebbero dovuto essere accettate e prese alla lettera, ma essere piuttosto confrontate con altre fonti più affidabili. La sentenza afferma che le informazioni offerte da ChatGPT sarebbero state «corroborate». Tuttavia, non c'è alcuna traccia esplicita nel testo che ci

permetta di concludere che il giudice Padilla o il suo cancelliere abbiano effettivamente controllato se le risposte di ChatGPT fossero accurate. Infatti, ho replicato le quattro domande poste dal giudice Padilla e il chatbot ha risposto in modo leggermente diverso, un risultato che non sorprende visto il funzionamento dello strumento. Inoltre, quando ho chiesto a ChatGPT di fornire esempi di sentenze della Corte costituzionale che giustificassero le sue risposte, il chatbot ha inventato i fatti e la logica determinanti e ha citato una sentenza che non esisteva (inventando i fatti e il verdetto del giudice).

Dato che ChatGPT e gli altri LLM attualmente disponibili sono evidentemente inaffidabili, poiché i loro risultati tendono a includere informazioni errate e false, i giudici avrebbero bisogno di molto tempo per verificare la validità dei contenuti generati dalla IA, annullando così qualsiasi significativo «risparmio di tempo». Come accade con l'IA in altri settori, sotto la narrazione di presunte «efficienze», i diritti fondamentali possono essere messi a rischio.

In terzo luogo, c'è il rischio che i giudici e i loro cancellieri si affidino eccessivamente alle raccomandazioni della IA, incorrendo nel cosiddetto «pregiudizio da automazione»[24].

Tutti i problemi sollevati da Gutiérrez non sono nuovi per chi sta leggendo questo libro, dato che ancora una volta si tratta di confrontarsi con allucinazioni, algoritmi *lossy* e database «poveri» o mal costruiti, su cui si allena ChatGPT. Il fatto è che, come sottolinea lo stesso Gutiérrez al termine della sua lunga riflessione, il chatbot di OpenAI è in grado di costruire rapidamente testi ben scritti, che appaiono del tutto verosimili, come se fossero vergati da una mano umana, velocizzando e alleviando il lavoro di chi se ne serve.

Se a questo si aggiunge che qualcuno, come abbiamo visto, fa circolare la notizia che questi strumenti informatici superano i test umani per l'accesso al mestiere di avvocato, essi cominciano ad apparire affidabili, facendo sì che nessuno si preoccupi più di verificare le loro affermazioni. Eppure, chi vorrebbe essere giudicato per mezzo di questo tipo di intelligenza artificiale? E più in generale, chi desidererebbe che nel processo per prendere decisioni che possono influire sulla propria vita, venissero considerati come interlocutori autorevoli degli strumenti come ChatGPT?

Conclusioni

Alla fine di questo capitolo, non ci sentiamo di aver fornito una risposta definitiva alla domanda di partenza, vale a dire se le macchine dotate di intelligenza artificiale siano intelligenti come o più di noi. Le differenze fra il nostro modo di «funzionare» e quello dei nostri strumenti informatici sono molte. Comunque, abbiamo mostrato che esistono diversi elementi teorici per supporre che l'intelligenza artificiale sia meno intelligente degli esseri umani.

La difficoltà del quesito ci ha permesso però di approfondire il discorso sull'interazione tra uomo e macchina. Ci siamo resi conto che un tema fondamentale è quello della forma convincente dei contenuti prodotti da strumenti come ChatGPT, che ci appaiono corretti grammaticalmente, coerenti con le nostre aspettative e sensati. Eppure abbiamo visto che queste tecnologie non sembrano dotate di alcuna capacità semantica, soprattutto se con questo termine si intende la qualità di poter verificare il rapporto tra

le parole che esse producono e la realtà a cui si riferiscono. Dunque, se i loro discorsi ci appaiono dotati di significato è per via dell'utilizzo esclusivo di regole sintattiche, risultato della formulazione statistica di un modello linguistico. Questo, insieme al fatto che gli algoritmi su cui si basano gli LLM sono di tipo *lossy*, genera le allucinazioni di cui abbiamo scritto, che rappresentano una sorta di «squarcio del velo di maya» che si può creare facilmente facendo ricorso a queste forme di «intelligenza» artificiale. Quando ci rendiamo conto della natura allucinatoria di ciò che ci dicono le macchine, allora capiamo che, forse, gli «allucinati» siamo noi che crediamo loro, spesso in maniera acritica, solo per il fatto che si esprimono correttamente. Esse, in realtà, non sbagliano, perché applicano una procedura, decisa sia in sede di programmazione sia in sede di autoapprendimento, fondata sull'unico criterio del calcolo statistico.

Per questo motivo, abbiamo dibattuto a lungo per capire se sia lecito chiamare gli strumenti come ChatGPT dei «pappagalli stocastici». La risposta determina, in un certo senso, la visione generale del rapporto tra questa tecnologia e la nostra società. Infatti, abbiamo visto che chi sostiene che questa sia una definizione corretta vuole sottolineare che si debba sempre diffidare di ciò che viene generato automaticamente dalla macchina per evitare di cadere nell'inganno di pensare di poter sostituire troppo facilmente le persone con un chatbot. Chi invece sostiene che parlare di pappagallo non renda giustizia a ciò che gli algoritmi degli LLM fanno per davvero, tiene di più all'idea che queste tecnologie – pur con tutte le precauzioni del caso, dato che il loro modo di funzionare richiede attenzione e consape-

volezza degli sbagli che possono commettere – regaleranno alla società delle grandi opportunità di crescita.

Ci rendiamo conto, dunque, che il contesto sociale nel quale queste tecnologie e il dibattito su di esse si sviluppano è fondamentale per determinare il loro significato e forse anche il loro futuro. In particolare, si rende necessario approfondire due questioni che in questo capitolo sono state solamente accennate. La prima ha a che vedere col fatto che i criteri sui quali vertono i calcoli statistici degli LLM potrebbero essere compromessi a monte, già nel momento della programmazione e dell'allenamento di questi ultimi, con il risultato di filtrare ciò che possono e non possono dire. La seconda questione, abbastanza simile nella sostanza, si riferisce al fatto che, come vedremo nelle pagine che seguono, i pappagalli stocastici hanno la tendenza a ripetere e riprodurre i valori delle aziende che li creano, portando avanti un certo tipo di ideologia. Nel prossimo capitolo approfondiremo questi temi, al fine di restituire un quadro più ampio del contesto socio-economico e politico nel quale strumenti come ChatGPT vengono realizzati e utilizzati. In continuità con ciò che abbiamo fatto sin qui, continueremo la nostra riflessione sull'impatto che determinate narrazioni e rappresentazioni di queste tecnologie hanno su di noi, sul nostro mondo e sul futuro che intendiamo costruire.

Note al capitolo

1. I Large Language Models sono dei modelli linguistici sviluppati da reti neurali artificiali attraverso un'elaborazione, basata sul *deep learning*, di un'enorme mole di parametri pre-trattati a partire da testi linguistici.
2. <https://openai.com/>.
3. Il tema dell'automazione è sempre stato al centro del dibattito sul rapporto uomo-macchina nella società, si pensi ad esempio al luddismo di inizio Ottocento. Tuttavia, è un dibattito ancora attualissimo ed è stato ridestato di recente da uno studio controverso di Frey e Osborne (2013).
4. Questo, ad esempio, è quanto viene sostenuto nel podcast *Singapore – Futuro anteriore*, a cura di Massimo Cerofolini, scaricabile sulla piattaforma RAI Play <https://www.rai.it/ufficiostampa/assets/template/us-articolo.html?ssiPath=/articoli/2023/11/Online-il-podcast-Singapore-futuro-anteriore-a31d2d8d-d787-41ac-9c95-dc78abd7a836-ssi.html>.
5. <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60673>.
6. Brano originale inglese: «OpenAI's ChatGPT, Google's Bard and Microsoft's Sydney are marvels of machine learning. Roughly speaking, they take huge amounts of data, search for patterns in it and become increasingly proficient at generating statistically probable outputs – such as seemingly humanlike language and thought. These programs have been hailed as the first glimmers on the horizon of artificial *general* intelligence – that long-prophesied moment when mechanical minds surpass human brains not only quantitatively in terms of processing speed and memory size but also qualitatively in terms of intellectual insight, artistic creativity and every other distinctively human faculty […]. However, the human mind is not, like ChatGPT and its ilk, a lumbering statistical engine for pattern matching, gorging on hundreds of terabytes of data and extrapolating the most likely conversational response or most probable answer to a scientific ques-

tion. […] it seeks not to infer brute correlations among data points but to create explanations».

7. In realtà, il funzionamento statistico di questi strumenti – basato su una tecnica chiamata «embedding» che qui non ci interessa approfondire (ma, se lo si volesse fare, si veda Kaplan, 2024; oppure <https://www.technologyreview.com/2015/09/17/166211/king-man-woman-queen-the-marvelous-mathematics-of-computational-linguistics> – non ha come unità minima la parola, ma i cosiddetti «token», che per semplificare potremmo individuare nelle sillabe delle parole stesse. Ciò che ChatGPT e i suoi epigoni sono programmati per saper riconoscere è, sostanzialmente, la probabilità che, nelle catene sintattiche dei testi su cui vengono allenati, una certa sillaba ha di essere collegata alle altre, fino a formare parole e finanche frasi o discorsi che ci appaiono di senso compiuto.

8. I «big data» sono, appunto, la grande mole di dati di cui scrive Chomsky, necessaria a strumenti come quelli dotati di intelligenza artificiale per funzionare.

9. Brano originale inglese: «We do not seek highly probable theories but explanations; that is to say, powerful and highly improbable theories».

10. Le intelligenze artificiali *forti* (*strong AI*) – o Artificial General Intelligence (AGI) – sono un altro modo per nominare le intelligenze artificiali di livello umano. Per rimanere nel contesto del testo di Searle, esse sono dei programmi che non si possono più ritenere tali, ma delle menti vere e proprie (Searle, 1980).

11. Il test di Turing prende il nome dal suo inventore, Alan Turing, considerato universalmente il padre dell'informatica, e consiste in un esperimento che mira a stabilire se un'intelligenza artificiale sia diventata cosciente. Esso si basa sull'idea che se un essere umano non riconosce più la differenza fra un suo simile e una macchina, allora quest'ultima deve aver sviluppato un grado di intelligenza simile al nostro.

12. <https://www.treccani.it/vocabolario/allucinazione/>.

13. <https://www.spreaker.com/podcast/dataknightmare-l-algoritmico-e-politico--1977562>.
14. <https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web>.
15. Brano originale inglese: «In 2013, workers at a German construction company noticed something odd about their Xerox photocopier: when they made a copy of the floor plan of a house, the copy differed from the original in a subtle but significant way. In the original floor plan, each of the house's three rooms was accompanied by a rectangle specifying its area: the rooms were 14,13 21,11 and 17,42 square meters, respectively. However, in the photocopy, all three rooms were labeled as being 14,13 square meters in size […]. Xerox photocopiers use a lossy compression format known as JBIG2, designed for use with black-and-white images. To save space, the copier identifies similar-looking regions in the image and stores a single copy for all of them; when the file is decompressed, it uses that copy repeatedly to reconstruct the image. It turned out that the photocopier had judged the labels specifying the area of the rooms to be similar enough that it needed to store only one of them –14,13 – and it reused that one for all three rooms when printing the floor plan».
16. Brano originale inglese: «This analogy to lossy compression is not just a way to understand ChatGPT's facility at repackaging information found on the Web by using different words. It's also a way to understand the 'hallucinations' or nonsensical answers to factual questions, to which large language models such as ChatGPT are all too prone. These hallucinations are compression artifacts, but – like the incorrect labels generated by the Xerox photocopier – they are plausible enough that identifying them requires comparing them against the originals, which in this case means either the Web or our own knowledge of the world. When we think about them this way, such hallucinations are anything but surprising; if a compression algorithm is designed to

reconstruct text after ninety-nine per cent of the original has been discarded, we should expect that significant portions of what it generates will be entirely fabricated».

17. Nell'articolo in cui presenta le sue riflessioni sull'esperimento della stanza cinese, Searle riporta una serie di obiezioni che gli sono state mosse dagli esperti di intelligenza artificiale con cui si è confrontato prima di pubblicarlo.

18. Searle, 1980; Cole, 2023.

19. Contro l'idea che i sistemi di IA basati sui Large Language Models siano pappagalli stocastici si veda anche Ciotti (2023) e Roncaglia (2023).

20. Nel momento in cui scriviamo, l'ultima versione disponibile di ChatGPT è GPT-4, ma il suo funzionamento è simile e queste argomentazioni rimangono degne di essere prese in considerazione.

21. Brano originale inglese: «Graded across the UBE components, in the manner in which a human tast-taker would be, GPT-4 scores approximately 297 points, significantly in excess of the passing threshold for all UBE jurisdictions. These findings document not just the rapid and remarkable advance of large language model performance generally, but also the potential for such models to support the delivery of legal services in society».

22. <https://aisnakeoil.substack.com/p/gpt-4-and-professional-benchmarks>.

23. <https://verfassungsblog.de/colombian-chatgpt/>.

24. Brano originale inglese: «The judges did not use ChatGPT in an informed or responsible manner […]. The short answer is that they used ChatGPT as if it was an oracle: a trustworthy source of knowledge that did not require any sort of verification. While the judges were transparent about the fact that they used the tool and included quotation marks to distinguish the content produced by ChatGPT, their use was not informed nor responsible.

There are three main reasons why the way ChatGPT was used by the judiciary in these cases is very concerning for Colombians and beyond. First, the stakes in judicial rulings are too high – especially when human rights are involved – to justify the use of unreliable and insufficiently tested technologies. Due to the way that LLMs like ChatGPT are developed and operate, these tools tend to produce incorrect and imprecise answers and confuse reality with fiction. Even the CEO of OpenAI acknowledged in December 2022, that 'ChatGPT is incredibly limited […] it's a mistake to be relying on it for anything important right now'. Furthermore, due to structural reasons, it is unlikely that these problems of LLMs will be solved soon.

Secondly, ChatGPT's answers should not have been accepted and taken at face value, but rather contrasted with other more reliable sources. The judgment states that the information offered by ChatGPT would be 'corroborated'. However, there is no explicit trace in the text that allows us to conclude that judge Padilla or his clerk effectively checked whether ChatGPT's responses were accurate. In fact, I replicated the four queries posed by judge Padilla and the chatbot answered slightly differently, a result that is not surprising given how the tool works. Furthermore, when I prompted ChatGPT to provide examples of case law from the Constitutional Court that justified its answers, the chatbot invented the facts and *ratio decidendi* of one ruling, and cited a judgment that did not exist (inventing the facts and the verdict).

If ChatGPT and other LLMs currently available are evidently unreliable, since their outputs tend to include incorrect and false information, then judges would require significant time to check the validity of the AI-generated content, thereby undoing any significant 'time savings'. As it happens with AI in other areas, under the narrative of supposed 'efficiencies', fundamental rights can be put at risk.

Finally, there is a risk that the judges and its clerks over-rely on the AI's recommendations, incurring what is known as 'automation bias'».

CAPITOLO SECONDO

I problemi sociali, economici e politici sollevati da ChatGPT

Gli LLM e l'intelligenza artificiale in senso lato sono tecnologie che, come abbiamo visto, stimolano una riflessione che non si limita a concentrarsi sulle questioni tecniche più immediate, ma ci spinge verso orizzonti lontani, inducendoci a ragionare sul nostro futuro e a toccare tematiche profonde, finanche quelle sulla nostra natura di esseri umani. A questo proposito, nel precedente capitolo ci siamo interrogati sul livello di «intelligenza» che uno strumento come ChatGPT potrebbe raggiungere, mettendo in luce come questi argomenti siano in grado di toccare corde non solo razionali, ma anche emotive. Probabilmente questo è inevitabile, visto che ciò di cui si discute è il nostro rapporto con la tecnologia e il tipo di società che, a partire da esso, immaginiamo di sviluppare, con tutte le speranze, ma anche le paure, che questo comporta. Alla fine, comunque, abbiamo capito che i ragionamenti più tecnici

e filosofici non possono essere separati da quelli economici, sociologici e politici. In questo capitolo, dunque, intendiamo approfondire questi ultimi tre aspetti del dibattito sull'avvento della IA generativa, mostrando ciò che si dice a proposito dei suoi costi, dei suoi benefici e del suo potere di plasmare il mondo che verrà.

Data la natura di questo libro, intendiamo perseguire il nostro obiettivo da un punto di vista specifico, che è quello dell'analisi delle narrazioni che vengono portate avanti dai protagonisti di questo genere di discussioni, sui media, ma anche nel settore della pubblicistica scientifica. Così facendo, speriamo di contribuire a rendere più chiara la struttura retorica dei discorsi che circolano in questi contesti, mostrando come essi producano i loro effetti di senso e domandandoci, soprattutto, perché li perseguano: a chi giova che si parli in un certo modo dell'intelligenza artificiale? A quali conseguenze può condurre l'affermazione di un tipo di argomentazione a scapito di un'altra?

L'hype culture e i Large Language Models

Come nel capitolo precedente, anche in questo caso ci sembra opportuno partire da un articolo esemplare. Esso è stato pubblicato il 24 marzo 2023 sul «New York Times» da Yuval Noah Harari, un noto storico israeliano[1]. Harari è famoso per essere l'autore di *Sapiens* (2014) e *Homo Deus* (2016), due libri in cui racconta la storia dell'umanità dal paleolitico ai giorni nostri, immaginando anche il futuro in funzione della tecnologia che abbiamo sin qui sviluppato. In particolare, in *Homo Deus* si sofferma a lungo sull'in-

telligenza artificiale tratteggiandola come la più grande invenzione mai realizzata dagli esseri umani, visto che, a suo modo di vedere, offre allo stesso tempo la possibilità di migliorare la vita delle persone, ma anche di distruggerla. Più o meno, questa è anche la sua linea argomentativa sul «New York Times», dove parte sottolineando come il grande potere di strumenti quali ChatGPT consiste nell'abilità di manipolare e generare linguaggio, che lo stesso Harari definisce come il «sistema operativo» della nostra cultura, da cui, per citare le sue precise parole, emergono «i miti, le leggi, gli dei e il denaro, l'arte e la scienza, l'amicizia e le nazioni e il codice informatico»[2]. Una macchina che maneggi con perizia questa facoltà così centrale per il dominio dell'uomo sul cosmo, può assurgere al ruolo che quest'ultimo ha faticosamente costruito per sé al centro del creato. Il rischio – che attraverso un percorso storico Harari collega al lavoro già compiuto in questi anni dagli algoritmi dei social media – sarebbe innanzitutto quello di farci manipolare, lasciando che l'IA generi e filtri per noi i contenuti che ci consentono di formarci un'idea del mondo. Ma, citando una ricerca compiuta intervistando settecento professionisti del settore, lo stesso Harari si spinge oltre, ricordando che il 48% di loro avrebbe stimato una probabilità pari al 10% che lo sviluppo delle intelligenze artificiali, così per come funzionano già oggi, possa portare all'estinzione dell'umanità. Mentre ci sarebbe un 5% di probabilità che ciò possa avvenire anche per l'arbitrio di una futuribile «super intelligenza»[3].

Tali preoccupazioni sono spesso considerate tutt'altro che fantascienza. Oltre ad Harari, anche alcuni luminari delle intelligenze artificiali si sono espressi per ribadire la

necessità di un intervento immediato per guidare la ricerca ed evitare il peggio in futuro. Ad esempio, Yoshua Bengio, considerato uno dei pionieri del *deep learning*[4], ipotizza l'avvento di quella che definisce una *rogue AI*[5], una sorta di IA maligna, più o meno super intelligente, in grado di prendere autonomamente il controllo della società e di disporre delle questioni umane come meglio crede, oppure di fornire ai suoi malintenzionati progettisti la capacità di farlo. Soprattutto nel primo caso, il pericolo starebbe nella creazione di sistemi dotati di agentività e finalizzati a perseguire obiettivi propri, da quelli più evidentemente pericolosi di natura militare ad altri le cui conseguenze potrebbero risultare inaspettate, come nell'eventualità in cui essi venissero concepiti per riprodursi, auto-correggersi, evolversi o cercare rinforzi positivi per il proprio operato, a prescindere dai risvolti pratici di tutto ciò[6]. Il timore di Bengio, sostanzialmente, è che, come nel mito di Prometeo, l'umanità perda la capacità di dominare il «fuoco» della propria tecnologia, venendone sopraffatta. Per questo suggerisce che l'intelligenza artificiale debba essere utilizzata solo per comprendere la realtà, senza che vi possa agire direttamente.

Una posizione simile a quella dei due studiosi che abbiamo appena citato viene assunta anche da Geoffrey Hinton, un altro luminare del settore. Come riporta il «Guardian»[7], quest'ultimo si è licenziato da Google per poter parlare liberamente dei rischi legati alle IA. Hinton ha preso la sua decisione a causa dell'implementazione di GPT-4 da parte di OpenAI nel chatbot del motore di ricerca Bing di Microsoft, un fatto che ha ritenuto azzardato, dato che, per entrare nel mercato anticipando la concorrenza, non sarebbe stato seguito l'iter di sviluppo tradizionale di

questo genere di tecnologie, di solito più lento e controllato, che dovrebbe garantire la sicurezza dei loro utenti. In particolare, Hinton teme che, per quanto ancora molto limitata, la capacità di ragionamento di Bing sia destinata a crescere esponenzialmente, rischiando di divenire molto superiore alla nostra. Questo, ancora una volta, fornirebbe ai proprietari di questo sistema un potere troppo grande (ad esempio, consentirebbe loro di produrre grandi quantità di testi apparentemente verosimili e in grado di manipolare le menti delle persone), che un giorno potrebbe addirittura sfuggire dalle mani degli esseri umani.

Le argomentazioni come quelle di Harari, Bengio e Hinton incarnano ormai un vero e proprio modello discorsivo, che da tempo ha la tendenza a riproporsi sempre uguale, soprattutto sugli organi di stampa *mainstream*. Il fatto che questo genere di contenuti sia sempre più frequente, sembra denotare la volontà da parte dei suoi promotori di creare un *hype*. Questo termine, di difficile traduzione, è stato coniato in ambito economico per aiutare gli investitori a capire quale sia il momento giusto per puntare sull'avvento di una nuova tecnologia, al netto di tutto ciò che si dice attorno a essa[8]. Nel Cambridge Dictionary, l'*hype* è descritto come «l'azione di rendere qualcosa molto più entusiasmante di quanto realmente sia»[9] per mezzo di una narrazione esasperata, spesso fantasiosa e per questo accattivante. Diffuso anche in coppia con altri sostantivi, come nelle locuzioni *hype culture* e *hype cycle*[10], si tratta di un concetto oggi molto utilizzato per descrivere tutti quegli avvenimenti, oggetti o individui che subiscono una sovraesposizione mediatica di tipo patemico, tale da comprometterne una comprensione razionale e scevra da pregiudizi.

In sostanza, sembra esistere un pregiudizio diffuso in merito alle intelligenze artificiali, frutto della tendenza a immaginarle e rappresentarle come strumenti molto più simili agli esseri umani di quanto lo siano davvero. Questo si può evincere facilmente, analizzando il linguaggio di Harari, il quale parla di tecnologie come se fossero persone – e come forse ci si aspetterebbe dal personaggio cattivo di un film – che operano o potrebbero operare in futuro per hackerare e manipolare la società, impadronendosene[11]. Questo modo di ragionare impedisce di mettere a fuoco le reali responsabilità per gli eventuali danni che questi ultimi possono procurarci, con la conseguenza non solo di non riuscire a punire o sanzionare i colpevoli, ma anche di confondere tutti noi che li vorremmo utilizzare. Invece di fomentare un *hype* ansiogeno, dunque, bisognerebbe capire chi sono gli individui, le aziende e le istituzioni che contribuiscono a sviluppare le intelligenze artificiali generative, per quali scopi le producono, a quale prezzo e con quali conseguenze.

Critica dell'hype

Dato che si potrebbe parlare in maniera molto diversa dell'intelligenza artificiale, viene da domandarsi come mai molte testate giornalistiche e tanti studiosi contribuiscano all'affermazione di un certo *hype*. A questo proposito, l'evento forse più eclatante dal punto di vista mediatico si è verificato in seguito alla pubblicazione, il 22 marzo 2023, di una lettera aperta finalizzata a raccogliere firme per fermare per qualche tempo gli esperimenti e i test sugli LLM generativi come ChatGPT e le sue evoluzioni[12]. Questa fortunata

petizione – tra i circa trentatremila firmatari spiccano di nuovo figure come Yoshua Bengio, ma anche il multimilionario Elon Musk – è stata scritta e diffusa dal Future of Life Institute (FLI), un'organizzazione no-profit la cui missione, come si legge nel suo sito internet[13], consiste nel promuovere ricerche e informazioni utili a preservare il futuro dell'umanità dalle minacce esistenziali conseguenti a un utilizzo poco lungimirante di tre strumenti: l'intelligenza artificiale, la biotecnologia e le armi nucleari. Nella lettera si legge esplicitamente che i più potenti sistemi basati sulla IA dovrebbero essere sviluppati solo quando saremo sicuri che i loro effetti saranno positivi e i loro rischi gestibili. Al momento, però, secondo i membri del FLI, non è così e questo comporterebbe grossi pericoli, tra cui, già adesso, la diffusione ad ampio raggio di disinformazione, misinformazione e ideologie autoritarie. In futuro, invece, potrebbero essere addirittura creati dispositivi informatici in grado di dominarci, sostituirci e cambiare il corso della storia.

Anche questa lettera è frutto dell'*hype* che è nato attorno alla spettacolarizzazione dell'intelligenza artificiale, oppure c'è un fondo di verità nelle sue parole? Appare difficile dirlo, dato che le tecnologie più terrificanti, tra quelle immaginate dagli autori di cui stiamo discutendo, non esistono ancora. Ma è certo che scritti del tenore della lettera del FLI creano più problemi di quelli che si propongono di evitare, paventando un rischio esistenziale soltanto presunto e nascondendo, invece, i danni certi che, per far funzionare l'intelligenza artificiale, ogni giorno subiscono tanti lavoratori, l'ambiente e anche gli utenti finali di questo genere di tecnologia. Inoltre, essi influenzano l'opinione pubblica, spingendola a pensare

che il problema sia sempre rappresentato dai meccanismi dei dispositivi informatici e mai da chi li crea.

A questo proposito, sempre nella lettera del FLI, si sottolinea come gli attuali sistemi di intelligenza artificiale stiano diventando via via più competitivi nello svolgere compiti che prima venivano assegnati agli esseri umani, poiché richiedevano e richiedono ancora una forma di intelligenza «generale»[14]. Ma, come abbiamo già mostrato nel precedente capitolo, molti, riferendosi alle IA generative, parlano piuttosto di *stochastic parrots*, di pappagalli stocastici. Dunque, etichettare come di «livello umano» queste tecnologie significa nascondere i termini di un dibattito che è ben lungi dall'essere concluso. Tanto più che, come abbiamo visto, molti altri ritengono che quando questi strumenti superano le persone in qualche test di intelligenza, lo fanno perché vengono allenati facendo loro apprendere le risposte alle domande di quel genere di prove.

La narrazione secondo cui gli attuali sistemi generativi si starebbero approssimando a realizzare l'intelligenza artificiale generale nasce probabilmente da un articolo scritto per far conoscere al mondo le meraviglie di GPT-4, la quarta versione di ChatGPT, intitolato *Sparks of AGI* (Bubeck et al., 2023), a opera di un team di Microsoft Research, la divisione ricerca del principale finanziatore di OpenAI. Nella prima stesura di questo lavoro, si riportava una definizione del concetto di «intelligenza» attinta da un editoriale del «Wall Street Journal» (Gottfredson, 1997), dove si possono trovare riferimenti razzisti secondo i quali i latini e gli afroamericani sarebbero meno intelligenti dei bianchi (Bender, 2023)[15]. Quando è stata fatta loro notare la gaffe, agli autori e alle autrici non è rimasto che cancellare la nozione di

intelligenza che avevano utilizzato, lasciando sprovvisto il loro testo di quell'inquadramento concettuale preliminare che dovrebbe rappresentare una delle colonne portanti della sua argomentazione (Bender, 2023). Se l'idea che ChatGPT e i suoi epigoni siano vicini a realizzare il sogno dell'intelligenza artificiale generale di livello umano si basa su articoli di questo tipo, essa è ancora lungi dall'essere verificata.

SALAMI

Le posizioni critiche nei confronti dell'*hype* che si è creato attorno all'intelligenza artificiale, seppure con diverse sfumature, si fondano sostanzialmente sulla convinzione che esso sia il frutto della scarsa conoscenza tecnica di questo genere di strumenti, oppure che si tratti di una forma palese di pubblicità, finalizzata a far acquisire valore alle aziende e ai protagonisti dello sviluppo di dispositivi come ChatGPT e GPT-4.

Per contrastare questa retorica si è fatto strada un altro appellativo per riferirsi alle intelligenze artificiali in tutte le loro forme. Si tratta del termine «SALAMI», coniato da Stefano Quintarelli[16], informatico, imprenditore, ex parlamentare e curatore di un libro sulla stessa IA (Quintarelli, 2020): un acronimo che significa *Systematic Approaches to Learning Algorithms and Machine Learning*[17] e che, anche per il sottile gioco di parole che innesca, dovrebbe depotenziare la narrazione dell'*hype*. Un salame, in fondo, oltre a essere un insaccato, nella vulgata comune è una persona ingenua, secondo una sfumatura di significato che associata a una IA ritenuta dotata di intelligenza generale di livello

umano ricorda piuttosto quei pappagalli stocastici di cui abbiamo più volte scritto qui.

Secondo Domenico Quaranta, uno degli studiosi che condividono l'utilizzo di questo termine, SALAMI nasce dalla mente di Quintarelli proprio per prendere le distanze dalla nozione antropomorfica di «intelligenza» associata ai dispositivi informatici:

> Quintarelli è convinto che il primo *bias* della IA sia il suo nome, che implicitamente suggerisce la possibilità che le macchine sviluppino una qualche forma di coscienza, emozioni e personalità, arrivando a superare i limiti dell'umano. Rinunciare al nome che gli diede nel 1956 il matematico americano Marvin Minsky sarebbe dunque il primo passo verso una necessaria demistificazione dell'Intelligenza Artificiale[18].

Questa interpretazione del pensiero di Quintarelli, del resto, è giustificata da quanto scrive lui stesso sul proprio blog:

> Il nome «intelligenza artificiale» sottende un pregiudizio implicito che non consente una percezione aderente alla realtà. Al contrario, il nome favorisce l'intuizione che le macchine sviluppino una qualche forma di coscienza, emozioni, che acquisiscano una «personalità» simile a quella umana e, in definitiva, che superino i limiti umani e sviluppino un'identità, un senso del sé superiore a quello umano. Avete visto i film, conoscete la narrazione… Ma si tratta solo di dispositivi che estraggono correlazioni dai dati e le usano per fare previsioni e un sacco di cose molto utili. E come le calcolatrici che fanno le radici quadrate, possono farlo a una scala e a una velocità di gran lunga superiori alle prestazioni umane. Aggiungendo un po' di logica e di casua-

lità, possono anche esibire alcuni comportamenti interessanti e originali. Tuttavia, le macchine non hanno la minima idea di cosa sia la realtà. Al massimo, imitano un modello della realtà, quindi sono a due passi di distanza dalla realtà[19].

Vedendola così, in effetti, l'IA assume connotazioni del tutto diverse ed evidentemente l'*hype* che può produrre si sgonfia.

Criti-hype

Come abbiamo anticipato, nonostante il proliferare di ragionamenti come quello di Quintarelli, le strategie discorsive di Harari, Bengio, Hinton e del FLI, che contribuiscono a produrre un *hype* attorno all'intelligenza artificiale generativa e alle varie forme che l'IA ha acquisito nel mondo contemporaneo, cominciano oggi a diventare una sorta di genere discorsivo, tanto che qualcuno ha proposto di classificare i vari articoli, saggi, libri e conferenze su questi temi con un altro termine: *criti-hype*. Lee Vinsel, professore di Scienza, Tecnologia e Società al Virginia Polytechnic Institute, ha coniato questo neologismo ispirandosi a un'espressione di un famoso storico della tecnologia, David C. Brock[20]. Quest'ultimo parla infatti di *wishful-worry*[21], ovvero di un fenomeno molto simile a quello che stiamo analizzando in questa sede, definendolo così:

Elaborati da scrittori e conferenzieri di ogni genere, gli scenari di desiderio-preoccupazione sono ambientati in un futuro vago, a una distanza non ben definita dal nostro presente, e presentati come questioni di cui essere seriamente preoccupati

(«Dobbiamo occuparci di questo!»). In realtà, funzionano come piacevoli distrazioni da quelle che si potrebbero definire le nostre *reali agonie*. I desideri-preoccupazioni sono problemi che sarebbe bello avere, in contrasto con le reali agonie del presente[22].

Il *criti-hype* produrrebbe, in sostanza, un allarmismo poco significativo, basato su preoccupazioni per un futuro che non esiste, distogliendo invece l'attenzione dalle problematiche reali legate allo sviluppo della tecnologia nel presente. Questo avverrebbe quando il progresso della tecnica introduce nelle nostre società nuovi strumenti dalle funzionalità promettenti (si pensi, ad esempio, alle preoccupazioni espresse da Platone nel *Fedro* a proposito dell'avvento della parola scritta su un supporto simile al libro, che secondo il filosofo greco avrebbe potuto atrofizzare la memoria delle persone, trasferendo il sapere al di fuori dei loro cervelli): evidentemente, lo stesso schema si starebbe riproducendo con l'intelligenza artificiale generativa.

Brock chiama in causa anche la fantascienza che istigherebbe le persone a prendere in seria considerazione un immaginario che parla per eccessi e che, per rivolgersi al grande pubblico, sposterebbe l'attenzione su questioni morali dalla natura decisamente generica, ma comprensibili e condivisibili da tutti, elidendo dai propri discorsi argomenti scottanti ma più divisivi. Questo tipo di critica viene condiviso anche da Frank Pasquale nel suo libro sulle nuove leggi della robotica (Pasquale, 2020) che chiude proprio con un capitolo nel quale attacca l'industria dell'intrattenimento per il modo sbagliato in cui essa rappresenta l'intelligenza artificiale e che non è sostanzial-

mente lontano da quello descritto da Renato Giovannoli nel suo *La scienza della fantascienza* (2015). Qui, lo studioso italiano sottolinea come da sempre l'IA e le macchine umanoidi stimolino la fantasia di chi le vede come amichevoli aiutanti al nostro servizio, ma anche di quelli che le vedono come moderni Frankenstein pronti a ribellarsi ai loro prometeici creatori o come subdoli e pericolosi *doppelganger* che sfruttano la loro capacità di somigliarci, per manipolarci e circuirci. In quest'ottica, le parole di Harari – che nel suo articolo parla esplicitamente del fatto che il nostro futuro potrebbe diventare come quello tratteggiato in un film di fantascienza come *The Matrix* (di Lana e Lilly Wachowski, 1999), oppure come quello della lettera del FLI, dove si paventa un avvenire distopico molto simile a quello di *1984* (Orwell, 2021), utilizzando locuzioni come «fine della nostra civilizzazione» o «sviluppo di menti non umane che ci potrebbero soverchiare, superare per intelligenza, rendere obsoleti e sostituire» – assumono un significato diverso.

Ethics washing

La chiave di lettura del *criti-hype* è sicuramente interessante, perché riesce a spiegare come mai alcuni immaginari hanno più presa di altri nel discorso mediatico sulla tecnologia. Tuttavia, non fornisce una risposta esaustiva alla domanda che abbiamo posto all'inizio di questo capitolo, riguardo alle ragioni non solo culturali, ma anche sociali ed economiche, per cui una certa impostazione della riflessione sull'intelligenza artificiale generativa si sta

imponendo rispetto alle altre. A questo proposito, in un articolo scientifico, Tafani prova a suggerire un altro tipo di ragionamento, sostenendo che le grandi corporation che investono miliardi nello sviluppo della IA avrebbero dei tornaconti nel trasformare il discorso sull'etica applicata a questi strumenti in un grande «specchietto per le allodole» (Tafani, 2023).

Tafani, in sostanza, si riferisce a un altro tema molto dibattuto oggi a proposito dell'intelligenza artificiale in generale, dunque anche di quella che si basa sugli LLM come ChatGPT: quello dell'*ethics washing*. Con questa locuzione si allude di solito al fatto che le aziende o le istituzioni che operano in questo ambito fingano interesse per i risvolti etici del proprio lavoro, con l'unico scopo di guadagnare reputazione e migliorare la propria immagine, dato che questo fa aumentare il valore del loro brand[23]. Esse, spesso, finanziano direttamente la ricerca sui loro stessi prodotti e servizi da parte di filosofi morali, giuristi e altri studiosi affini, indirizzandola in maniera tale che si occupi di argomenti non troppo problematici per il buon andamento del loro business, ma di grande impatto culturale ed emotivo, come quelli trattati da Harari e dal FLI. In questo modo, dimostrando una certa sensibilità verso valori un po' generici ma di portata quasi universale, tali soggetti illuderebbero le anime candide di cui parla la stessa Tafani, potendo continuare a lavorare come sempre.

Alcuni hanno inquadrato questa manovra come una scelta di marketing, che sostanzialmente trasformerebbe l'etica della tecnica in una merce. Essa sarebbe la conseguenza dell'affermarsi di un modello culturale «lungotermista»[24]. Con questo termine, ci si riferisce a una narrazione che

mira a far sembrare alcune eventuali minacce di un lontano futuro, come i rischi esistenziali paventati dal FLI e da Harari, qualcosa di molto più imminente e, di conseguenza, più preoccupante di quanto lo sia davvero. Un diffuso consenso dell'opinione pubblica a questa visione garantirebbe facili giustificazioni di accentramenti di potere economico e politico nelle mani degli unici attori sociali in grado di poter creare materialmente delle soluzioni efficaci. In questo modo, le aziende detentrici di quelle stesse tecnologie che si dimostrano problematiche diventerebbero le stesse che hanno i mezzi per risolverlo, chiudendo un cerchio ben poco virtuoso, agli occhi di chi le critica, ma sicuramente molto redditizio.

Una delle componenti socio-economiche che rende questa situazione inaccettabile sarebbe, tra l'altro, la condizione di scarso finanziamento in cui versano le università, che impedirebbe loro di svolgere ricerche indipendenti e di rivestire il ruolo di controllore disinteressato al servizio delle pubbliche istituzioni (Tafani, 2023).

Di fronte ad analisi di questo tipo, qualcuno si potrebbe chiedere se sia giusto condannare a priori il desiderio delle grandi aziende di occuparsi di questioni etiche relative al proprio operato e se le denunce come quelle degli studiosi di cui abbiamo scritto o del FLI non siano comunque qualcosa di positivo, visto che senza di esse il panorama della critica all'intelligenza artificiale perderebbe rilevanza mediatica, lasciando il campo all'entusiastica propaganda di chi la deve vendere. Ma sempre Tafani, nell'articolo che qui stiamo commentando, sostiene che per evitare di scadere nella pura retorica sarebbe necessario realizzare un impianto normativo concreto, promulgare leggi in grado

di garantire la salvaguardia dei diritti di tutti coloro che lavorano nell'ambito dell'intelligenza artificiale, che la utilizzano o che ne subiscono il ricorso da parte delle varie istituzioni che puntellano la nostra vita civile, anche se questo potrebbe comportare delle perdite economiche per le aziende che operano in questo settore o una diminuzione del potere di chi si serve di queste tecnologie. Come vedremo nelle prossime pagine, il modo di funzionare della IA crea già oggi diversi problemi ambientali (Crawford, 2021), di sfruttamento dei lavoratori (Casilli, 2019), di ingiustizie e di discriminazioni (O'Neil, 2016; Chun, 2021): di questo, dunque, ci si dovrebbe occupare, senza bisogno di sconfinare nella fantascienza o di guardare troppo avanti.

Al momento, le poche regolamentazioni che, in giro per il mondo, esplicitano la necessità di affrontare questi temi, come l'Artificial Intelligence Act[25] proposto dalla Commissione Europea nel 2021 e da poco promulgato, richiedono a chi sviluppa o utilizza la stessa intelligenza artificiale per lo svolgimento di pubbliche funzioni di autocertificare la sua adeguatezza a una serie di principi etici che dovrebbero tutelare i cittadini. Questo, però, secondo molti, non basterebbe, come si evince dall'articolo di Tafani che qui stiamo commentando.

Oltre il criti-hype

Concordando evidentemente con le posizioni critiche che abbiamo esposto fin qui, Kapoor e Narayanan, in un articolo pubblicato su internet[26], si riferiscono ancora una volta alla nota lettera del FLI di cui abbiamo scritto, defi-

nendola «fuorviante» e confezionando un'ironica tabella a doppia entrata. Da un lato vi inseriscono i tre problemi più significativi che, a loro modo di vedere, oggi ci vengono posti dall'utilizzo dell'intelligenza artificiale: la misinformazione, l'impatto di questa tecnologia sul lavoro e quello sulla sicurezza dei dati personali. Dall'altro lato, mostrano la differenza tra come questi temi vengono trattati da chi vuole creare *hype* e i rischi reali di cui si occupano gli studiosi.

Così facendo, Kapoor e Narayanan polemizzano con coloro che sostengono che dietro ai sistemi di IA che producono contenuti non aderenti alla realtà, spesso razzisti o oltranzisti, ci sia la volontà malevola di qualcuno – o addirittura della stessa IA, trasformatasi in un'intelligenza artificiale generale – di propagare la disinformazione e di dare origine a qualche distopia. Piuttosto, essi sottolineano come il vero problema sia che gli individui, utilizzando ingenuamente i generatori di testi come ChatGPT, hanno la tendenza ad affidarvisi troppo, come abbiamo già scritto nel primo capitolo, fino a perdere il proprio spirito critico e a smettere di pensare che anche questo tipo di informazione debba sempre essere verificato. Inoltre, i due autori parlano di come la narrazione di un futuro del lavoro nel quale le macchine saranno in grado di svolgere tutte le nostre attività e ci sostituiranno quasi completamente sia priva di solide basi empiriche, mentre sono molto attestati gli studi che mostrano come le grandi aziende di tutto il mondo, dunque anche quelle che operano nell'ambito dell'intelligenza artificiale, erodano il potere e i diritti dei lavoratori, servendosi a questo scopo anche degli strumenti che questi ultimi contribuiscono a sviluppare[27]. Infine,

Kapoor e Narayanan sostengono che il continuo parlare di rischi esistenziali e di messa in sicurezza di IA che, altrimenti, prenderebbero il controllo del mondo, impedisce in realtà di concentrarsi sui problemi normativi legati alla sicurezza dei dati, al diritto alla privacy e alla tutela del copyright che tutte le piattaforme digitali, dunque anche strumenti come ChatGPT, pongono per loro natura.

Questi tre gruppi di problemi sono trattati in maniera estremamente sintetica nell'articolo che abbiamo appena citato. Approfondiremo alcuni di essi nelle pagine che seguono, ma qui ci interessa soprattutto mettere in evidenza la retorica di questo genere di discorsi. Essi, ormai, proprio come quelli sull'*hype*, stanno diventando sempre più frequenti, quando si tratta di parlare dell'intelligenza artificiale in generale e di quella generativa in particolare. Autori come Kapoor e Narayanan, Tafani, Vinsel, Brock, Quintarelli e tutti gli altri che abbiamo citato in questo capitolo si battono per decostruire una narrazione che ritengono, per l'appunto, «fuorviante», proponendone altre che ci sembrano più concrete e significative.

L'intelligenza artificiale generativa e il futuro del lavoro

Come abbiamo visto, al di là dell'*hype*, uno dei temi più rilevanti da affrontare con l'avvento dei sistemi di intelligenza artificiale, anche generativa, è quello del loro rapporto col mondo del lavoro umano. Abbiamo altresì mostrato che i più allarmisti ritengono che questi dispositivi siano già in grado di sostituirci in molte mansioni e che in futuro lo saranno sempre di più. Ma anche chi cri-

tica queste posizioni, ritenendole esagerate, riconosce che queste tecnologie stanno portando alcuni cambiamenti che devono essere valutati.

A questo proposito, è esplicativa del contrasto di opinioni che stiamo tratteggiando la contrapposizione tra le posizioni espresse da Eloundou e un gruppo di ricercatori di OpenAI, OpenResearch e della Pennsylvania State University in un articolo intitolato *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models* (Eloundou et al., 2023), e le riflessioni del sociologo Antonio Casilli, molto critiche nei confronti di questi studiosi[28].

Nella maniera tipica dei discorsi che producono *hype*, Eloundou e i suoi colleghi scrivono:

> La nostra analisi indica che circa il 19% dei posti di lavoro ha almeno il 50% delle mansioni esposte se si considerano le capacità attuali dei modelli [di LLM] e gli strumenti che si può prevedere che saranno costruiti su di essi. Le valutazioni umane indicano che solo il 3% dei lavoratori statunitensi è esposto a LLM per oltre la metà delle proprie mansioni, se si considerano le capacità linguistiche e di codice esistenti, senza software o modalità aggiuntive. Tenendo conto di altri modelli generativi e tecnologie complementari, le nostre stime indicano che fino al 49% dei lavoratori potrebbe avere metà o più dei propri compiti esposti agli LLM. […] Il contributo principale di questo lavoro è quello di fornire una serie di misurazioni del potenziale di impatto degli LLM e di dimostrare come si possono utilizzare gli LLM per sviluppare tali misurazioni in modo efficiente e su scala. Inoltre, mostriamo il potenziale di uso generale degli LLM. Se «le GPT sono GPT», la traiettoria

finale dello sviluppo e dell'applicazione degli LLM può essere difficile da prevedere e regolamentare per i responsabili politici. Come per altre tecnologie di uso generale, gran parte del potenziale di questi algoritmi emergerà in un'ampia gamma di casi d'uso economicamente validi, compresa la creazione di nuovi tipi di lavoro (Acemoğlu, Restrepo, 2018). La nostra ricerca serve a misurare ciò che è tecnicamente fattibile ora, ma necessariamente non coglierà l'evoluzione del potenziale d'impatto degli LLM nel tempo[29].

La struttura dell'argomentazione di questo articolo è ancora una volta ben codificata, riprendendo quella di molti altri testi di questo genere: data la potenza dell'intelligenza artificiale – in questo caso di quella che utilizza gli LLM – essa già adesso può svolgere molti lavori umani. L'importanza dei numeri e delle percentuali non è da trascurare, dato che fornisce un'aura di autorevolezza, di realismo e di scientificità al discorso. Infine, il fatto che la rapidità dello sviluppo tecnologico non sia stimabile con precisione, pur essendo facilmente percepibile nel suo andamento storico, lascia immaginare che certe cifre, già di per sé impressionanti, in futuro lo saranno ancora di più. Tutto questo, che è da accettare come un dato di fatto, non si può e non si deve fermare, per chi porta avanti questo tipo di discorsi. Alle istituzioni, dunque, non resterebbe che prenderne atto, facendo in modo che, come è sempre avvenuto col progresso, tutto cambi per il meglio.

Ma Antonio Casilli, un sociologo molto esperto di questi temi, in un'intervista volta a commentare i contenuti di questo articolo, mette subito in guardia tutti[30]. Egli,

in pratica, ritiene di trovarsi ancora una volta di fronte al fenomeno dell'*hype*. Scimmiottando il più famoso articolo di Frey e Osborne (2013) su come il computer e l'informatizzazione avrebbero trasformato il mercato del lavoro, i quattro autori di questo studio su ChatGPT e sui suoi epigoni si baserebbero su fonti incerte, dati poco verificabili e metodologie astruse. Inoltre, essi vogliono rappresentare le varie professioni come il risultato di una sommatoria di tutti i compiti che le lavoratrici o i lavoratori devono svolgere. Ma questa visione sarebbe estremamente riduzionistica, presupponendo che le persone e le loro competenze siano qualcosa di modellizzabile, riproducibile come se si trattasse di un software in cui si possono installare determinate funzioni. Seguendo questo ragionamento, l'individuo nella sua complessità viene svalutato. Di lui rimangono, per l'appunto, solo un ruolo e una sfera d'azione, che si attiva con dei dati in input, producendo un output grazie a un algoritmo. Una rappresentazione, questa, che sarebbe evidentemente utile alle aziende che realizzano strumenti informatici per l'automazione del lavoro e che dovrebbe ingolosire anche quelle che li acquistano, con la prospettiva di aumentare la loro produttività. Ma, negli intenti dei fautori di questo genere di argomentazioni, esse dovrebbero convincere anche gli stessi lavoratori, i quali dovrebbero accogliere con gioia il fatto di essere affiancati da macchine capaci di sgravarli dai loro compiti più noiosi e ripetitivi: così facendo, infatti, essi si potrebbero dedicare a mansioni più soddisfacenti, senza temere di essere per forza sostituiti (Acemoğlu, Restrepo, 2019).

La cultura *task-oriented* e la promessa di un futuro del lavoro migliore perché automatizzato è un tema molto caro

a Casilli. Egli vi ha dedicato un libro, *En attendant les robots* (Casilli, 2019), in cui ha sostenuto che, come nel titolo della famosa opera teatrale di Samuel Beckett, *Aspettando Godot*, la speranza dell'avvento dell'automazione totale grazie all'intelligenza artificiale, così come è spesso raccontata, è destinata a non avverarsi mai. Essa rimarrà sempre parziale e imperfetta, ma quella che oggi si sta affermando nell'ambito del digitale produce sofferenza, precariato e ingiustizie di diversa natura, determinando la comparsa sul mercato globale di nuove forme di schiavitù.

Questa situazione è descritta plasticamente, ad esempio, da un'inchiesta del «Times»[31] proprio sull'operato di OpenAI, che ha utilizzato stuoli di lavoratori e di lavoratrici in Kenya per ripulire i *training set* di ChatGPT da dati legati ad abusi sessuali su minori, stupri, incesti e omicidi, allenando il dispositivo a riconoscerli e classificarli correttamente. Queste persone, però, pur svolgendo un compito molto delicato e usurante dal punto di vista psicologico, sono state sottoposte a turni massacranti e pagate meno di due euro l'ora, senza dotarle di un adeguato supporto.

Questi lavori, pur esistendo da molti anni, hanno avuto un'impennata con l'affermazione sul mercato delle intelligenze artificiali. Si tratta di «professioni» – pomposamente definite di Reinforcement Learning from Human Feedback (RLHF), apprendimento supplementare tramite il feedback umano, a cui uno studioso tecno-entusiasta come Kaplan si riferisce tratteggiandole come il mestiere delle «Mary Poppins» dei giorni nostri, poiché consisterebbero nel «fare da tutor ai loro allievi elettronici, spiegando loro ogni sfumatura di come si deve interagire con gli umani»

(Kaplan, 2024, posizioni nel Kindle 2633-2634) – che obbligano a svolgere compiti iper-parcellizzati, micro-transati e spesso estremamente de-mansionanti, in cambio di una paga misera che costringe a ritmi disumani per ottenere il minimo indispensabile per vivere.

In nome del progresso e del mito dell'automazione, dentro a grandi capannoni come quelli di Amazon nelle periferie delle nostre città (Delfanti, Posada, 2021), oppure nelle sedi dei sub-fornitori delle big tech companies nel sud del mondo, esiste una società fatta di persone – «cottimisti sottopagati assunti e licenziati sulle piattaforme digitali» (Casilli, 2019) – che soffrono e che stanno lavorando, in effetti, per essere sostituite. Esse, infatti, accettano di operare per procurare alla IA i dati necessari per svolgere i loro stessi micro-compiti, in modo che un giorno quest'ultima sarà capace di prendere o riporre un prodotto da uno scaffale, impacchettarlo e spedirlo, oppure saprà riconoscere un contenuto scabroso su internet, evitando di farvi riferimento nelle sue interazioni con gli esseri umani. Per fare in modo che questo avvenga, inoltre, ai lavoratori viene spesso chiesto di svolgere questi stessi compiti come se fossero dei robot, indossando telecamere e pesanti tecnologie che consentano di trasformare in dati ogni loro azione, oppure «processando» grandi moli di testi e immagini problematici, in un circolo vizioso molto lontano dalla visione edulcorata dei promotori della piena automazione, i quali dichiarano, come OpenAI, di volersi assicurare che «l'intelligenza artificiale generale riesca a giovare a tutta l'umanità»[32].

Ambiente, diversità ed equità

Che lo sviluppo della IA da parte delle grandi aziende non sia quell'attività filantropica che esse tentano di descrivere, ma sottintenda tutti i conflitti di interesse e di potere di cui è puntellata la nostra società, si evince anche dalle vicende di Timnit Gebru e Margaret Mitchell, due ricercatrici esperte di etica dell'intelligenza artificiale che sono state licenziate da Google per non aver voluto ritirare un articolo che espone diverse problematiche legate all'utilizzo dei chatbot come Google Bard, i quali funzionano secondo principi molto simili a ChatGPT[33].

L'articolo in questione è già stato citato in questo libro e si intitola *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* (Bender et al., 2021). Esso è stato pubblicato grazie all'intervento di Emily Bender, coautrice e professoressa di Linguaggio computazionale alla University of Washington, che si è rivolta al Massachusetts Institute of Technology per cercare di promuoverlo nonostante il tentativo di censura da parte di Google[34]. Al suo interno, si leggono alcune riflessioni critiche sull'impatto ambientale e sociale delle intelligenze artificiali generative. La ricerca spazia da uno studio approfondito sulle enormi emissioni di anidride carbonica che vengono prodotte per allenare strumenti come BERT, il modello linguistico della stessa Google, fino alle analisi sul razzismo, sul sessismo e sul classismo strutturalmente incorporati nei dataset da cui attingono questi dispositivi, con tutte le conseguenze che questo comporta.

Per quanto riguarda i problemi ambientali, Bender, Gebru, McMillan-Major e Mitchell riprendono una linea

argomentativa molto diffusa tra gli studiosi di intelligenza artificiale (Crawford, 2021; Becker, 2023). Esse sottolineano come, in media, un essere umano produca cinque tonnellate di anidride carbonica all'anno, mentre l'allenamento di una IA che usa modelli come BERT ne richiede duecentottantaquattro. Inoltre, la maggior parte dell'energia utilizzata per queste operazioni non proviene da fonti rinnovabili. Dunque, sarebbe necessario promuovere una sorta di *green AI*, così da favorire il ricorso a strumenti che ottimizzino il rapporto costi ambientali-benefici di questo genere di tecnologie e che appartengano ad aziende o istituzioni che utilizzino delle metriche per calcolarlo. Tali soggetti, inoltre, dovrebbero operare in zone del mondo che siano *carbon friendly* e dovrebbero anche spendersi per promuovere politiche di lobbying e di sensibilizzazione dell'opinione pubblica verso queste questioni ecologiche.

Un'altra problematica importante sollevata dal funzionamento della IA generativa è quella della capacità dei database su cui quest'ultima si allena di rendere la pluralità delle posizioni che si possono esprimere sui temi su cui essa viene sollecitata a «esprimersi». Abbiamo già osservato nel capitolo precedente che la strategia attraverso cui dispositivi come ChatGPT si relazionano col mondo di cui parlano è di tipo *lossy*, vale a dire che questi strumenti non processano tutti i dati che abbiamo a disposizione sul mondo stesso, ma solo una parte, divenendo così imprecisi e rischiando di produrre allucinazioni. Per risolvere questo problema, forse qualcuno potrebbe pensare che basti procurarsi database molto capienti, ma Bender, Gebru, McMillan-Major e Mitchell dissentono. Esse sottolineano che avere accesso a grandi moli di dati non significa

entrare in contatto con quelli giusti. Piuttosto, si corre il rischio di servirsi di informazioni che si riferiscono a una visione egemonica delle cose. Ad esempio, è risaputo che se per allenare l'IA si usa ciò che si trova su internet, l'accesso alla rete è molto diseguale e privilegia chi se lo può permettere[35]. Inoltre, se nella stessa internet, come è avvenuto per allenare GPT-2, la seconda versione di ChatGPT, si utilizzano i contenuti di Reddit, la nota piattaforma americana per l'aggregazione di notizie e la discussione tra i suoi utenti, le statistiche indicano che questi ultimi sono per il 67% uomini e per il 64% giovani tra i 18 e i 29 anni. Se si ricorre a Wikipedia, come si fa sempre con gli LLM, solo il 15% delle persone che vi scrivono sono donne. Se, come si è fatto per GPT-3, si filtrano i dati di Common Crawl[36] per eliminare automaticamente parole sconvenienti, maleducate o sessiste, si rischia di cancellarne alcune che in certi contesti non sono tali, come nel caso del termine *twink* nella comunità LGBTQ+[37]. Se, per parlare dell'attualità, si prendono in considerazione solo i siti dei giornali e della stampa in generale, quasi certamente non si rappresentano tutti quei movimenti non violenti che portano avanti le proprie cause in maniera poco notiziabile. Se non si ri-allenano continuamente le IA – e di solito questo non viene fatto spesso, dati i costi, sia in termini economici che ambientali – si rischia di non riprodurre in tempo reale i rapidi cambiamenti nel modo di vedere le cose al di fuori dei database utilizzati a questo scopo (le autrici di cui stiamo scrivendo fanno esplicitamente riferimento, in questo caso, a come l'ondata di emozione legata al movimento Black Lives Matter abbia indotto una super-produzione improvvisa di con-

tenuti in suo favore su internet, con una velocità che non avrebbe potuto essere intercettata da nessuna intelligenza artificiale allenata in un periodo anche solo di poco antecedente).

Infine, le studiose di cui qui ci stiamo occupando sottolineano l'importanza di contrastare i *bias* della IA generativa. Esse ricordano che gli LLM incorporano pregiudizi nei confronti di certi tipi di persone (BERT di Google, ad esempio, produrrebbe contenuti più negativi nei confronti dei disabili rispetto ad altri strumenti simili), stereotipi (sempre BERT tenderebbe ad associare più facilmente alla malattia mentale l'abuso di droga, gli homeless e coloro che si servono impropriamente di armi da fuoco), fake news e altri dati di natura «tossica». Questo problema è consustanziale a questo genere di tecnologia, tanto che anche quando si tenta di combatterlo, ciò non fa che aumentare l'attenzione per la tutela di certi gruppi, sfavorendone altri (Bender et al., 2021: 615). Per tale ragione, sarebbe necessario dichiarare sempre con quali criteri si opera per ridurre i *bias* del proprio sistema di IA generativa, investendo fondi per certificare il proprio operato dal punto di vista etico e prendendosi la responsabilità legale dell'eventuale malfunzionamento del dispositivo. Del resto, se ChatGPT o Bard, come spesso accade, vengono impiegati dai loro utenti per farsi un'idea sul mondo circostante e sulle opinioni che al proposito vengono espresse nel proprio contesto socioculturale, deve risultare evidente in che modo viene costruita questa rappresentazione.

I nefasti effetti psicologici della IA generativa

Molti dei problemi sollevati da Bender, Gebru, McMillan-Major e Mitchell hanno a che vedere con i possibili effetti culturali del massiccio ricorso a strumenti come ChatGPT e i suoi simili, nel caso in cui essi non si rivelino pluralisti, bilanciati e rispettosi nel modo di parlare delle minoranze, eccetera. I testi prodotti da questi dispositivi, infatti, potrebbero influenzare il pensiero dei loro destinatari, riproducendo modelli di ragionamento e posizioni disfunzionali al vivere civile, oppure eticamente criticabili. Naturalmente, il fatto che questo tipo di contenuti circoli liberamente all'interno di una società non significa che i membri di quest'ultima debbano essere razzisti o totalitaristi, dato che ognuno è libero di formarsi un'opinione diversa e di respingere ciò che gli viene detto. Questo discorso, però, si fa più controverso quando a interagire con un'intelligenza artificiale generativa sono i minorenni o persone con qualche problema psicologico.

A questo proposito, è interessante citare i contenuti di un'altra petizione[38], lanciata dai suoi sostenitori proprio per reagire alla lettera del FLI sulle minacce esistenziali della IA della quale abbiamo tanto scritto in questo capitolo, sfruttandone l'*hype* e richiamando l'attenzione su un problema più concreto: quello dei suicidi istigati dai chatbot come ChatGPT. I promotori e i firmatari di quest'altro testo si riferiscono al noto caso di un giovane che si è tolto la vita dopo aver chiesto aiuto a uno di questi strumenti, invece di rivolgersi a una figura umana competente[39]. A loro modo di vedere, è necessario riflettere sull'effetto che una IA «travestita da uomo» può sortire sulla psiche, soprattutto quella

dei più deboli[40]. In certi casi, infatti, le competenze linguistiche degli LLM, che sono in grado di simulare le riflessioni delle persone in carne e ossa, possono trasformarsi in pericolose armi manipolatorie, soprattutto quando confermano a qualcuno le sue visioni disforiche della realtà, come del resto si può facilmente verificare, dato che a interrogare questi strumenti, proponendo i temi e il tenore delle discussioni, sono proprio coloro che vogliono sentirsi rispondere in un certo modo.

Questo genere di riflessioni, in sostanza, riprende un filone di studi molto conosciuto nell'ambito dell'intelligenza artificiale, vale a dire quello della *deceitful AI*[41] (Natale, 2021). Gli autori che se ne occupano riflettono su come i progettisti di questa tecnologia cerchino di nascondere il fatto che i loro dispositivi siano macchine, tentando di «camuffarli» da esseri umani. Essi si interrogano anche sugli effetti di queste scelte di design: se le persone vengono deliberatamente indotte a confondersi, scambiando un chatbot per un interlocutore degno di fede, allora è necessario renderle più consapevoli circa il significato dell'esperienza di interazione linguistica che viene proposta loro, sia per mezzo di disclaimer più efficaci, sia attraverso corsi di educazione informatica.

L'intelligenza artificiale, inoltre, dovrebbe essere regolamentata per legge. Offrire una spiegazione facilmente comprensibile sui limiti di questi prodotti dovrebbe essere responsabilità delle aziende, le quali dovrebbero essere perseguite in caso di inadempienza. A questo proposito, Replika, un chatbot basato su un sistema GPT, è stato bloccato in Italia dal Garante della privacy il 2 febbraio 2023, per via dei numerosi dubbi sul modo in cui

esso presentava i suoi servizi. Creato dalla Lukas Inc., esso è nato per rispondere al crescente bisogno di aiuto psicologico a causa della solitudine al tempo dei social media. Negli app store da cui può essere scaricato, viene presentato come un chatbot capace di migliorare l'umore e il benessere emotivo dei suoi utenti, aiutandoli a comprendere più a fondo i loro pensieri e i loro sentimenti, a tenere a bada lo stress, calmare l'ansia, perseguire i propri obiettivi essendo sorretti da una forma di pensiero positivo, socializzare, cercare l'amore, eccetera. Tuttavia, numerose evidenze portate alla luce dagli utenti stessi, dalla stampa e dalle autorità competenti hanno dimostrato che le sue risposte erano assolutamente inidonee rispetto al grado di sviluppo e auto-consapevolezza dei minori a cui il servizio si indirizzava[42]. Inoltre, per indurre il desiderio di servirsene più a lungo, Replika era dotato di meccanismi di *gamification*[43], una sorta di premi che sbloccavano interazioni più approfondite e soddisfacenti con il dispositivo, in funzione del tempo passato a utilizzarlo. Questo, naturalmente, poteva provocare qualche forma di dipendenza nei soggetti più condizionabili. Perciò, ora questo sistema di intelligenza artificiale ha un meccanismo di verifica dell'età di chi vi entra in contatto e può essere usato solo da maggiorenni, che danno o meno il consenso al trattamento dei loro dati personali attraverso il contratto di utilizzo.

Sempre a proposito di questi temi, può essere utile parlare di Wisa[44], un altro chatbot adottato in Gran Bretagna, nel sistema sanitario nazionale, per fare il *triage* dei pazienti minorenni con possibili problemi psichiatrici[45]. Questo strumento utilizza la tecnica psicoterapeutica detta Cognitive Behavioral Therapy (CBT) ed è supportato da evi-

denze cliniche raccolte per mezzo di ricerche *peer-reviewed*, condotte in alcune fra le più importanti università al mondo come quelle di Cambridge, Harvard e Washington, che ne dimostrano l'efficacia. L'idea, sostanzialmente, è di sostituire il lavoro di una persona specializzata con quello della macchina, che viene descritta dai suoi produttori come «an artificial intelligence-based emotionally intelligent service», dunque un servizio basato sull'intelligenza artificiale che è capace di comprendere le emozioni. Tutto questo, in qualche modo, insieme alla sua capacità di utilizzare il linguaggio naturale e di simulare interazioni «umane», dovrebbe rendere accettabile il ricorso a uno strumento del genere al posto del giudizio di un medico. Ma, come scrive uno studioso di IA come Pasquale (2020), introdurre questo genere di tecnologia negli ospedali, magari per risparmiare sul costo del personale e per velocizzare i servizi, potrebbe instradarci verso un futuro in cui i più abbienti, recandosi in strutture private, potranno ancora permettersi di consultare professionisti in carne e ossa, magari supportati dalla stessa intelligenza artificiale, mentre chi non avrà le stesse possibilità economiche dovrà accontentarsi di interagire solo con un chatbot. Certo, se quest'ultimo funzionerà a dovere e non produrrà allucinazioni (che però, come abbiamo visto, sono consustanziali all'architettura di questo genere di sistemi), magari tutto ciò non creerà problemi, ma ancora una volta è lecito pensare che dietro all'idea di avallare procedure del genere ci sia quella cultura *task oriented* che abbiamo avuto già modo di criticare nelle pagine precedenti: come nel mondo del lavoro, anche in quello della salute una rappresentazione povera e schematica dell'essere umano di tipo behaviorista –

l'idea che quest'ultimo sia come una macchina che risponde automaticamente agli stimoli che gli vengono mandati da un sistema di intelligenza artificiale che deve testare la sua salute psichica – giustificherebbe il ricorso alla macchina stessa per surrogare le interazioni tra persone in carne e ossa. In effetti, questo modello di pensiero è molto radicato nella nostra società e, come abbiamo mostrato, induce tanti individui a considerare IA come Wysa e Replika delle soluzioni efficaci per trattare le fragilità umane, nonostante i ragionevoli dubbi che tutto questo può sollevare.

Conclusioni

I problemi sociali, economici e politici portati dall'avvento di strumenti come ChatGPT sono sicuramente più numerosi e sfaccettati di quelli che abbiamo presentato in queste pagine, ma per gli obiettivi che ci siamo posti in questo libro la carrellata di questioni che abbiamo affrontato ci sembra sufficiente. Come abbiamo anticipato, infatti, ciò che ci premeva era mostrare i diversi modi in cui può essere narrato il senso dell'intelligenza artificiale generativa, così come viene sviluppata e utilizzata oggi. A questo proposito, risulta evidente che al momento si confrontino due grandi modelli di pensiero: uno più entusiastico e focalizzato sul futuro, intento a immaginare come questa tecnologia, se usata correttamente, migliorerà le nostre vite; l'altro più polemico, concentrato sul presente e deciso a mettere a nudo le ingiustizie e gli errori che si stanno compiendo in nome di un progresso di stampo capitalistico che potrebbe non rivelarsi tale. Noi, per tutte le ragioni che abbiamo

riportato, sentiamo di condividere le preoccupazioni di chi porta avanti questo secondo tipo di argomentazioni.

In ogni caso, nessuno può eludere il problema politico sollevato dall'intelligenza artificiale, che in fondo è lo stesso che ogni società deve affrontare quando sviluppa una tecnologia che promette di radicarsi al suo interno, producendo effetti che devono essere governati. In questo senso, Winner, in un articolo molto citato negli studi sulla scienza e sulla tecnica (Winner, 1980), propone di considerare l'esempio dei ponti costruiti dall'architetto Robert Moses fra gli anni Venti e Settanta del Novecento, per collegare il distretto di Long Island agli altri della città di New York. Essi sono stati progettati volutamente per essere bassi, al fine di bloccare il passaggio dei bus urbani e disincentivare la circolazione di alcuni ceti sociali non sufficientemente abbienti per permettersi un'automobile privata. Queste persone non avrebbero dovuto raggiungere facilmente la costa, in modo che il valore immobiliare delle case in quella zona rimanesse alto e ne potessero godere solo i ricchi. Spesso, dunque, sono motivazioni politiche quelle che ispirano la particolare organizzazione di un sistema tecnico o di un dispositivo tecnologico. Certi interessi si imprimono nella tecnologia, che deve essere valutata per come influenza la vita delle persone.

Lo stesso, come abbiamo visto, può essere detto a proposito di come un sistema sanitario nazionale stabilisce di accogliere le persone al suo interno, servendosi o meno di medici in carne e ossa o di sistemi di intelligenza artificiale, oppure di come una banca può stabilire di relazionarsi con i suoi clienti che le chiedono un mutuo, ancora una volta ricorrendo a una IA programmata a questo scopo o impie-

gando un funzionario che ascolti le storie degli individui che si rivolgono a lui e che sia in grado di valutare caso per caso. La letteratura critica nei confronti della stessa IA è piena di studi che mettono in discussione l'opportunità di ricorrere a una macchina quando si affrontano problematiche delicate come queste. Oppure come il fatto di suggerire a un datore di lavoro che deve decidere a chi riconoscere la possibilità di fare un colloquio con lui quale curriculum debba prendere in considerazione, eccetera (O'Neil, 2016; Zuboff, 2019; Crawford, 2021).

La politica, in sostanza, deve produrre regole per il vivere civile e, soprattutto, deve promulgare leggi, che gli Stati devono far rispettare. Di questo tema abbiamo già dibattuto un po' in queste pagine, ma non siamo ancora andati abbastanza a fondo. Nel prossimo capitolo, dunque, ci occuperemo più nello specifico di alcuni problemi legali sollevati dall'utilizzo dell'intelligenza artificiale generativa.

Note al capitolo

1. <https://www.nytimes.com/2023/03/24/opinion/yuval-harari-ai-chatgpt.html>.
2. Brano originale inglese: «Language is the operating system of human culture. From language emerges myth and law, gods and money, art and science, friendships and nations and computer code».
3. Per «super intelligenza artificiale» si fa riferimento a una IA che superi il livello umano di intelligenza.
4. Apprendimento profondo.
5. <https://yoshuabengio.org/2023/05/22/how-rogue-ais-may-arise/>.
6. A questo proposito, si parla di solito del *problema dell'allineamento* dell'intelligenza artificiale ai nostri valori e al nostro modo di pensare (Bostrom, 2014).
7. <https://www.theguardian.com/technology/2023/may/02/geoffrey-hinton-godfather-of-ai-quits-google-warns-dangers-of-machine-learning>.
8. <https://www.gartner.com/en/research/methodologies/gartner-hype-cycle>.
9. <https://dictionary.cambridge.org/it/dizionario/inglese/hype#google_vignette>.
10. Le cui traduzioni sono rispettivamente «cultura dell'*hype*» e «ciclo dell'*hype*».
11. Brano originale inglese: «AI is seizing the master key […] it can now hack and manipulate the operating system of civilization», <https://www.nytimes.com/2023/03/24/opinion/yuval-harari-ai-chatgpt.html>.
12. <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>.
13. <https://futureoflife.org/>.
14. Brano originale inglese: «Contemporary AI systems are now becoming human-competitive at general tasks».
15. <https://medium.com/@emilymenonbender/talking-about-a-schism-is-ahistorical-3c454a77220f>.

16. <https://blog.quintarelli.it/2019/11/lets-forget-the-term-ai-lets-call-them-systematic-approaches-to-learning-algorithms-and-machine-inferences-salami/> .
17. Approccio sistematico allo studio degli algoritmi e del *machine learning*.
18. <https://flash---art.it/article/intelligenza-artificiale/>.
19. Brano originale inglese: «The name 'artificial intelligence' has an implicit bias that does not allow for a cognitive perception adherent to reality. On the contrary, the name favours the suggestion of the possibility of machines to develop some form of consciousness, emotions, acquire a 'personality' similar to humans and, ultimately overcome human limitations and developing a self superior to humans. You've seen the movies, you know the narrative… But they are only devices that extract correlations from data and use those correlations to make predictions and a load of very useful things. And as having calculators doing square roots, they can do it at a scale and speed far exceeding human's performances. By throwing in some logic and randomness they can also exhibit some interesting and original behaviors. Yet machines have no clue of what reality is. At best, they mimic a model of the reality, so they are two steps away from reality»; <https://blog.quintarelli.it/2019/11/lets-forget-the-term-ai-lets-call-them-systematic-approaches-to-learning-algorithms-and-machine-inferences-salami/>.
20. <https://sts-news.medium.com/youre-doing-it-wrong-notes-on-criticism-and-technology-hype-18b08b4307e5>.
21. Letteralmente, «preoccupazione desiderosa». Si potrebbe tradurre come desiderio e preoccupazione insieme.
22. Brano originale inglese: «Crafted by writers and talkers of all stripes, wishful-worry scenarios are set in a future at some vague distance from our present and presented as matters for serious concern («We have to deal with this!»). In truth, they function as pleasant distractions from what one might call our *actual agonies*. Wishful-

worries are problems that it would be nice to have, in contrast to the actual agonies of the present»; <https://lareviewofbooks.org/article/our-censors-ourselves-commercial-content-moderation/>.
23. <https://www.carnegiecouncil.org/explore-engage/key-terms/ethics-washing>.
24. <https://www.agendadigitale.eu/cultura-digitale/ai-diffidiamo-dal-lungotermismo-plasma-il-dibattito-sugli-interessi-della-silicon-valley/>.
25. <https://artificialintelligenceact.eu/ai-act-explorer/>.
26. <https://www.aisnakeoil.com/p/a-misleading-open-letter-about-sci>.
27. Kapoor e Narayanan propongono come esempio quello che sta succedendo al mondo dell'arte con lo sviluppo e la diffusione di generatori di immagini, quali ad esempio Dall-e, Stable Diffusion e Midjourney, come si legge nelle accuse rivolte contro le società proprietarie (Vincent, 2023); <https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart>.
28. <https://www.editorialedomani.it/tecnologia/intelligenza-artificiale-chat-gpt-open-ai-mondo-del-lavoro-precario-robot-icwwg8ht>.
29. Brano originale inglese: «Our analysis indicates that approximately 19% of jobs have at least 50% of their tasks exposed when considering both current model capabilities and anticipated tools built upon them. Human assessments suggest that only 3% of US workers have over half of their tasks exposed to LLMs when considering existing language and code capabilities without additional software or modalities. Accounting for other generative models and complementary technologies, our human estimates indicate that up to 49% of workers could have half or more of their tasks exposed to LLMs. […] This paper's primary contributions are to provide a set of measurements of LLM impact potential and to demonstrate the use case of applying LLMs to develop such measurements efficiently and at scale. Additionally, we showcase the general-purpose potential of LLMs. If «GPTs are GPTs», the eventual trajectory of LLM development and application may be challenging for

policymakers to predict and regulate. As with other general-purpose technologies, much of these algorithms' potential will emerge across a broad range of economically valuable use cases, including the creation of new types of work (Acemoğlu, Restrepo, 2019). Our research serves to measure what is technically feasible now, but necessarily will miss the evolving impact potential of the LLMs over time».

30. <https://www.editorialedomani.it/tecnologia/intelligenza-artificiale-chat-gpt-open-ai-mondo-del-lavoro-precario-robot-icwwg8ht>.

31. <https://time.com/6247678/openai-chatgpt-kenya-workers/?fbclid=PAAabiRhajL9uq0H8WJVr38iZdo0UmZ7DrRPjbWq4lu8Oyhvj9CNtu5mG9uJs>.

32. La frase originale si può trovare in bella mostra sulla pagina di approfondimenti del sito internet di OpenAI <https://openai.com/> e recita testualmente: «Our mission is to ensure that artificial general intelligence benefits all of humanity».

33. <https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/>; <https://www.nytimes.com/2021/02/19/technology/google-ethical-artificial-intelligence-team.html>.

34. <https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/>.

35. <https://data.worldbank.org/indicator/IT.NET.USER.ZS?end=2017&locations=US&start=2015>; <https://www.pewresearch.org/internet/fact-sheet/internet-broadband/>.

36. Common Crawl, <https://commoncrawl.org>, è un'organizzazione senza scopo di lucro che esegue il monitoraggio del web raccogliendone i contenuti e mettendo a disposizione del pubblico i suoi archivi e i suoi dataset. L'archivio web di Common Crawl consiste in petabyte di dati raccolti dal 2008.

37. Wikipedia definisce così il significato di questa parola: «*Twink* o *twinkie* è un termine di origine anglosassone che fa parte del linguaggio gay. Indica un giovane attraente o un uomo gay che dimostra un'età

adolescenziale, con fisico magro, sottile e longilineo (ectomorfo), glabro e con pochissimi peli sul corpo»; <(https://it.wikipedia.org/wiki/Twink_(linguaggio_gay)>.

38. <https://www.law.kuleuven.be/ai-summer-school/open-brief/open-letter-manipulative-ai>.

39. <https://www.brusselstimes.com/belgium/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt>.

40. <https://www.forbes.com/sites/lanceeliot/2023/03/01/generative-ai-chatgpt-as-masterful-manipulator-of-humans-worrying-ai-ethics-and-ai-law/?sh=7eec05231d66>.

41. IA ingannevole.

42.<https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9852214>.

43. Per *gamification* si intende l'utilizzo di tecniche, strumenti e modelli di interazione cari soprattutto ai videogiochi. Nello specifico, tale attività consiste nel proporre meccanismi di ricompense, premi e incentivi di diversa natura (accumulo di punti, denaro virtuale e classifiche) per spronare e, al contempo, intrattenere l'utente.

44. <https://www.wysa.com>.

45. <https://www.wysa.com/nhs>.

CAPITOLO TERZO

ChatGPT e il diritto

In questo capitolo, esamineremo ChatGPT dalla prospettiva legale, analizzando come i principi giuridici e le normative vigenti si applicano all'uso delle opere creative e dei dati personali di milioni di persone per l'addestramento degli algoritmi. Pur senza pretendere di risolvere questioni legali complesse, attualmente oggetto di numerosi contenziosi e procedimenti da parte delle autorità garanti, ci proponiamo di fornire alcuni elementi essenziali per una comprensione critica, e il più possibile di ampio respiro, degli interrogativi sollevati dal problematico rapporto tra ChatGPT e il diritto. Molti di questi interrogativi affondano le loro radici nella storia della cosiddetta «rivoluzione digitale» (Balbi, 2022) e nelle nuove forme di capitalismo dell'informazione che sono nate e si sono sviluppate con essa.

La discussione che segue tralascia volutamente la questione della presunta «personalità giuridica» dei sistemi di

intelligenza artificiale, cioè la supposta capacità di esercitare diritti. Si tratta di suggestioni ampiamente discusse in letteratura, e che offrono spunti per sviluppare scenari avvincenti (di cui il libro di Kaplan, 2024, è particolarmente prodigo), ma che distolgono inevitabilmente l'attenzione dal presente. Infatti, i problemi legali più urgenti attualmente non riguardano la possibilità ipotetica che alcuni diritti vengano esercitati in sostituzione degli esseri umani, ma piuttosto le violazioni reali dei diritti umani stessi. Il capitolo si concentra dunque su due ambiti giuridici particolarmente rilevanti, oggi, per ChatGPT: il diritto d'autore e la protezione dei dati personali. Iniziamo con il primo.

Il problema del diritto d'autore nell'era del digitale

Il rapporto tra diritto d'autore e nuove tecnologie è, per usare un eufemismo, complesso. È ben noto che l'avvento delle tecnologie digitali prima, e di internet subito dopo, ha trasformato la riproduzione e circolazione di opere creative quali testi, immagini, musica e film in attività a buon mercato e alla portata di tutti, innescando un conflitto con gli interessi dei titolari dei diritti. Questi ultimi – editori, case discografiche e case di produzione cinematografica – hanno reagito esercitando forte pressione sui legislatori per inasprire le norme a tutela del copyright e adattarle al nuovo contesto tecnologico. Ciò ha provocato malcontento nel pubblico, che vedeva minacciato il grande potenziale di internet e del digitale per l'accesso senza barriere al sapere e alla cultura. Una frustrazione condivisa anche dagli sviluppatori di nuovi servizi su internet, che proliferavano

proprio grazie alla circolazione di opere protette da copyright tra gli utenti. Così, tra utenti e sviluppatori di nuove tecnologie si è creata una naturale convergenza di interessi, che le grandi piattaforme di internet hanno successivamente sfruttato a proprio vantaggio, anche mediante interventi incisivi nel processo legislativo. Ad esempio, nel 2011, le grandi aziende tecnologiche statunitensi, con Google in testa, riuscirono a bloccare l'adozione da parte del Congresso USA di una controversa legge che avrebbe inasprito le misure contro la pirateria online. Fondamentale per il successo di questa opposizione fu il coinvolgimento di migliaia di siti web, che chiusero temporaneamente in segno di protesta, e la mobilitazione di milioni di cittadini che si attivarono firmando petizioni, scrivendo ai membri del Congresso e addirittura scendendo in piazza a manifestare[1]. Qualche anno più tardi, Google tentò una simile operazione in Europa, finanziando campagne di attivisti che si opponevano alla proposta di una nuova direttiva sul copyright e sollecitando gli utenti a inviare email ai propri rappresentanti al Parlamento Europeo per bloccare la legge[2]. La mobilitazione non ebbe però lo stesso successo e la direttiva fu infine approvata nel 2019[3]. Attualmente, le grandi aziende tecnologiche, con Amazon, Apple e Meta in prima linea, stanno arruolando artisti e scrittori in una campagna di lobbying per il libero utilizzo di opere protette da copyright al fine dell'addestramento dei sistemi di intelligenza artificiale generativa. La campagna, denominata *Generate and Create* intende evidenziare «come gli artisti utilizzano l'IA generativa per migliorare la loro produzione creativa» e «mostrare come l'IA abbassi le barriere per la produzione di arte»: tutto ciò allo scopo di «difendere il principio del *fair use* nella legge sul copyright»[4].

Ora, se è vero che un regime di copyright troppo restrittivo può limitare diritti fondamentali come l'accesso alla cultura, all'informazione e all'istruzione, soprattutto in un contesto di «transizione digitale» in cui tali diritti si esercitano prevalentemente su internet, è però altrettanto vero che l'indebolimento del copyright può favorire la creazione di monopoli tecnologici che sono altrettanto (se non più) dannosi per l'esercizio di quegli stessi diritti. Il fenomeno economico che sta alla base di questo problema è ben descritto da Jonathan Barnett in un ampio studio dedicato alla proprietà intellettuale nella nuova struttura d'impresa:

> Un regime di diritto d'autore forte è spesso liquidato in modo sommario nelle discussioni popolari e in gran parte della letteratura accademica come un meccanismo di ricerca di rendita che aumenta i già elevati margini di profitto dei proprietari dei contenuti. Tuttavia, si dimentica che un regime di diritto d'autore debole crea una distorsione che danneggia i modelli di business basati sulla monetizzazione dei contenuti indipendenti, favorendo le piattaforme che distribuiscono gratuitamente i contenuti per promuovere le vendite di beni complementari [come la pubblicità]. In questo contesto, le piattaforme, avendo un vantaggio competitivo, riescono a ottenere profitti positivi (Barnett, 2021)[5].

È dunque un errore ridurre la questione «copyright e nuove tecnologie» a un conflitto tra ricchi proprietari di contenuti e «poveri» utenti finali, o tra vecchio capitalismo *rentier* e nuova imprenditoria dell'innovazione. Si tratta piuttosto della transizione tra due modalità di estrazione capitalistica del valore: dal modello basato sulla vendita

di contenuti a un modello in cui il valore viene estratto direttamente dagli utenti, a cui i contenuti sono offerti gratuitamente. Questo processo non scardina, ma anzi mantiene e rafforza le rendite di posizione tipiche del capitalismo maturo. Non a caso l'economista Yanis Varoufakis ha descritto l'attuale sistema economico, dominato dalle grandi aziende tecnologiche, come «tecnofeudalesimo»: un sistema in cui pochi «signori della tecnologia» non mirano più a ricavare profitto dal lavoro dei propri dipendenti, ma estraggono rendite da una massa di «servi digitali» che postano, scorrono e acquistano o vendono ininterrottamente su un numero limitato di piattaforme sempre più grandi e sempre più performanti (Varoufakis, 2023).

Non sorprende quindi che le grandi piattaforme tecnologiche, dopo aver combattuto la «guerra contro il copyright» (Baldwin, 2014) nei parlamenti e nelle aule di tribunale, abbiano finito per stringere alleanze con i grandi proprietari di contenuti per spartirsi la rendita derivante dall'estrazione di valore dagli utenti (Giblin, Doctorow, 2022)[6]. In questo contesto si inserisce anche l'attuale fase del conflitto tra copyright e nuove tecnologie, che vede di nuovo contrapposti i proprietari di contenuti – portatori di interessi molto eterogenei – ai «signori della tecnologia», ora impegnati nella corsa all'intelligenza artificiale. In questa nuova fase, i contenuti protetti da copyright non vengono più utilizzati soltanto per attirare utenti verso i loro servizi ed estrarre dati sul loro utilizzo, ma anche per sviluppare servizi in grado di produrre contenuti simili, entrando in diretta competizione con quelli originali.

ChatGPT e il diritto d'autore: i termini della questione

Non sorprende, dunque, che OpenAI sia al centro di numerose azioni legali intraprese da titolari di copyright negli Stati Uniti. Tra queste spicca la causa intentata dal «New York Times» a dicembre del 2023 che fa seguito ad altri procedimenti avviati in forma di *class action* da autori indipendenti[7]. La questione centrale in queste cause è se l'utilizzo di opere protette da copyright per «addestrare» l'algoritmo di ChatGPT sia una violazione dei diritti esclusivi dei titolari. Da un lato, gli autori e gli altri detentori di diritti sostengono di dover autorizzare e ricevere compensi per l'uso dei propri materiali a fini di addestramento. Essi adducono, in particolare, che i testi prodotti da ChatGPT siano tutt'altro che creazioni originali, ma siano invece delle contraffazioni mascherate delle loro opere. Dall'altro lato, OpenAI insiste che questo tipo di utilizzo sia un uso lecito, assimilabile a quello che hanno sempre fatto gli autori umani da che mondo è mondo: leggere e istruirsi su altre opere prima di creare le proprie. Insomma, ChatGPT non «copia» (termine che fa subito sobbalzare i titolari del *copy*right), bensì «impara» e lascia intatto ciò da cui ha imparato:

> ChatGPT non *copia* né memorizza le informazioni di addestramento in un database. Al contrario, *impara* le associazioni tra le parole […]. Non fa «copia e incolla» con le informazioni di addestramento – proprio come una persona che ha *letto* un libro e lo ripone, i nostri modelli non hanno accesso alle informazioni di addestramento dopo aver *imparato* da loro[8].

Come abbiamo già osservato in precedenza in questo libro, l'impiego di un linguaggio antropomorfizzante per descrivere il funzionamento di dispositivi tecnici – «addestramento», «apprendimento», «generazione», eccetera – è spesso funzionale all'elusione di responsabilità da parte dei produttori di quei dispositivi. Qui, in particolare, l'analogia con il supposto «processo creativo» umano, fondato sul *leggere* e *imparare* da opere precedenti, suggerirebbe un uso perfettamente legittimo che non viola alcun diritto d'autore. Ma è proprio così?

Per cercare di rispondere, è utile richiamare brevemente alcuni principi fondamentali del diritto d'autore. In estrema sintesi, il diritto d'autore o copyright è un sistema giuridico che distingue tra usi legittimi e usi vietati delle opere creative. Questi ultimi comprendono la riproduzione, totale o parziale, dell'opera protetta. In tutti i sistemi nazionali vigono alcuni principi comuni, come il fatto che il divieto di riproduzione si applica solo alla forma espressiva e non alle «idee» espresse, le quali possono essere liberamente fatte proprie da chiunque. Tuttavia, è importante considerare che la forma espressiva di un testo non si limita alla sequenza esatta delle parole, ma include anche elementi «interni» spesso di complessa valutazione, come la struttura espositiva o narrativa, la sequenza degli argomenti (in un saggio), l'intreccio dei fatti e dei personaggi (in un romanzo), e così via. Questo significa che per violare il diritto d'autore non è necessario fare «copia e incolla» dell'opera altrui, ma è sufficiente riprodurre elementi originali della sua forma espressiva.

Va aggiunto, inoltre, che il diritto d'autore contempla situazioni in cui è comunque permesso «copiare», sia in

termini di copia letterale che di riproduzione della forma interna. Si pensi ai casi della citazione, che comporta la copia letterale di una parte di un'altra opera, o della parodia, che spesso implica la ripresa integrale della forma espressiva dell'opera parodiata. O ancora al cosiddetto «pastiche», dove un autore si diverte a estrarre brani di opere preesistenti e arrangiarli in una nuova composizione. Il confine preciso di queste cosiddette «eccezioni» rappresenta uno degli ambiti in cui le leggi nazionali mostrano le maggiori differenze. Negli Stati Uniti, viene applicata la dottrina del *fair use*, che permette l'utilizzo di opere protette in molteplici situazioni senza richiedere autorizzazione, considerando fattori come lo scopo dell'uso, la natura dell'opera, la quantità riprodotta e l'impatto dell'uso sul mercato potenziale dell'opera originale. Nella determinazione del *fair use* è cruciale il riferimento ai casi precedenti, secondo la tradizione del *common law*[9]. D'altra parte, l'Unione Europea adotta un sistema di eccezioni specifiche e ben definite, che includono situazioni come la citazione, l'uso a scopo didattico o di ricerca, la parodia e, particolarmente rilevante per il nostro discorso, l'estrazione di testo e dati («text and data mining»).

La questione legale relativa a ChatGPT presenta dunque due aspetti: da un lato, occorre determinare se l'utilizzo di opere protette per l'«addestramento» – ovvero per consentire a ChatGPT di «imparare a scrivere» emulando autori umani – rientri nel *fair use*; dall'altro, stabilire se i testi generati da ChatGPT rappresentino riproduzioni di opere protette.

Opere creative «in» e «out»

Spesso, nei discorsi legali sull'intelligenza artificiale generativa, si fa distinzione tra le questioni concernenti l'«input» (cioè la fase di addestramento) e quelle relative all'«output», ossia il risultato finale del processo. La distinzione è sicuramente utile per l'analisi legale, anche se, come vedremo in seguito, non deve farci perdere di vista l'effetto complessivo dell'utilizzo delle opere protette.

Per quanto riguarda la fase di «input», osserviamo che l'addestramento di questi sistemi richiede l'impiego di grandi quantità di opere creative, quali testi, immagini, brani musicali e altro ancora, a loro volta contenuti in banche dati e spesso raccolti su internet mediante tecniche di *web scraping*[10]. Sia la riproduzione di tali opere sia la loro estrazione da banche dati potrebbero costituire una violazione di diritti esclusivi[11]. Questo coinvolge non solo gli sviluppatori diretti dei sistemi di IA, ma anche altri attori della catena del valore dell'IA, come i creatori di database utilizzati per l'addestramento di tali sistemi. Tuttavia, come vedremo meglio più avanti, tali attività possono rientrare tra gli usi consentiti dal *fair use* americano o dalle eccezioni per «text and data mining» dell'Unione Europea.

Se però passiamo a considerare la fase di «output», ossia la produzione di testi, immagini e altro materiale che simula creazioni umane, la questione si complica sensibilmente. Infatti, anche nel caso in cui le operazioni nella fase di «input» siano lecite (in virtù di un'eccezione o del *fair use*), il prodotto generato dall'intelligenza artificiale potrebbe comunque violare i diritti sulle opere utilizzate come materiale di addestramento. Ciò può accadere soprattutto se il

sistema produce contenuti identici o molto simili a quelli usati nella fase di addestramento. Ma questa non è l'unica possibilità. Anche la produzione di «opere derivate», ossia modificazioni, elaborazioni o trasformazioni di una o più opere precedenti, può costituire una violazione[12]. Solo un esame puntuale dei singoli output può stabilire in modo conclusivo se vi sia o meno una violazione. Ma dal momento che ChatGPT produce output ogni volta diversi, anche in risposta allo stesso prompt, l'analisi dei singoli casi dovrà inscriversi in un esame complessivo del funzionamento del sistema e del modo in cui è stato progettato. Le questioni rilevanti per l'analisi legale sono molteplici. Ad esempio: può il sistema impedire che i testi «ingurgitati» durante la fase di addestramento vengano «rigurgitati» tali e quali? È in grado di escludere la produzione di contenuti che rielaborano, modificano o trasformano opere protette da diritti? Come reagisce ai prompt che inducono alla violazione del copyright? E se il sistema prevede dei filtri sui prompt, quanto efficaci sono questi filtri? Quanto è facile aggirarli?

Abbiamo accennato in precedenza al fatto che, nel diritto d'autore, anche la copia della forma espressiva è permessa in alcuni casi, come ad esempio la citazione, la parodia e il pastiche. La tentazione di applicare queste categorie agli output di ChatGPT (e della IA generativa in generale) è molto forte, e ben supportata dalla tendenza naturale ad attribuire qualità umane alle macchine. Del resto, i testi prodotti da ChatGPT non sono proprio un «pastiche» di milioni di testi ingeriti in fase di «apprendimento», spesso risultante in una parodia di ragionamento umano? In questo modo, però, travisiamo completamente la ragion d'essere di queste categorie giuridiche. Le eccezioni al

diritto d'autore, così come il *fair use*, sono strumenti a tutela di diritti fondamentali quali il pieno esercizio della libertà di espressione e di pensiero. Si tratta di diritti che appartengono alla persona umana e che solo come tali possono essere esercitati, nell'interesse generale della collettività. Riferirsi agli output di un sistema automatico come «parodie», «pastiche» o «citazioni» può essere accettabile nel discorso colloquiale, influenzato dal linguaggio antropomorfizzante che pervade la narrativa dominante sull'IA, ma non ha alcun fondamento dal punto di vista legale e del diritto d'autore (Westkamp, 2024).

Se la distinzione tra «input» e «output» può aiutare l'analisi dei casi concreti, è però opportuno ricordare che le corti conducono un esame globale che va al di là delle singole fasi del processo. L'esame considera infatti gli effetti complessivi dell'utilizzo, i rischi che esso comporta per i titolari dei diritti e se tali rischi siano accettabili in relazione agli eventuali benefici per il pubblico. Questo tipo di esame è particolarmente rilevante per le corti americane, le quali applicano la dottrina del *fair use* basandosi sui precedenti giurisprudenziali.

L'incerta scommessa sul fair use negli Stati Uniti

L'opinione prevalente negli Stati Uniti è che l'utilizzo di opere protette per addestramento (ossia la fase di «input») sia da considerarsi un uso legittimo. Secondo l'orientamento giuridico statunitense, infatti, gli usi cosiddetti «non espressivi» delle opere protette, cioè quelli che mirano esclusivamente all'estrazione di elementi non tutelati dal

copyright (dati, informazioni, eccetera), e che non diffondono al pubblico la forma espressiva dell'opera, sono generalmente considerati *fair use*, anche nel caso in cui comportino la riproduzione integrale delle opere e abbiano una finalità commerciale. I casi più rilevanti spesso citati a sostegno di questo orientamento sono Google Books e iParadigms. Nel primo caso, che consiste in realtà in due procedimenti distinti, le corti hanno stabilito che la digitalizzazione di milioni di libri per svolgere funzioni istituzionali delle biblioteche (causa *Authors Guild v HathiTrust*), o per creare una funzionalità di ricerca per gli utenti di internet che mostra solo brevi frammenti di ogni opera (causa *Authors Guild v Google*), costituisce *fair use*. Nel caso iParadigms si trattava invece dell'utilizzo di un software, chiamato Turnitin, per rilevare i casi di plagio negli elaborati degli studenti. Anche in questa circostanza il giudice ha ritenuto che la riproduzione integrale degli elaborati al solo scopo di confrontarli elettronicamente con le fonti disponibili su internet sia da considerarsi *fair use*. In entrambi i casi, le corti evidenziavano la natura «trasformativa» degli utilizzi effettuati, rispettivamente, dal motore di ricerca Google e dal software Turnitin, i quali non compromettevano la funzione espressiva delle opere protette e, dunque, non incidevano sul loro valore economico[13]. Né in un caso né nell'altro, infatti, le copie delle opere protette erano rese visibili al pubblico. Come ebbe a dire un ingegnere di Google: «Non scannerizziamo tutti quei libri per farli leggere agli umani; lo facciamo per farli leggere alla nostra intelligenza artificiale» (citato in Borghi, Karapapa, 2013, p. 14).

Pur mostrando chiare analogie con questi precedenti, l'u-

tilizzo di opere protette come materiale di addestramento per l'IA generativa si distingue per due aspetti significativi, che riguardano sia la fase di addestramento («input») sia, soprattutto, l'effetto finale del processo («output»). Innanzitutto, gli utilizzi posti in essere da servizi come Google e Turnitin non erano tali da privare i titolari dei diritti di qualche possibilità di sfruttamento economico delle loro opere, poiché non esisteva un mercato realistico per tali forme di utilizzo. Nei casi Google Books, le corti erano addirittura persuase che il servizio sviluppato dal motore di ricerca non solo non danneggiasse il mercato dei libri, ma anzi potesse avere un impatto positivo su di esso, poiché attraeva potenziali lettori grazie alla funzionalità di ricerca. Ora però il contesto è cambiato: l'utilizzo di opere protette per addestrare i sistemi di IA generativa ha un valore economico ormai ampiamente riconosciuto, al punto che molti titolari di diritti stanno già traendo vantaggio da contratti di licenza a questo scopo[14]. Inoltre, come si vedrà nel prossimo paragrafo, il legislatore europeo ha espressamente introdotto un meccanismo legale che, di fatto, istituisce un mercato potenziale per l'utilizzo di opere e banche dati a fini di addestramento della IA generativa. È dunque diventato più arduo sostenere che l'uso non autorizzato per questa finalità non incide sul valore di mercato di quelle opere.

Se il mutato contesto di mercato è un fattore che le corti statunitensi dovranno tenere in dovuta considerazione, ancor più determinante per la *fair use analysis* è la valutazione dell'effetto complessivo dell'utilizzo alla luce del suo «output». Non solo i prodotti della IA generativa – comunque li si voglia definire – appartengono al medesimo genere

dei loro «input», ma potrebbero anche assomigliare molto o addirittura essere indistinguibili dalle opere utilizzate in fase di addestramento. Questo li metterebbe in diretta competizione economica con tali opere, diventando potenziali sostituti e indebolendo uno dei fattori cruciali nella determinazione del *fair use*. Non è un caso che, per vanificare la difesa del *fair use*, i titolari dei diritti abbiano insistito proprio sulla qualità degli output di ChatGPT. Ad esempio, nel procedimento avviato dal «New York Times» contro OpenAI alla fine del 2023, sono stati presentati diversi esempi in cui ChatGPT fornisce risposte che sono un «copia e incolla» di articoli di giornale accessibili solo attraverso un paywall. È da notare che nell'atto di citazione, lungo 69 pagine, la parola «*verbatim*» («alla lettera») appare per ben 69 volte, a riprova dell'insistenza sul fatto che ChatGPT non sia (soltanto) un produttore di opere di genere affine, ma effettivamente «rigurgiti» sistematicamente copie letterali di opere «ingerite» in fase di addestramento.

In risposta a tali azioni legali, e forse anche alle preoccupazioni del pubblico, OpenAI ha implementato filtri di moderazione sulle risposte. Nelle prime versioni di ChatGPT lanciate temerariamente sul mercato a partire dalla fine del 2022, il servizio poteva essere facilmente sollecitato a restituire ampie citazioni di testi protetti da copyright con un buon grado di accuratezza. Ora, invece, ChatGPT risponde generalmente a tali richieste con l'affermazione: «Mi dispiace, non posso fornire un estratto letterale da [nome dell'opera] in quanto potrebbe essere protetto da copyright» (solitamente seguita dall'offerta di fornire una «sintesi generale» o di discutere «le idee» dell'autore in questione). È però tutta da verificare l'effica-

cia di questi filtri e se rendano il sistema sufficientemente immune da richieste che potrebbero portare a una violazione del copyright. La letteratura abbonda di esempi di prompt in grado di aggirare i filtri di ChatGPT (come di altri sistemi di IA generativa) e ottenere copie quasi perfette di opere protette (Marcus, Southen, 2024)[15].

In sintesi, la scommessa di molti sviluppatori di IA generativa sul fatto che l'uso su larga scala di opere creative sia protetto dal *fair use* potrebbe dimostrarsi azzardata. Assistiamo qui a un'evoluzione per certi versi simile a quella già vissuta nella fase precedente della «rivoluzione digitale», quella che vedeva contrapposti i titolari dei diritti alle piattaforme di condivisione. Anche ora, le cause legali in corso negli Stati Uniti potrebbero preludere a un accordo tra grandi proprietari di contenuti e grandi sviluppatori di IA generativa, spesso riconducibili agli stessi «tecnofeudatari» di prima (Giblin, Doctorow, 2022).

La soluzione legislativa nell'Unione Europea

Nei paesi dell'Unione Europea le cause per violazione del diritto d'autore contro soggetti impegnati nella IA generativa sono molto meno numerose[16]. Ciò è anche dovuto al fatto che il legislatore europeo ha anticipato i pronunciamenti delle corti con una soluzione legislativa che non scontenta troppo le parti in causa. La Direttiva 790 del 2019 sul diritto d'autore nel mercato unico digitale ha infatti introdotto due eccezioni specifiche per la riproduzione di opere protette e l'estrazione da banche dati a fini di «text and data mining». L'articolo 3 consente il «text

and data mining» a scopo di ricerca non commerciale, mentre l'articolo 4 prevede un'eccezione più ampia che include anche le finalità commerciali. Durante un'audizione presso il Parlamento Europeo sull'AI Act (entrato in vigore a luglio del 2024), la Commissione ha precisato che la definizione di «text and data mining» comprende anche l'uso per l'addestramento dei sistemi di intelligenza artificiale generativa. Ma l'articolo 4 contiene una restrizione significativa, nota come «diritto di riserva» o «opt-out». Applicando una specifica dicitura sull'opera o sulla banca dati, il titolare può riservarsi il diritto di effettuare riproduzioni per «text and data mining» a scopo commerciale. In questi casi, per poter utilizzare tali opere ai fini di addestramento, si dovrà negoziare una licenza con i rispettivi titolari dei diritti. Solo le opere prive di questa riserva possono essere utilizzate liberamente.

La norma è applicabile nei paesi dell'Unione Europea. Tuttavia, l'AI Act del 2024 estende l'obbligo di rispettare il «diritto di riserva» a tutti gli operatori che commercializzano i loro prodotti nell'Unione Europea, indipendentemente dalla sede legale o dal luogo in cui avviene l'addestramento[17]. Questo implica che OpenAI dovrà adeguarsi al diritto europeo riguardo a questo aspetto cruciale per il funzionamento del proprio prodotto. Tanto più che il sistema di opt-out introdotto dal legislatore europeo è già stato adottato nella pratica, soprattutto dai grandi gruppi editoriali per «proteggere» i loro ricchi cataloghi e, in alcuni casi, utilizzarli per sviluppare autonomamente dei servizi specializzati basati su Large Language Models (LLM)[18]. Ma anche creatori e artisti indipendenti possono esercitare questa opzione. Curiosamente, al momento si ha

notizia soltanto di imprese statunitensi che offrono servizi di intermediazione a questo scopo. Ad esempio, la startup californiana Spawning AI ha lanciato un motore di ricerca, chiamato «Have I Been Trained?», che permette di identificare e gestire l'uso di immagini in dataset pubblici utilizzati per l'addestramento di modelli di IA generativa[19]. Gli autori che scelgono di escludere le proprie opere da queste banche dati, in base ad accordi contrattuali volontari con i gestori, non vedranno più le loro opere utilizzate per addestramento di IA generativa.

Inizialmente, la previsione del diritto di riserva nella legge europea è stata oggetto di critiche da parte dei commentatori, principalmente perché veniva interpretata (non senza ragione) come il risultato delle pressioni esercitate sul legislatore europeo da parte della potente lobby dei proprietari di contenuti. In seguito, anche alla luce dei contenziosi aperti negli Stati Uniti, sono però emersi pareri meno sfavorevoli. Un *policy brief* di Communia, un'associazione non governativa attiva da anni nella difesa del pubblico dominio, sostiene che la previsione dell'opt-out rappresenta una soluzione ragionevole per affrontare le questioni sollevate dall'uso su larga scala di opere protette per l'addestramento dell'intelligenza artificiale. Tale approccio, secondo Communia:

> Costituisce un quadro lungimirante per affrontare le problematiche derivanti dall'uso su larga scala di opere protette da copyright per l'addestramento di sistemi di *machine learning* (ML). È importante sottolineare che garantisce un equilibrio equo tra gli interessi dei titolari dei diritti da una parte e quelli dei ricercatori e degli sviluppatori di ML dall'altra (Communia, 2023).

In altre parole, l'opt-out consente ai creatori di decidere autonomamente se permettere o meno l'utilizzo delle loro opere per l'addestramento della IA, offrendo una flessibilità che potrebbe non solo mitigare il contenzioso tra i grandi proprietari di contenuti e le grandi aziende tecnologiche, ma anche promuovere una maggiore consapevolezza tra i singoli autori e, più ampiamente, tra i cittadini europei. A questi ultimi si rivolge in particolare la legislazione sulla protezione dei dati personali.

ChatGPT e la protezione dei dati personali

I sistemi di IA generativa funzionano grazie all'immissione di un'enorme quantità di dati di diversa provenienza, molti dei quali vengono semplicemente raccolti su internet mediante tecniche di *web scraping* – ossia l'estrazione di dati da pagine web accessibili al pubblico e la loro raccolta in banche dati. Tra la moltitudine di dati così «rastrellati» sul web si trovano non solo opere protette dal diritto d'autore, ma anche dati personali, ossia informazioni che riguardano direttamente o indirettamente le persone fisiche. Nel nostro ordinamento, chiunque raccolga e tratti dati personali di cittadini dell'Unione Europea deve rispettare le regole imposte dal regolamento generale sulla protezione dei dati personali, meglio noto come GDPR[20]. È importante ricordare che il GDPR si applica anche ai soggetti che hanno sede legale al di fuori dell'Unione, come è il caso di molte *big tech corporations* inclusa OpenAI.

Perché proteggere i dati personali?

Fino a qualche anno fa, nel discorso pubblico sulle trasformazioni economiche e sociali indotte dalla «transizione digitale» andava molto di moda ripetere la frase «I dati sono il nuovo petrolio!». Come ogni metafora, anche questa infelice espressione racchiudeva al contempo un elemento di realtà e uno di inganno. La realtà è che, allo stesso modo del petrolio, i dati nella loro forma grezza sono poco utili. Per sfruttare appieno il loro potenziale, per far sì che facciano «girare il motore dell'economia (digitale)», non basta «estrarli»: è necessario raffinarli, filtrarli e renderli utili, attraverso processi che richiedono sempre più dati e sempre maggiore potenza di calcolo. È questo binomio di fattori interdipendenti – *big data* e potenza di elaborazione – che conferisce ai dati lo status di risorsa primaria per l'economia digitale.

Ma – ed è qui che risiede l'inganno – a differenza del petrolio o del gas, i dati non sono affatto una risorsa naturale. Sono, in larga misura, il prodotto dell'azione umana nella sua complessa rete di attività, esperienze e relazioni. Una rete che non è più limitata all'«online». Per effetto della natura ubiquitaria dei dispositivi «smart», ogni singolo aspetto della vita umana – dalle semplici funzioni corporee alle complesse operazioni intellettuali –, così come ogni interazione sociale, si trasforma in una fonte preziosa di «dati grezzi». Analogamente all'estrazione del petrolio, la datificazione dell'esistenza umana richiede mezzi tecnici appropriati per costringere la «natura» a cedere il suo prezioso tesoro. Il dispositivo tecnico che sta oggi al centro di questa impresa estrattiva è lo smartphone (De Martin,

2023). Mentre perseguono la datificazione a oltranza e l'estrazione di dati, i dispositivi «smart», con lo smartphone in prima linea, insidiano direttamente e in modo nascosto i diritti individuali e collettivi. Come ha scritto Shoshana Zuboff nel suo imponente studio sul «capitalismo della sorveglianza», le tecnologie di estrazione dei dati sono in ultima analisi funzionali alla «pretesa unilaterale», avanzata da questa nuova forma di capitalismo, di «considerare l'esperienza umana come materia prima, liberamente disponibile, da tradurre in dati comportamentali» (Zuboff, 2019). Una pretesa e una traduzione che si verificano, ad esempio, scoprendo sempre nuovi terreni per l'estrazione dei dati, spingendo individui e gruppi verso comportamenti che massimizzano le opportunità di sorveglianza, indirizzando le azioni individuali e sociali verso risultati prevedibili e, infine, modellando il comportamento umano su larga scala.

L'intelligenza artificiale generativa, di cui ChatGPT è oggi il prodotto di punta, è l'espressione più avanzata di questo capitalismo estrattivo e della sua pretesa di libero e incondizionato utilizzo di dati come «materia prima grezza». Una pretesa che si pone in aperta contraddizione con la necessità di affermare che i dati costituiscono un bene comune che non può essere detenuto da privati. Dal punto di vista giuridico, la sfida normativa si gioca soprattutto sul terreno della protezione dei dati *personali*, ossia quelli che rivelano direttamente o indirettamente informazioni sulla persona. È un terreno che, nel nostro ordinamento, è presidiato dal GDPR e dalle autorità garanti preposte alla sua applicazione.

Dal provvedimento del Garante italiano alla task force europea su ChatGPT e oltre

Il 30 marzo 2023, con un provvedimento che ha suscitato grande scalpore, il Garante italiano per la protezione dei dati personali ha disposto, nei confronti di OpenAI, la limitazione del trattamento dei dati degli utenti italiani. Nel provvedimento, il Garante rilevava una serie di inadempienze da parte del produttore di ChatGPT, tra cui la mancanza di un'adeguata informativa privacy, l'assenza di sistemi per verificare l'età dei minori, la carenza di misure tecniche per evitare la perdita di dati, ma soprattutto riscontrava «*l'assenza di una base giuridica che giustifichi la raccolta e la conservazione massiccia di dati personali, allo scopo di 'addestrare' gli algoritmi sottesi al funzionamento della piattaforma*»[21].

Come approfondiremo in seguito, la presenza di una «base giuridica» per il trattamento dei dati è un requisito fondamentale e inderogabile del nostro ordinamento. Al provvedimento del Garante è seguito quindi un blocco cautelare con la richiesta a OpenAI di adeguarsi al GDPR e l'avvio di un'istruttoria. Quest'ultima si è conclusa il 2 novembre 2024 con un provvedimento «correttivo e sanzionatorio» che obbliga OpenAI a realizzare una campagna informativa della durata di sei mesi e a pagare una multa di 15 milioni di euro[22].

Il provvedimento del Garante italiano non è un'iniziativa isolata, ma si inserisce nel contesto di una crescente attenzione a livello europeo per l'uso dei dati personali nei prodotti della IA generativa. L'iniziativa del Garante italiano ha infatti indotto il Comitato europeo per la protezione dei dati (EDPB), che coordina le attività delle autorità

garanti nei paesi membri dell'Unione[23], a creare una task force su ChatGPT. I lavori di questa task force si sono conclusi a maggio 2024 con la pubblicazione di un rapporto che identifica i principi di protezione dei dati personali rilevanti per ChatGPT e offre indicazioni su come valutarne la liceità[24]. Nel suo rapporto, la task force EDPB distingue cinque fasi nel processo di generazione dei testi in cui avviene un trattamento dei dati personali. Queste fasi sono:

1. Raccolta dati per addestramento: include attività come il *web scraping* o il riutilizzo di set di dati esistenti.
2. Pre-elaborazione: comprende il filtraggio e la pulizia dei dati raccolti.
3. Addestramento: fase in cui il modello viene effettivamente addestrato utilizzando i dati pre-elaborati.
4. Prompt e output di ChatGPT: interazione dell'utente con il modello e generazione delle risposte.
5. Addestramento di ChatGPT con prompt: ulteriore addestramento del modello basato sui prompt forniti dagli utenti.

Per ciascuna di queste fasi, la task force indica le condizioni di liceità del trattamento di dati personali e le garanzie che dovrebbero essere presenti per evitare violazioni. È una valutazione complessa che chiama in causa diversi principi fondamentali del GDPR. Per i nostri scopi, ci concentreremo solo sulle due fasi più problematiche del trattamento dei dati personali, ossia la prima (raccolta) e la quarta (interazione con l'utente) – entrambe oggetto di estesa disamina da parte del Garante. Iniziamo dall'interazione di ChatGPT con l'utente finale.

L'interazione con l'utente e l'informativa privacy di ChatGPT

L'articolo 4 del GDPR stabilisce che operazioni come la raccolta, la conservazione (anche temporanea) o la diffusione di dati personali costituiscono un «trattamento» di quei dati e individua nel «titolare del trattamento» il soggetto legalmente responsabile. Il titolare del trattamento (o *data controller*) è «la persona fisica o giuridica […] che, singolarmente o insieme ad altri, determina le finalità e i mezzi del trattamento di dati personali»[25]. Tra le sue responsabilità rientra anche quella di informare i soggetti interessati riguardo alle finalità, ai mezzi, alle basi giuridiche e ad altri aspetti del trattamento dei dati, nonché di consentire agli stessi soggetti l'esercizio dei diritti previsti dal regolamento, come ad esempio il diritto di accesso ai propri dati.

Ora, un metodo con cui le grandi aziende tecnologiche tentano di eludere le norme del GDPR è quello di negare di essere titolari del trattamento. In sostanza, esse avanzano la teoria che il «titolare del trattamento» non è l'azienda tecnologica che sviluppa e diffonde il servizio, ma l'utilizzatore del servizio stesso, sul quale dunque incombono tutti gli obblighi previsti dal regolamento. Ad esempio, recentemente Microsoft ha respinto le richieste di accesso ai propri dati da parte degli utenti della piattaforma 365 Education, impiegata da molte scuole in tutto il mondo, invitandoli a rivolgersi direttamente alla loro scuola: secondo Microsoft, infatti, il titolare del trattamento dei dati non è Microsoft ma la scuola che adotta la piattaforma[26].

In modo analogo, nella versione iniziale dei Termini di Servizio, oggetto del provvedimento del Garante italiano,

OpenAI non forniva alcuna informazione sulle finalità, sui mezzi e sulle basi giuridiche del trattamento dei dati, ma soprattutto trasferiva ogni responsabilità del trattamento stesso sugli utenti. OpenAI chiedeva semplicemente agli utenti che intendessero utilizzare ChatGPT per trattare dati personali di «fornire adeguate informative privacy e ottenere i consensi necessari», oltre a dichiarare a OpenAI che il trattamento era conforme alle leggi vigenti[27].

In pratica, l'azienda si cautelava rispetto a possibili trattamenti illeciti negando di essere il titolare del trattamento e trasferendo ogni responsabilità sugli utenti del servizio: se un utente avesse inserito un prompt contenente dati personali, o anche solo suscettibile di generare una risposta che li contenesse (come ad esempio: «scrivi una poesia sulla mia compagna di banco»), e se tale trattamento fosse stato contrario alle leggi vigenti (ad esempio, se l'utente non avesse ottenuto il consenso della persona oggetto del prompt, o se il prompt avesse generato contenuti offensivi), sarebbe stato l'utente stesso a risponderne, e non OpenAI. Ma non solo. Se OpenAI fosse stata citata in giudizio dalla persona oggetto del trattamento, avrebbe potuto rivalersi sull'utente per aver violato i propri *Terms of Service*.

Ora, la pretesa di OpenAI di scaricare interamente sull'utente la responsabilità del trattamento dei dati personali è contraria non solo al testo del Regolamento, ma anche alla giurisprudenza in materia. Nel caso giudiziario noto come *Google Spain* del 2014, in cui è stato introdotto il principio del «diritto all'oblio» applicato ai risultati delle ricerche su internet, la Corte di Giustizia della UE ha stabilito che un motore di ricerca è titolare del trattamento dei dati personali presenti sui siti web da esso indicizzati. Per la Corte,

il motore di ricerca pone in essere un trattamento ulteriore e distinto da quello del gestore del sito web, nella misura in cui consente di rendere i dati accessibili «a qualsiasi utente di internet che effettui una ricerca a partire dal nome della persona interessata, anche a quegli utenti che non avrebbero altrimenti trovato la pagina web su cui questi stessi dati sono pubblicati»[28]. Il fatto che le informazioni siano state messe in circolazione dal sito web non solleva il motore di ricerca dalla responsabilità, ma al massimo configura una responsabilità condivisa tra il gestore del sito web e il motore di ricerca. In nessun caso, però, la responsabilità può cadere sull'utente di internet che effettua la ricerca su una certa persona.

Se, dunque, anche un semplice motore di ricerca è considerato titolare del trattamento di tali dati, e risponde legalmente di tale trattamento, è molto difficile pensare che un sistema di IA generativa – il quale manipola dati personali in modo ben più profondo – possa eludere le responsabilità che derivano dalle proprie operazioni.

Il Garante italiano, naturalmente, non ha neppure preso in considerazione l'ipotesi che OpenAI non fosse il titolare del trattamento e ha svolto un'analisi dettagliata della Privacy policy in vigore nel marzo del 2023 rilevando, tra le altre cose, che non veniva fornita «nessuna informazione in relazione al trattamento dei dati personali dei non utenti che vengono trattati per l'addestramento del LLM»[29]. Tuttavia, a seguito dei rilievi del Garante nel primo provvedimento, OpenAI ha modificato la propria policy includendo un'informativa specifica per gli utenti dei paesi UE[30]. In essa si dichiara di raccogliere tre categorie di «dati dell'utente», e cioè i dati personali forniti dall'utente stesso

(quando si iscrive al servizio e crea l'account), i dati ricevuti «automaticamente dall'utilizzo dei Servizi da parte dell'utente» (informazioni tecniche sull'utilizzo del servizio), e infine «Dati Personali che riceviamo da altre fonti», comprendenti «informazioni pubbliche disponibili su internet» che OpenAI utilizza «in particolare per sviluppare i modelli che alimentano i nostri Servizi». È su questa terza categoria di dati che si concentrano i rilievi giuridici più significativi. Ma proprio su questo aspetto, l'informativa è quanto mai stringata, limitandosi a dichiarare che il trattamento avviene sulla base del «legittimo interesse» di OpenAI e rinviando a un testo in inglese, in forma di FAQ, per ulteriori dettagli[31].

La base giuridica del trattamento dei dati raccolti su internet

Nel suo provvedimento sanzionatorio, il Garante ha rilevato che OpenAI avrebbe dovuto individuare una base giuridica per il trattamento dei dati personali raccolti da internet prima di iniziare a utilizzarli per l'addestramento del proprio modello. Tuttavia, dall'istruttoria non è emerso alcun documento che dimostri che tale base giuridica fosse stata identificata prima dell'avvio del trattamento, in particolare al momento della raccolta dei dati. Di conseguenza, il Garante ha contestato la violazione degli articoli 5 e 6 del GDPR, in quanto OpenAI non ha dimostrato di aver identificato correttamente la base giuridica prima dell'inizio del trattamento dei dati per l'addestramento del modello[32].

Ma che cos'è la base giuridica del trattamento e perché è così importante individuarla? Tocchiamo qui un punto cruciale del nostro ordinamento sulla protezione dei dati, in cui trovano espressione i diritti fondamentali della persona presidiati dal GDPR. Il trattamento dei dati personali è lecito unicamente a condizione che ricorra almeno una delle basi giuridiche elencate in modo esaustivo all'articolo 6 del Regolamento. Queste comprendono il consenso dell'interessato, la necessità (per concludere un contratto, o adempiere a un obbligo di legge o per salvare la vita di una persona) o il legittimo interesse del titolare del trattamento. Quest'ultimo può essere invocato come legittima base giuridica solo «a condizione che non prevalgano gli interessi o i diritti e le libertà fondamentali dell'interessato»[33], cioè della persona titolare dei dati. Ciò riguarda non solo le informazioni «private», ossia quelle che riguardano più da vicino la tutela della privacy, ma anche le informazioni che la persona ha reso pubbliche, ad esempio condividendole volontariamente su internet. I dati personali non perdono la loro natura quando vengono divulgati direttamente dalla persona a cui si riferiscono. È pertanto un errore ritenere che l'utilizzo di informazioni «che si trovano su internet» sia sempre e comunque consentito. Come ha scritto il Garante italiano in un altro procedimento:

La pubblica disponibilità di dati in internet non implica, per il solo fatto del loro pubblico stato, la legittimità della loro raccolta da parte di soggetti terzi. Infatti, ogni dato che viene pubblicato online subisce tale operazione di trattamento (segnatamente, la diffusione) sulla scorta di una base giuridica e per finalità determinate e legittime stabilite e perseguite dal titolare

del trattamento che ne ha disposto la pubblicazione. […] [L]a pubblicazione in internet di dati personali da parte del soggetto cui si riferiscono, ad esempio nell'ambito di un social media network, non comporta, di per sé, una condizione sufficiente per legittimarne il libero riutilizzo da parte di soggetti terzi[34].

Il principio che sottende questa restrizione nel trattamento dei dati «pubblici» è semplice: quando abbiamo condiviso i nostri dati, ad esempio foto e video sul nostro social media preferito, lo abbiamo fatto per uno scopo definito (più o meno chiaro e consapevole), che viene sancito con l'espressione del consenso alla loro diffusione. Una volta resi pubblici, i dati possono essere sottoposti a ulteriori trattamenti per i quali non abbiamo dato espressamente il consenso, ma che possiamo ragionevolmente aspettarci una volta che abbiamo diffuso quelle foto o quei video su internet. Ad esempio, è ragionevole attendersi che i motori di ricerca trattino i nostri dati per rendere tali informazioni reperibili online[35]. La «ragionevole aspettativa» della persona è un elemento fondamentale per stabilire se un trattamento di dati personali senza il consenso della persona possa rientrare nel legittimo interesse del titolare del trattamento. Nel provvedimento citato in precedenza, il Garante italiano ha stabilito che gli usi delle immagini pubblicate sui social da parte di Clearview AI, una compagnia australiana di riconoscimento facciale, non rientravano nelle ragionevoli aspettative degli utenti. L'azienda in questione raccoglieva sistematicamente immagini di persone su internet, le elaborava con tecniche biometriche e le incrociava con altri dati personali disponibili online per addestrare il proprio sistema a riconoscere

l'identità di una persona a partire da un'immagine catturata, ad esempio, da una telecamera di sorveglianza (Hill, 2023). Per il Garante:

> L'eventuale natura pubblica delle immagini non è sufficiente a far ritenere che gli interessati possano ragionevolmente attendersi un utilizzo per finalità di riconoscimento facciale, per giunta da parte di una piattaforma privata, non stabilita nell'Unione e della cui esistenza e attività la maggior parte degli interessati è ignaro[36].

Pur senza voler tracciare analogie superficiali tra la condotta di OpenAI e quella di Clearview, che è stata oggetto di pesanti sanzioni da parte del Garante italiano e non solo[37], è però evidente la natura problematica delle tecnologie basate sul rastrellamento indiscriminato di dati personali. I problemi sono tanto più seri in quanto, tra i dati personali raccolti a tappeto su internet, si trovano anche dati considerati particolarmente sensibili sotto il profilo dei diritti e delle libertà fondamentali, come quelli che rivelano l'appartenenza a un partito politico, a un sindacato, o l'orientamento sessuale, o l'adesione a un credo religioso. Il trattamento di questi dati è espressamente vietato dal GDPR, anche in presenza di un legittimo interesse del titolare, fatta eccezione per i dati «resi manifestamente pubblici dall'interessato»[38]. Un'eccezione che mira principalmente a garantire la libertà di informazione e che non svincola il titolare del trattamento dal rispetto della «ragionevole aspettativa» dell'interessato e dagli altri obblighi imposti dal Regolamento[39].

Possiamo davvero ritenere che l'uso di dati personali per addestrare un algoritmo che genera informazioni

molto spesso inesatte e talvolta offensive o diffamatorie[40] – quelle che OpenAI, nei soliti termini antropomorfici, chiama «allucinazioni» – sia un uso conforme alle aspettative di chi, consapevolmente o meno, condivide i propri dati personali online[41]?

Epilogo

Secondo la narrativa dominante propagata fin dall'inizio della «rivoluzione digitale», una narrativa che si riflette sia nel discorso popolare che in molta letteratura accademica, il diritto è sempre in ritardo rispetto allo sviluppo tecnologico. Il compito principale dei giuristi, dei legislatori e delle corti sarebbe innanzitutto quello di rimuovere gli ostacoli all'«innovazione», considerata il valore più alto da proteggere nell'interesse di tutti. Con il favore di questa narrativa, si lanciano nuovi dispositivi, prodotti e modelli di business in aperto contrasto con le leggi esistenti scommettendo sul fatto che, alla fine, il diritto cederà alla tecnologia e all'innovazione. La scommessa sull'evoluzione futura del diritto crea una «bolla giuridica» che, al pari delle bolle speculative, genera euforia e aspettative irrealistiche sull'effettiva tenuta legale di questi dispositivi (Giraudo, 2022). Gli esempi abbondano. All'inizio degli anni Duemila, progetti come Napster, Grokster e molti alti sfidavano apertamente le leggi sul copyright, consentendo agli utenti di condividere liberamente file musicali e altri contenuti protetti su internet. Nonostante la loro popolarità, la scommessa di poter operare entro il quadro legale esistente fu persa, e alla fine furono costretti a chiudere i battenti. Con l'avvento

delle *big tech* del capitalismo della sorveglianza il metodo non è cambiato, ma la posta in gioco è diventata così alta che queste aziende sono diventate troppo grandi per fallire, anche quando perdono la scommessa sul diritto futuro (Giraudo, Fosch-Villaronga, Malgieri, 2024). Infine, con l'avvento dell'intelligenza artificiale, la bolla giuridica si è ingigantita a tal punto che soggetti come Clearview AI, le cui pratiche illegali sono state sanzionate in tutto il mondo, continuano a prosperare diffondendo i propri dispositivi aggirando divieti e sanzioni e promuovendoli dove la legge ancora non è arrivata (Hill, 2023)[42].

L'intelligenza artificiale generativa e ChatGPT sono l'ultimo prodotto di un capitalismo estrattivo che, per operare a pieno regime, deve reclamare l'esperienza umana depositata in dati personali e opere creative come una «terra di nessuno» da occupare e sfruttare prima degli altri e il più in fretta possibile. Vale a dire: anticipando i tempi di reazione del diritto. Ma è anche, come abbiamo visto, un prodotto che esaspera il conflitto tra i molti che forniscono la necessaria materia prima di questo processo estrattivo e i pochissimi che ne controllano i mezzi di estrazione del valore e beneficiano di tale rendita di posizione. Dal modo in cui le corti e le autorità garanti risolveranno questo conflitto si comprenderà, forse, se il diritto può ancora avere un ruolo nella costruzione del *mondo in cui vogliamo vivere* o se è definitivamente ridotto a un discorso obsoleto al traino della tecnologia e degli interessi dominanti.

Note al capitolo

1. Per una ricostruzione degli eventi si veda la pagina di Wikipedia <https://en.wikipedia.org/wiki/Protests_against_SOPA_and_PIPA>. Una discussione critica approfondita è svolta da Julie Cohen (2019, pp. 124-126.
2. *Google funds website that spams for its causes*, «The Times», 6-8-2018, <https://www.thetimes.com/article/google-funds-activist-site-that-pushes-its-views-rg2g5cr6t>.
3. La direttiva 790 del 2019 sul diritto d'autore nel mercato unico digitale ha introdotto, all'articolo 17, obblighi più stringenti per le grandi piattaforme di condivisione riguardo al controllo e alla rimozione di contenuti protetti.
4. *Big Tech Launches Campaign to Defend AI Use*, «The Hollywood Reporter», 6-6-2024, <https://www.hollywoodreporter.com/business/business-news/big-tech-lobby-ai-use-1235916540>.
5. Traduzione nostra.
6. Maurizio Borghi, *Why tech giants have little to lose (and lots to win) from new EU copyright law*, «The Conversation», 19-9-2018, <https://theconversation.com/why-tech-giants-have-little-to-lose-and-lots-to-win-from-new-eu-copyright-law-103374>.
7. Per un elenco aggiornato delle cause in corso negli Stati Uniti contro sviluppatori di sistemi di intelligenza artificiale generativa si può consultare il sito *ChatGPT Is Eating the World*, <https://chatgptiseatingtheworld.com>.
8. *How ChatGPT and Our Language Models Are Developed*, <https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed>. Corsivi nostri.
9. Il *common law* è un sistema giuridico basato sulla giurisprudenza, ovvero sulle decisioni dei giudici nei casi anteriori, che creano dei precedenti più o meno vincolanti per i casi futuri.

10. Ossia la raccolta automatica di dati e altri contenuti da pagine web allo scopo di estrarre informazioni.
11. I paesi dell'Unione Europea riconoscono, oltre al diritto d'autore sulle opere creative, un diritto *sui generis* sulle banche dati.
12. Il diritto esclusivo di elaborare l'opera, previsto all'art. 18 della legge italiana sul diritto d'autore, comprende «le traduzioni in altra lingua, le trasformazioni da una in altra forma letteraria o artistica, le modificazioni e aggiunte che costituiscono un rifacimento sostanziale dell'opera originaria, gli adattamenti, le riduzioni, i compendi, le variazioni non costituenti opera originale» (art. 4 l.a.).
13. Nel caso di Turnitin, peraltro, le opere in questione (elaborati di studenti) erano prive di un valore economico effettivo.
14. Bill Rosenblatt, *The Media Industry's Race to License Content for AI*, «Forbes», 18-7-2024, <https://www.forbes.com/sites/billrosenblatt/2024/07/18/the-media-industrys-race-to-license-content-for-ai>.
15. *AI chatbots' safeguards can be easily bypassed, say UK researchers*, «The Guardian», 20-5-2024, <https://www.theguardian.com/technology/article/2024/may/20/ai-chatbots-safeguards-can-be-easily-bypassed-say-uk-researchers>.
16. Al momento della pubblicazione di questo libro sono note solo due cause avviate in Germania: la prima, conclusasi con una sentenza di primo grado il 27 settembre 2024 presso il tribunale di Amburgo, tra un titolare di diritti su immagini fotografiche e LAION, un database creato tramite *web scraping* e utilizzato da sviluppatori di IA generativa; la seconda, intentata dalla società di gestione collettiva GEMA contro OpenAI presso la corte di Monaco di Baviera nel novembre 2024.
17. Regolamento (UE) 2024/1689 sull'intelligenza artificiale, art. 53(1)(c).
18. Fra i molti esempi possiamo menzionare Lexis+ di LexisNexis per la professione legale (https://www.lexisnexis.com/html/lexisnexis-generative-ai-story/) e Scopus AI di Elsevier per la ricerca in vari ambiti

scientifici, entrambi riconducibili al gruppo editoriale RELX. Per quanto riguarda le immagini, Getty Images, la rinomata agenzia fotografica che ha intentato azione legale contro Stability AI nel Regno Unito, ha collaborato con NVIDIA per creare una propria piattaforma di IA generativa, <https://www.gettyimages.it/ia/generazione/informazioni>.
19. <https://haveibeentrained.com/>.
20. «General Data Protection Regulation» (Regolamento UE 2016/679 sui dati personali).
21. *Intelligenza artificiale: il Garante blocca ChatGPT. Raccolta illecita di dati personali. Assenza di sistemi per la verifica dell'età dei minori*, comunicato stampa, 31-3-2023, <https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9870847>.
22. *ChatGPT, il Garante privacy chiude l'istruttoria*, comunicato stampa, 20-12-2024, <https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/10085432>.
23. European Data Protection Board (EDPB), <https://www.edpb.europa.eu>.
24. EDPB, *Report of the work undertaken by the ChatGPT Taskforce*, May 2024, <https://www.edpb.europa.eu/our-work-tools/our-documents/other/report-work-undertaken-chatgpt-taskforce_en>.
25. GDPR, art. 4(7). Dal «titolare del trattamento» (*data controller*) si distingue il «responsabile del trattamento» (*data processor*), definito all'art. 4(8) come «la persona fisica o giuridica […] che tratta dati personali per conto del titolare del trattamento», ossia il soggetto che esegue le operazioni tecniche di trattamento secondo le istruzioni del *data controller* e su cui incombono obblighi legali molto meno gravosi.
26. *Microsoft violates children's privacy – but blames your local school*, 4-6-2024, <https://noyb.eu/en/microsoft-violates-childrens-privacy-blames-your-local-school>. L'associazione None of your business (NOYB) ha presentato, contro Microsoft, un esposto al Garante austriaco della protezione dei dati.

27. Garante per la Protezione dei Dati Personali, provvedimento del 2 novembre 2024 [10085455], par. 3.1.3. La policy contestata (in vigore il 14 marzo 2023) recitava: «If you use the Services to process personal data, you must provide legally adequate privacy notices and obtain necessary consents for the processing of such data, and you represent to us that you are processing such data in accordance with applicable law» («Se utilizzi i Servizi per trattare dati personali, devi fornire avvisi sulla privacy legalmente adeguati e ottenere i consensi necessari per il trattamento di tali dati, dichiarando a noi che stai trattando tali dati in conformità con la legge applicabile»). Il testo della policy è reperibile all'indirizzo <https://www.linkedin.com/posts/piatesdorf_openai-products-privacy-terms-april-2023-activity-7049344427445673984-WHu0>.
28. Corte di Giustizia dell'Unione Europea, causa C-131/12 (*Google Spain contro Mario Costeja Gonzalez*), 2014, par. 36.
29. Provvedimento del 2 novembre 2024 [10085455], par. 3.1.3.
30. La versione pubblicata al momento di scrivere queste pagine era quella in vigore a partire dal 15-2-2024: <https://openai.com/it/policies/eu-privacy-policy>.
31. <https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed>. Senza entrare in un esame approfondito del testo, osserviamo solo che alla domanda «How does the development of ChatGPT comply with privacy laws?» («Come si conforma lo sviluppo di ChatGPT alle leggi sulla privacy?») segue un paragrafo che inizia con «We train information lawfully» («Utilizziamo le informazioni a fini di addestramento in modo legale»), che si conclude con un rinvio, per «more details» (maggiori dettagli), alla Privacy Policy… Insomma, cerchiamo nell'Informativa Privacy informazioni sulla base giuridica del trattamento dei dati e veniamo indirizzati a un testo che, per ulteriori dettagli, rinvia di nuovo all'Informativa Privacy!

32. Provvedimento del 2 novembre 2024 [10085455], par. 3.1.2.
33. GDPR, art. 6 (f).
34. Ordinanza ingiunzione nei confronti di Clearview AI – 10 febbraio 2022 [9751362]
35. Così la Corte di Giustizia europea nel caso *Google Spain* ricordato poco fa. Per una critica dell'approccio della corte su questo punto si rimanda a Guido Scorza, *AI, è ora di difendere i nostri dati dalla raccolta massiva: ecco come*, «Agenda Digitale», 18-6-2024, <https://www.agendadigitale.eu/sicurezza/privacy/linsostenibile-scorrettezza-del-training-degli-algoritmi-come-difendere-i-nostri-dati/>.
36. Ordinanza ingiunzione nei confronti di Clearview AI – 10-2-2022 [9751362].
37. Clearview AI ha ricevuto sanzioni da venti milioni di euro dal Garante italiano, francese, greco e svedese, mentre il Garante austriaco ha ordinato la cancellazione dei dati. Anche fuori dall'Unione Europea (Regno Unito, Canada e altri paesi) l'azienda è stata oggetto di provvedimenti e sanzioni (Pathak, 2022).
38. GDPR, art. 9(e).
39. L'EDPB spiega che l'eccezione al divieto di trattamento dev'essere interpretata in modo restrittivo: si applica solo ai dati sensibili che l'interessato ha volontariamente reso pubblici, ad esempio configurando espressamente il proprio profilo social, con piena consapevolezza delle conseguenze di tale diffusione. EDPB, <https://www.edpb.europa.eu/system/files/2021-04/edpb_guidelines_082020_on_the_targeting_of_social_media_users_en.pdf>. Nella sentenza *Meta contro Bundeskartellamt* del 4-7-2023 la Corte europea ha sancito che se un insieme di dati contiene sia dati sensibili che non sensibili, e questi non possono essere separati al momento della raccolta, allora il trattamento dell'insieme è vietato ai sensi dell'articolo 9, paragrafo 1, del GDPR, se non ricorre una delle deroghe previste dal paragrafo 2 (Causa 252/21, par. 89).

40. OpenAI è stata citata in giudizio per diffamazione negli Stati Uniti e in Australia, e forse anche in altre giurisdizioni. Per un esame critico si rimanda a Volokh, 2023.
41. Il Garante italiano si è espresso con cautela su questo punto, riconoscendo che OpenAI, a seguito del provvedimento del marzo 2023, ha «implementato alcune misure per ridurre gli effetti dell'inaccuratezza degli output». Ciò tuttavia lascia la questione giuridica «tutt'altro che risolta»; ma, trattandosi di una «violazione continuativa» e a seguito dello stabilimento di OpenAI in Irlanda, il Garante ha deciso di non pronunciarsi su questo aspetto, trasmettendo gli atti alla competente autorità nazionale (provvedimento del 2 novembre 2024 [10085455], par. 6.6).
42. Ad esempio, Clearview ha donato il proprio sistema di riconoscimento facciale al governo ucraino, il quale lo impiega per individuare «soldati e spie» (russe) sul proprio territorio (Hill, 2023, cap. 24). Si veda anche *Exclusive: Ukraine has started using Clearview AI's facial recognition during war*, «Reuters», 22-3-2022, <https://www.reuters.com/technology/exclusive-ukraine-has-started-using-clearview-ais-facial-recognition-during-war-2022-03-13/>. La notizia è rilanciata sul sito ufficiale di Clearview: <https://www.clearview.ai/ukraine>.

CONCLUSIONI

Quale rivoluzione digitale?

Come abbiamo anticipato all'inizio di questo libro, molta della pubblicistica contemporanea su ChatGPT e i suoi epigoni – anche su quei sistemi che lavorano sulle immagini o sui suoni, oppure su quelli multimodali, capaci di servirsi di ogni tipo di linguaggio – sostiene che ci troviamo di fronte all'ennesimo passaggio epocale della rivoluzione digitale. Secondo alcuni, dopo l'invenzione del computer e di internet, avremmo dato origine a una nuova tecnologia informatica capace di realizzare addirittura il sogno dell'intelligenza artificiale generale:

> Abbiamo dunque finalmente trovato il sacro graal delle AI, la cosiddetta «Artificial General Intelligence», macchine all'altezza dell'intelligenza umana, se non superiori? La AGI, da non confondere con le GAI[1], da tempo è il sogno irraggiungibile degli scienziati, protagonista di un'infinità di libri e film di fanta-

scienza. La risposta è: «sì, ma». Per i loro scopi pratici, questi sistemi sono «cervelli sintetici» effettivamente versatili, ma questo non vuol dire che siano dotati di una «mente» come quella degli umani. Non hanno obiettivi e desideri indipendenti, non hanno pregiudizi e aspirazioni, emozioni e sensazioni. Queste restano caratteristiche puramente umane. Tali programmi sanno però comportarsi come se disponessero davvero di questi tratti caratteriali, a patto di essere addestrati con i giusti dati e venire programmati per perseguire i giusti obiettivi. […]

Probabilmente siamo alle soglie di un nuovo Rinascimento: un enorme movimento culturale e filosofico. Caratteristico del Rinascimento, dal quattordicesimo al diciassettesimo secolo, è stato lo spostarsi del focus dalla religione (e da Dio) alle conquiste secolari degli esseri umani, con progressi enormi nell'arte, nella scienza, nella tecnologia e nella conoscenza. L'AI generativa potrebbe innescare un nuovo scarto culturale, spostando il focus sulle macchine, permettendoci di utilizzare la potenza dell'intelligenza sintetica come strumento per accelerare il progresso (Kaplan, 2024, posizioni nel Kindle 84-105).

Secondo i loro più entusiastici cantori, le IA generative produrranno sulla nostra società un impatto superiore a quello della ruota, della stampa, della lampadina, della pennicillina, del telefono o dell'energia nucleare (Kaplan, 2024). Le versioni future di queste tecnologie, infatti,

ci faranno da assistenti personali. Alcune prenderanno appunti per noi, altre ci rappresenteranno sui forum, promuoveranno i nostri interessi, gestiranno le nostre comunicazioni e ci avviseranno dei pericoli imminenti. Saranno il volto delle agenzie governative, di aziende e organizzazioni. Connesse a una

rete di sensori, monitoreranno il mondo circostante per avvisarci di possibili disastri ambientali come trombe d'aria, incendi nei boschi e fuoriuscite di materiali tossici. E in caso di crisi, potremmo delegare loro il compito di agire immediatamente, ad esempio facendo atterrare un aereo in una situazione di emergenza o salvando un bambino che si è avventurato in mezzo al traffico [Kaplan, 2024, posizioni nel Kindle 79-84].

Chi non potrà permettersi cure mediche di alto livello erogate da esseri umani, potrà ottenere precise diagnosi e cure affidabili dall'intelligenza artificiale. Chi non potrà accedere a un'istruzione adeguata fornita da docenti in carne e ossa, avrà il suo tutor personale sintetico che lo formerà nella maniera migliore. Chi, per colpa dei suoi insegnanti negligenti o di un sistema scolastico deficitario, non avrà appreso correttamente come si legge, come si comprende e come si compone un testo multimodale, lo potrà fare con semplicità e immediatezza grazie a questo genere di strumenti. Chi si sentirà solo, avrà «qualcuno» con cui comunicare. Chi vorrà programmare un computer, non dovrà imparare alcun linguaggio informatico, ma potrà chiedere a un LMM di farlo per lui, domandandoglielo nella propria lingua. Chi avrà bisogno di una buona consulenza legale, la potrà ottenere a poco prezzo. Chi vorrà fare arte, non avrà che da esplicitare a una IA generativa che tipo di immagine, musica, poesia o altro tipo di opera desidera produrre e questa la realizzerà. Infine, chi saprà fare bene tutte queste cose o chi saprà già come procurarsele, servendosi di queste tecnologie migliorerà comunque la propria produttività, risparmierà tempo e denaro, saprà come accedere ai servizi di maggiore qualità, eccetera. Insomma, sarà

davvero una rivoluzione e i suoi benefici si riverbereranno nella vita di tutti. Certo, dovremo produrre tanti piccoli e grandi aggiustamenti nelle nostre esistenze. Magari qualcuno perderà il lavoro e dovrà imparare come procurarsene uno più al passo con i tempi. Dovremo fare in modo che le IA non producano allucinazioni, non si ribellino ai loro creatori, rimangano sempre allineate con i nostri valori, e così via. Ma varrà assolutamente la pena di fare questo sforzo, visti gli innumerevoli vantaggi che ne trarremo.

Naturalmente, anche noi speriamo che tutto questo accada, ma riportando quali sono le diverse critiche che vengono rivolte oggi a ChatGPT e alle intelligenze artificiali generative, crediamo di aver dimostrato che il presente è molto diverso da questi scenari utopistici. A questo proposito, la campagna di marketing del tecno-entusiastico libro di Kaplan – *Generative AI* (2024) – da cui abbiamo tratto le ultime citazioni appena commentate, recita che non saranno queste tecnologie a toglierci il lavoro, ma i nostri colleghi che hanno imparato a conoscerle e a utilizzarle prima di noi. Essa, in sostanza, rivela che non viviamo in un'epoca particolarmente rivoluzionaria, ma in un periodo storico dominato ormai da decenni dal neoliberismo, in cui persone e aziende che si fanno concorrenza per produrre profitti investono molto tempo e denaro in questo tipo di strumenti, scommettendo che ne trarranno un grande tornaconto. Per ottenerlo, portano avanti tutte le pratiche di cui abbiamo scritto, dalla creazione di *hype*, allo sfruttamento della manodopera a basso costo nel sud del mondo, dal consumo smodato di energia, all'eccessiva emissione di sostanze inquinanti, dal confezionamento di contenuti che appaiono significativi alla maggioranza delle

persone abbienti che possono pagare a caro prezzo i servizi dell'intelligenza artificiale, alla scelta di trascurare gli interessi delle minoranze più svantaggiate o meno rappresentate, dall'attività di lobbying sui governi affinché non legiferino per bloccare lo sfruttamento di dati e prodotti dell'ingegno, al tentativo di eludere le leggi ora in vigore.

Ma se anche si trattasse di una rivoluzione, dovremmo comunque chiederci, come abbiamo fatto nell'introduzione di questo libro, se essa ci porti avanti o se ci riporti indietro, se ci apparecchi un futuro davvero radioso, oppure cupo. A questo proposito, in un altro nostro lavoro[2], abbiamo riscontrato che gli studiosi di questa rivoluzione digitale che staremmo attraversando sono soliti tratteggiarla in quattro maniere diverse, a seconda che essa, i suoi strumenti e il nostro modo di utilizzarli ci conducano, a loro modo di vedere, verso un avvenire individualistico, tribale, responsabile o inclusivo. Nel primo caso, la concepiamo come il frutto dell'invenzione di potenti tecnologie che ci consentono di realizzarci, per l'appunto, come individui, in una società liberale non molto diversa da quella in cui viviamo oggi, ma più florida e funzionale, in cui ciascuno è responsabile per sé stesso ma non verso gli altri, dato che ognuno ha i mezzi informatici e le opportunità per mettere a frutto i propri talenti. Nel secondo caso, invece, pensiamo la rivoluzione digitale come qualcosa che è per pochi, come se le macchine su cui si regge servissero a gruppi più o meno grandi – élite, nazioni, blocchi di paesi alleati, in ultima istanza moderne «tribù» – per trarre dei vantaggi nei confronti di chi non ne fa parte e costruire un mondo basato sui propri valori, che non contempla il fatto di farsi carico di quelli altrui. Nel terzo

caso, piuttosto, immaginiamo di voler essere responsabili e di servirci della tecnologia per risolvere i grandi problemi collettivi dell'umanità, come quelli ecologici, la carenza di cibo, le malattie, eccetera. Nell'ultimo, infine, concepiamo la rivoluzione digitale come un processo che, grazie ai suoi strumenti che ci consentono di superare i nostri *bias* e le nostre ideologie, ci conduce verso una migliore comprensione del punto di vista degli altri, verso l'incontro e l'armonizzazione con le loro prospettive, nella direzione di una società, per l'appunto, più inclusiva. L'intelligenza artificiale generativa, dunque, intesa oggi come il motore di tutto ciò, può essere inquadrata in ognuno di questi modi e, a seconda del modello che si utilizza per pensare al futuro che grazie a essa si vorrebbe costruire, può venire progettata, utilizzata e interpretata diversamente.

Nelle pagine di questo libro, in sostanza, abbiamo mostrato come questi quattro modelli diano forma anche ai discorsi di chi si occupa di ChatGPT e dei Large Language Models. Il dibattito è molto acceso, tra chi li vede come sistemi per l'*empowerment* individuale e chi come i mezzi per l'instaurazione di tecnocrazie elitistiche, tra chi è convinto che risolveranno i più svariati problemi dell'umanità e chi ritiene che piuttosto li acuiranno. Oppure, ancora, tra chi si accontenterebbe che servissero per ridurre, invece che per esasperare, alcune forme di disuguaglianza. Noi, come abbiamo dichiarato già nel titolo del nostro lavoro, intendiamo essere critici nei confronti di coloro che pensano l'IA generativa come uno strumento per portare avanti visioni del mondo di natura individualistica, «tribale» o finanche soluzionistica, come se bastasse affidarsi a queste tecnologie per migliorare la realtà in cui viviamo e la

nostra condizione. Ci piacerebbe, piuttosto, che essa diventasse l'ambito del confronto dialettico, anche aspro ma costruttivo, tra le posizioni dominanti dei soggetti e delle istituzioni che la stanno realizzando o che si possono permettere di utilizzarla, e quelle di chi ne vede tutti i limiti, sia per come è fatta, sia per il tipo di futuro poco desiderabile verso cui ci potrebbe condurre se non venissero affrontati e risolti i problemi che abbiamo sollevato.

Note alle Conclusioni

1. Le Generative Artificial Intelligence (intelligenze artificiali generative).
2. Antonio Santangelo, *La rivoluzione digitale e il futuro. Narrazioni a confronto*, «Lexia», vol. 45-46, pp. 356-374.

Postfazione

di *Marco Ricolfi*

A che cosa può servire una postfazione? Forse a tirare le somme del lavoro che precede e immaginare direzioni di ricerca e di marcia ulteriori.

Il lavoro si compone di tre blocchi: uno fenomenologico (che cos'è ChatGPT), l'altro antropologico-politico (che impatto ha sulle nostre società), l'ultimo legal-istituzionale (quali sono i punti di crisi giuridici). Questi sono presentati con la tecnica della meta-narrazione e quindi attraverso un'esposizione polifonica delle diverse facce del dibattito in corso.

Il lavoro pone all'ordine del giorno una questione che si affaccia con urgenza crescente nella sequenza delle parti che lo compongono, quella delle norme, delle regole, dei diritti: quale regolazione per l'intelligenza artificiale e in particolare per ChatGPT?

Tre sono i livelli a cui questo interrogativo va analizzato

e affrontato: quello della *velocità* del cambiamento tecnologico, dell'individuazione della sede istituzionale o *foro* deputato a selezionare la risposta normativa e, infine, della ricognizione degli *interessi* contrapposti a quelli dei protagonisti dell'innovazione tecnologica.

Sul primo tema, quello della velocità del cambiamento, il lavoro si sofferma ripetutamente: per segnalare come l'*hype*, l'intonazione iperbolica che caratterizza le narrazioni prevalenti, possa essere talora fuorviante, talaltra smentita dai fatti; ma anche per affacciare il dubbio che le risposte della società siano destinate ad arrivare sempre in ritardo, in una sorta di fatica di Sisifo normativo. È sicuramente vero che, a partire da una data ben identificabile, il gennaio del 2007, quando ha fatto la sua prima comparsa lo smartphone della Apple, il cambio della velocità di marcia della tecnologia è stato nettissimo, mentre le risposte della società, intese come istituzioni legislative e corti, sono fin qui state esitanti e tardive. Non sembra però fuori luogo assumere, almeno come ipotesi, che questo divario possa essere in qualche misura colmato e coltivare la speranza che la società civile, rimasta «spiazzata» da un cambiamento di carattere inusitato, possa rimettersi al passo.

Va in ogni caso affrontata la seconda questione. Anche immaginando che la regolazione possa riprendere controllo dell'iniziativa dei protagonisti dell'innovazione tecnologica, ci si deve domandare in quale sede, o, meglio, in quali sedi, questo possa avvenire. Tradizionalmente, la regolazione dei poteri privati ha avuto luogo a livello nazionale e locale. Nell'Unione Europea, che, però, si ricordi, abbraccia una quota sempre decrescente del PIL e della popolazione mondiale, si è affiancata una dimensione regionale, che

talora ha registrato qualche successo (l'espressione «effetto Bruxelles» è nata proprio per segnalare l'impatto transnazionale della disciplina europea dei dati, GDPR). Infine, vi è la dimensione internazionale e multilaterale.

Ora, il significato del livello nazionale non può qui essere del tutto trascurato. Il conflitto fra ChatGPT e regole sulla *data protection* è stato, un po' sorprendentemente, innescato da un Garante nazionale, quello italiano (come ricorda il cap. 3, § 8, del lavoro); ma si è subito allargato alla dimensione UE. Se, come si è accennato, la disciplina del GDPR ha avuto qualche successo, si sta tuttavia diffondendo la consapevolezza che la regolazione di fenomeni come il *machine learning* basato sull'uso dei dati disponibili in rete, lo *scraping* («raschiatura») dei dati, l'erogazione di servizi di IA generativa imperniati sulla potenza di fuoco computazionale, non può essere affidata al comando verticale di un unico regolatore, anche se regionale e non solo nazionale.

Esiste infatti una letteratura importante che mostra come, in società algoritmiche come la nostra, la protezione dei diritti fondamentali non possa essere affidata alla regolazione tradizionale, *top down*, ma debba ricorrere a forme di regolazione *by design*, che si valgono di tecniche di co- e di auto-regolazione (Cleynenbreugel, 2023).

Il fatto è che, come è stato osservato, abbiamo assistito a una svolta infrastrutturale dell'infosfera. Da Negroponte in poi, ci siamo convinti che si fosse passati dagli atomi ai bit come sostituenti dei nostri sistemi. Ora dobbiamo constatare che, una volta avviluppata la realtà in *bit*, sono balzate in primo piano le forche caudine costituite di hardware fatto dei vecchi atomi: dai cavi sottomarini ai data center su cui è allocato il – materialissimo – *cloud* (Floridi, 2024).

Di questo la regolazione, anche europea, prende atto divenendo regolazione *by design* (come mostra l'art. 53 dell'AI Act appena adottato in materia di copyright). Le norme nuove non pretendono di sindacare la liceità dello *scraping* che sia avvenuto, ad esempio, in Giappone; ma impongono obblighi di comportamento e di auto-certificazione in capo di chi si valga del *machine learning* così conseguito (Peukert, 2024, discorre al riguardo di meta-obbligazioni; dubbi sull'idoneità di meccanismi di auto-certificazione sono ben esposti al § 5 del cap. 2 del lavoro).

Considerazioni come queste ci porterebbero a dire che per la regolazione della IA diventano sempre meno adatti fori, sedi, legislative o normative, nazionali ma anche regionali; e che si imporrebbe una collocazione in fori multilaterali, che concernono non solo il miliardo di abitanti del Nord del mondo, ma anche gli altri sette miliardi. Vi è una difficoltà: che l'IA sempre è stata anche, o forse soprattutto, militare, e oggi più che mai. Quindi ogni questione che la riguardi ha una dimensione geopolitica, presentemente condizionata dalla rivalità fra USA e Cina. La UE, che pur qualche segno di risveglio lo ha dato (vedi, ad esempio, cap. 3, § 5), resta relegata a un ruolo marginale.

Questa riflessione è il miglior punto di passaggio alla terza e ultima questione, quella della rilevazione degli interessi in campo. Il lavoro non trascura l'analisi dei molteplici interessi messi a repentaglio dalle modalità di funzionamento di ChatGPT e più in generale dell'AI; si potrebbe anche dire che qui si concentra l'attenzione degli autori. Mi pare tuttavia che a questo proposito resti moltissimo lavoro da fare. Nella letteratura in materia è infatti troppo frequente il cortocircuito logico che si trova fra le analisi di questi fenomeni e la

richiesta, vigorosa e spesso ben argomentata, di innovazioni normative, di regolazione. Il fatto è che il più delle volte nel ciò fare si tralascia l'analisi delle categorie di soggetti controinteressati ai poteri privati che mettono in campo le tecnologie nuove; analisi che però è indispensabile perché le regole non si generano da sole ma vengono adottate se reclamate, e reclamate in modo efficace. Dove naturalmente l'azione collettiva dei portatori di interessi contrapposti a quelli dei giganti dell'informazione trova sempre l'ostacolo rilevato molto tempo fa da Mançur Olson che pochi attori concentrati e dotati di risorse importanti normalmente prevalgono sui molti o moltissimi attori dispersi.

È almeno dal febbraio del 2019 che il Centro Nexa ha mostrato particolare interesse per questi *countervailing powers*. I lavoratori che perdono salario e potere, che appartengono alla gig economy, i *digital slaves*. Gli agricoltori privati del diritto di riparazione dei trattori che pur hanno acquistato. Tutti gli acquirenti di *digital devices*. Le categorie discriminate per ragioni di razza o di sesso o di codice postale. E così via.

Si tratta di lavorare sull'omogeneità e sulla frammentazione di questi portatori di interessi. Quanto all'omogeneità, il *digital slave* eritreo ha moltissimo in comune con quello francese o italiano; l'agricoltore francese con quello sudafricano. Quanto alla frammentazione, si tratta di prendere atto del fatto che oggi manca una classe sociale generale, come era il proletariato nella seconda metà dell'Ottocento e nella prima metà del Novecento, prima della fine della fabbrica fordista. Gli agricoltori non sono lavoratori dei servizi, la discriminazione algoritmica colpisce gruppi di soggetti anche molto diversi fra di loro.

Tenderei a pensare che qui ci sia ancora molto lavoro da fare, anche a partire dalle pagine di questo lavoro, così puntuali nell'individuare i portatori di interessi messi a repentaglio, violati o discriminati.

Naturalmente, il giurista vede innanzitutto gli istituti dell'azione collettiva, dalla *class action* alla *representative action*. E intravede la possibilità di aggregarli transnazionalmente, *cross-border*, come alcune direttive UE consentono.

Ma, oltre alla rilevazione dell'omogeneità degli interessi, della loro frammentazione e delle tecniche per superarla, resta il tema, politico, delle creazione di coalizioni fra categorie e gruppi di interessi e di alleanze trasversali. Sicuramente i partiti tradizionali sono a ciò inadeguati; sindacati e ONG hanno i loro problemi. Ma la costruzione di poteri contrapposti, *countervailing*, è istanza troppo importante per non puntarci su. Con molta umiltà: soprattutto noi europei dobbiamo comprendere che facciamo parte sì del miliardo di persone del Nord del mondo; ma anche e prima ancora degli otto miliardi di persone che sono esposte a questo frangente tecnologico.

Ringraziamenti

Come abbiamo dichiarato all'inizio, questo libro è il frutto dell'intelligenza collettiva di una serie di persone che, nel Centro Nexa su Internet e Società del Politecnico di Torino e nella mailing list da esso creata, dibattono da tempo su ChatGPT e l'IA generativa, criticandone con grande consapevolezza gli aspetti più problematici. In particolare, intendiamo ringraziare per aver condiviso con noi i loro saperi Daniela Tafani, Enrico Nardelli, Roberto Dolci, Alberto Cammozzo, Giacomo Tesio, Stefano Quintarelli, Giuseppe Attardi, Antonio Casilli, Marco Fioretti, Stefano Zacchiroli, Marco Calamari, Giovanni Biscuolo, Guido Vetere, Maria Chiara Pievatolo, Carlo Blengino, Ugo Pagallo, Massimo Durante. Ringraziamo anche i direttori del Centro Nexa, Juan Carlos De Martin e Marco Ricolfi, per averci spronato e sostenuto dal momento del concepimento del nostro progetto alla sua pubblicazione, corredata anche dalla loro prefazione e postfazione. Ringraziamo i fellow e i garanti dello stesso Centro Nexa, come Antonio Vetrò e Anna Masera, che hanno letto e commentato il nostro testo. Infine, ringraziamo Giulio De Petra, per averci suggerito l'idea di scriverlo.

Bibliografia

ACEMOĞLU Daron, RESTREPO Pascual, «Artificial Intelligence, Automation, and Work», in Ajay Agrawal, Joshua Gans, Avi Goldfarb (a cura di), *The Economics of Artificial Intelligence: An Agenda*, National Bureau of Economic Research, University of Chicago Press, Chicago, 2019.

ARENDT Hannah, *On Revolution*, Penguin, New York, 1963.

BALBI Gabriele, *L'ultima ideologia: Breve storia della rivoluzione digitale*, Laterza, Roma-Bari, 2022.

BALDWIN Peter, *The Copyright Wars: Three Centuries of Trans-Atlantic Battle*, Princeton University Press, Princeton, 2014.

BARNETT Jonathan, *Innovators, Firms, and Markets: The Organizational Logic of Intellectual Property*, Oxford University Press, New York, 2021.

BASSETT Caroline, *The computational therapeutic: exploring Weizenbaum's ELIZA as a history of the present*, «AI & Society», vol. 34, 2019, pp. 803-812, <https://doi.org/10.1007/s00146-018-0825-9>.

BEALS Ralph L., HOIJER Harry, *An Introduction to Anthropology*, The Macmillan Company, New York, 1953.

BECKER Cristoph, *Insolvent. How to Reorient Computing for Just Sustainability*, The MIT Press, Cambridge, 2023.

BENDER Emily M., GEBRU Timnit, MCMILLAN-MAJOR Angelina, MITCHELL Margaret, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big*, FACCT '21, Association for Computing Machinery, New York, 2021, <doi: 10.1145/3442188.3445922>.

BODEN Margaret A., *L'intelligenza artificiale*, il Mulino, Bologna, 2019.

BORGHI Maurizio, *Why tech giants have little to lose (and lots to win) from new EU copyright law*, «The Conversation», 19-9-2018, <https://theconversation.com/why-tech-giants-have-little-to-lose-and-lots-to-win-from-new-eu-copyright-law-103374>.

BORGHI Maurizio, KARAPAPA Stavroula, *Copyright and mass digitization*, Oxford University Press, Oxford, 2013.

BOSTROM Nick, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, Oxford, 2014.

BUBECK Sébastien et al., *Sparks of Artificial General Intelligence: Early experiments with GPT-4*, 2023, <https://arxiv.org/pdf/2303.12712>.

CASILLI Antonio, *En attendant les robots. Enquête sur le travail du clic*, Éditions du Soleil, Paris, 2019.

CHOMSKY Noam, *A Review of Skinner's Verbal Behavior*, «Language», vol. 35, n. 1, 1959, pp. 26-58.

CHUN Wendy Hui Kyong, *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition*, The MIT Press, Cambridge, 2021.

CIOTTI Fabio, *I Large Language Models e i «problemi difficili»*, <https://infouma.hypotheses.org/2429>, 2023.

COHEN Julie, *Between Truth and Power: The Legal Constructions*

of Informational Capitalism, Oxford University Press, New York, 2019.

COLE David, «The Chinese Room Argument», in Edward N. Zalta, Uri Nodelman (a cura di), *The Stanford Encyclopedia of Philosophy*, Stanford University, Stanford, 2023, <https://plato.stanford.edu/archives/win2020/entries/chinese-room/>.

CORDESCHI Roberto, *The Discovery of the Artificial: Behavior, Mind and Machines Before and Beyond Cybernetics*, Kluwer Academic, Dordrecht, 2002.

CRAWFORD Kate, *Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence*, Yale University Press, New Haven-London, 2021.

DA EMPOLI Stefano, *L'economia di ChatGPT. Tra false paure e veri rischi*, Egea, Milano, 2023.

DE BAGGIS Mafe, PULIAFITO Alberto, *In principio era ChatGPT. Intelligenze artificiali per testi, immagini, video e quel che verrà*, Apogeo, Milano, 2023.

DELFANTI Alessandro, POSADA Julian, *Boxing AI and Amazon Fulfillment Centers*, in Elinor Wahal (a cura di), *Unboxing AI. Understanding Artificial Intelligence*, Fondazione Giangiacomo Feltrinelli, Milano, 2021, pp. 21-33.

DE MARTIN Juan Carlos, *Contro lo smartphone. Per una tecnologia più democratica*, ADD, Torino, 2023.

ELOUNDOU Tyna, MANNING Sam, MISHKIN Pamela, ROCK Daniel, *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models*, working paper, 2023, <https://arxiv.org/pdf/2303.10130>.

FERRARO Giulio, *Fondamenti di teoria sociosemiotica. La teoria «neoclassica»*, Aracne, Roma, 2012.

FLORIDI Luciano, *Etica dell'intelligenza artificiale. Sviluppi, opportunità, sfide*, Cortina, Milano, 2022.

FLORIDI Luciano, *The hardware turn in the digital discourse: an analysis, explanation, and potential risk*, «Philosophy and Technology», 12-2-2024, <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4723312>.

FLORIDI Luciano, CABITZA Federico, *Intelligenza Artificiale. L'uso delle nuove macchine*, Bompiani, Milano, 2021.

FREY Carl B., OSBORNE Michael A., *The Future of Employment: How Susceptible are Jobs to Computerisation?*, working paper, Oxford Martin School, University of Oxford, Oxford, 2013.

FRIGERIO Aldo, *Filosofia del linguaggio*, Apogeo, Milano, 2011.

FRISCHMANN Brett, SELINGER Evan, *Re-engeneering Humanity*, Cambridge University Press, Cambridge, 2020.

GARDNER Howard, *Frames of Mind. The Theory of Multiple Intelligences*, Basic Books, New York, 1983.

GIBLIN Rebecca, DOCTOROW Cory, *Chokepoint Capitalism. How Big Tech and Big Content Captured Creative Labor Markets and How We'll Win Them Back*, Beacon Press, Boston, 2022.

GIOVANNOLI Renato, *La scienza della fantascienza*, Bompiani, Milano, 2015.

GIRAUDO Marco, «Legal Bubbles», in Alan Marciano, Giovanni Battista Ramello (a cura di), *Encyclopedia of Law and Economics*, Springer, New York, 2022, <https://doi.org/10.1007/978-1-4614-7883-6_786-1>.

GIRAUDO Marco, FOSCH-VILLARONGA Eduard, MALGIERI Gianclaudio, *Competing Legal Futures*, «German Law Journal», 2024.

GOTTFREDSON Linda S., *Mainstream Science on Intelligence: An Editorial With 52 Signatories, History, and Bibliography*, «Intelligence», vol. 24, n. 1, Ablex Publishing Corporation, NY, 1997, pp. 13-23.

GRICE Paul H., *Logic and Conversation*, «Syntax and semantics», vol. 3, *Speech Acts*, pp. 41-58, 1975.

HARARI Yuval Noah, *From Animals Into Gods. A Brief History of Humankind*, Harvill Secker, London, 2014.

HARARI Yuval Noah, *Homo Deus. A Brief History of Tomorrow*, Harvill Secker, London, 2016.

HARARI Yuval Noah, *Why technology favors tyranny*, «The Atlantic», October issue, *Is Democracy Dying?*, Washington, 2018.

HILL Kashmir, *Your Face Belongs to Us: The Secretive Startup Dismantling Your Privacy*, Simon & Schuster, New York, 2023.

HOFFMAN Reid, *Impromptu. Amplifying Our Humanity Through AI*, Dallepedia LLC, 2023.

HORCHAK Oleksandr V., GIGER Jean-Christophe, CABRAL Maria, POCHWATKO Grzegorz, *From demonstration to theory in embodied language comprehension. A review*, «Cognitive Systems Research», vol. 29-30, 2014, pp. 66-85, doi:10.1016/j.cogsys.2013.09.002.

HUME David, *The Natural History of Religion*, A. and H. Bradlaugh Bonner, London, 1757 (trad. it. di Alba Graziano, *La religione naturale*, Editori Riuniti, Roma, 1995).

HUTTER Marcus, *Towards a Universal Theory of Artificial Intelligence based on Algorithmic Probability and Sequential Decision Theory*, «Lecture Notes in Artificial Intelligence», LNAI 2167, ECML 2001, pp. 226-238, 2001, <https://doi.org/10.48550/arXiv.cs/0012011>.

KAPLAN Jerry, *Generative AI. Conoscere, capire e usare l'intelligenza artificiale generativa*, Luiss University Press, Roma, 2024.

KATZ Daniel M., BOMMARITO Michael J., GAO Shang, ARREDONDO Pablo, *GPT-4 Passes the Bar Exam*, 2023, <http://dx.doi.org/10.2139/ssrn.4389233>.

LANGLEY Patrick, *The Cognitive Systems Paradigm*, «Advances in Cognitive Systems», vol. 1, 2012, pp. 3-13.

LEGG Shane, HUTTER Marcus, *A Collection of Definitions of Intelligence*, in Ben Goetzel, Pei Wang (a cura di), *Advances*

in Artificial General Intelligence. Concepts, Architecture, and Algorithms, vol. 157, 2007, pp. 17-24.

LIETO Antonio, *Cognitive Design for Artificial Minds*, Routledge, London-New York, 2021.

LINDZEY Gardner, THOMPSON Richard F., SPRING Bonnie, *Psychology*, Worth Publisher, New York, 1975.

MARCUS Gary, SOUTHEN Reid, *Generative AI Has a Visual Plagiarism Problem. Experiments with Midjourney and DALL-E 3 show a copyright minefield*, «IEEESPECTRM», 6-1-2024, <https://spectrum.ieee.org/midjourney-copyright>.

MORRIS Charles W., *Foundations of the Theory of Signs*, University of Chicago Press, Chicago, 1938.

NATALE Simone, *Deceitful Media. Artificial Intelligence and Social Life After the Turing Test*, Oxford University Press, New York, 2021.

NUMERICO Teresa, *Big data e algoritmi. Prospettive critiche*, Carocci, Roma, 2021.

O'NEIL Cathy, *Weapons of math destruction: how big data increases inequality and threatens democracy*, Crown, New York, 2016.

ORWELL George, *1984* (1949), Einaudi, Torino, 2021.

PASQUALE Frank, *New Laws of Robotics: Defending Human Expertise in the Age of AI*, Harvard University Press, Cambridge 2020.

PATHAK Gaurav, *Manifestly Made Public: Clearview and GDPR*, «European Data Protection Law Review», vol. 8, n. 3, 2022.

PETERS Uwe, *What Is the Function of Confirmation Bias?*, «Erkenn», vol. 87, 2022, pp. 1351-1376, <https://doi.org/10.1007/s10670-020-00252-1>.

PEUKERT Alexander, *Copyright in the Artificial Intelligence Act – A Primer*, «GRUR International», vol. 73, n. 6, giugno 2024, <https://ssrn.com/abstract=4771976> o <http://dx.doi.org/10.2139/ssrn.4771976>.

POPPER Karl R., *Conjectures and Refutations: The Growth of*

Scientific Knowledge, Basic Books, London-New York, 1962.

QUINTARELLI Stefano (a cura di), *Intelligenza artificiale. Cos'è davvero, come funziona, che effetti avrà*, Bollati Boringhieri, Torino, 2020.

RUSSELL Stuart, NORVIG Peter, *Intelligenza artificiale. Un approccio moderno*, Pearson Prentice Hall, Milano-Torino, 2010.

RONCAGLIA Gino, *L'architetto e l'oracolo. Forme digitali del sapere da Wikipedia a ChatGPT*, Laterza, Roma, 2023.

SADIN Éric, *L'intelligence artificielle ou l'enjeu du siècle: anatomie d'un antihumanisme radical*, L'Echappée, Paris, 2021.

SANTANGELO Antonio, *Sociosemiotica dell'audiovisivo*, Aracne, Roma, 2013.

SAUSSURE Ferdinand de, *Cours de Linguistique Générale*, Librairie Payot & Cie, Paris, 1916.

SEARLE John R., *Minds, brains, and programs*, «Behavioral and Brain Sciences», vol. 3, n. 3, 1980, pp. 417-424, <doi:10.1017/S0140525X00005756>.

TAFANI Daniela, *L'«etica» come specchietto per le allodole. Sistemi di intelligenza artificiale e violazioni dei diritti*, «Bollettino telematico di filosofia politica», 2023, pp. 1-13, <https://doi.org/10.5281/zenodo.7799775>.

TEGMARK Max, *Life 3.0, Being Human in the Age of Artificial Intelligence*, Knopf, New York, 2017.

VAN CLEYNENBREUGEL Pieter, *By-design regulation and European Union law: opportunities, challenges and the way ahead*, in Maurizio Borghi, Roger Brownsword (a cura di), *Law, Regulation and Governance in the Information Society. Informational Rights and Informational Wrongs*, Routledge, London, 2023, 49 ff.

VAN ROOIJ Iris, GUEST Olivia, ADOLFI Federico G., DE HAAN Ronald, KOLOKOLOVA Antonina, RICH Patricia, *Reclaiming*

AI as a theoretical tool for cognitive science, 2023, <https://doi.org/10.31234/osf.io/4cbuv>.

VAROUFAKIS Yanis, *Tecnofeudalesimo. Cosa ha ucciso il capitalismo*, La Nave di Teseo, Milano, 2023.

VIOLI Patrizia, *Significato ed esperienza*, Bompiani, Milano, 1997.

VOLOKH Eugene, *Large Libel Models? Liability for AI Output*, UCLA School of Law, Public Law Research Paper n. 23-17, 2023, <https://ssrn.com/abstract=4546063>.

WINNER Langdon, *Do Artifacts Have Politics?*, «Daedalus», *Modern Technology: Problem or Opportunity?*, vol. 109, n. 1, 1980, pp. 121-136.

ZUBOFF Shoshana, *The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power*, Profile Books, London, 2019.

Sitografia

BEDOYA Alvaro M., *«Early Thoughts on Generative AI», Prepared Remarks of Commissioner Alvaro M. Bedoya Before the International Association of Privacy Professionals*, «Federal Trade Commission», 5-4-2023, <https://www.ftc.gov/system/files/ftc_gov/pdf/Early-Thoughts-on-Generative-AI-FINAL-WITH-IMAGES.pdf>.

BENDER Emily, *Talking about a 'schism' is ahistorical*, «Medium», 5-7-2023, <https://medium.com/@emilymenonbender/talking-about-a-schism-is-ahistorical-3c454a77220f>.

BENNATO Davide, *AI, diffidiamo dal «lungotermismo»: fa gli interessi della Silicon Valley*, «Agenda Digitale», 24-5-2023, <https://www.agendadigitale.eu/cultura-digitale/ai-diffidiamo-dal-lungotermismo-plasma-il-dibattito-sugli-interessi-della-silicon-valley/>.

BROCK David C., *Our Censor, Ourselves: Commercial Content Moderation*, «The Los Angeles Review of Books»,

25-7-2019, <https://lareviewofbooks.org/article/our-censors-ourselves-commercial-content-moderation/>.

CASILLI Antonio, *L'intelligenza artificiale renderà sempre più precario il mondo del lavoro*, «Domani», 23-3-2023, <https://www.editorialedomani.it/tecnologia/intelligenza-artificiale-chat-gpt-open-ai-mondo-del-lavoro-precario-robot-icwwg8ht>.

CHIANG Ted, *ChatGPT is a blurry jpeg of the web. OpenAI's chatbot offers paraphrases, whereas Google offers quotes. Which do we prefer?,* «The New Yorker», 9-2-2023, <https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web>.

CHOMSKY Noam, *The False Promise of ChatGPT*, «The New York Times», 8-3-2023, <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>.

COMMUNIA, *Policy paper #15 on using copyrighted works for teaching the machine*, 26-4-2023, <https://communia-association.org/policy-paper/policy-paper-15-on-using-copyrighted-works-for-teaching-the-machine/>.

ELIOT Lance, *Generative AI ChatGPT As Masterful Manipulator Of Humans, Worrying AI Ethics And AI Laws*, «Forbes», 1-3-2023, <https://www.forbes.com/sites/lanceeliot/2023/03/01/generative-ai-chatgpt-as-masterful-manipulator-of-humans-worrying-ai-ethics-and-ai-law/?sh=7eec05231d66>.

EMERGING TECHNOLOGY FOR THE ARXIV, *King–Man+Woman=Queen: The Marvelous Mathematics of Computational Linguistics*, «MIT Technology Review», 17-9-2015, <https://www.technologyreview.com/2015/09/17/166211/king-man-woman-queen-the-marvelous-mathematics-of-computational-linguistics>.

FIORE Francesca, *Il condizionamento operante: il paradigma sperimentale di Skinner*, «State of Mind», 19-11-2015, <https://www.stateofmind.it/2015/11/condizionamento-operante-skinner/>.

FUTURE OF LIFE INSTITUTE, *Pause Giant AI Experiment: An Open Letter*, «Futureoflife», 22-3-2023, <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>.

GARANTE PER LA PROTEZIONE DEI DATI PERSONALI, *Provvedimento del 2 febbraio 2023*, «garanteprivacy.it», <https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9852214>.

GEBRU Timnit, BENDER Emily M., MCMILLAN-MAJOR Angelina, MITCHELL Margaret, *Statement from the listed authors of Stochastic Parrots on the «AI pause» letter*», «Dair», 31-3-2023 <https://www.dair-institute.org/blog/letter-statement-March2023/>.

GRUPPO DI ESPERTI DI ALTO LIVELLO SULL'INTELLIGENZA ARTIFICIALE, *Una definizione di IA: principali capacità e discipline scientifiche*, 2019, <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60673 >.

GUTIÉRREZ Juan D., *ChatGPT in Colombian Courts: Why we need to have a conversation about the digital literacy of the judiciary*, «Verfassungsblog», 23-2-2023, <https://verfassungsblog.de/colombian-chatgpt/>.

HAO Karen, *We read the paper that forced Timnit Gebru out of Google, Here's what it says*, «MIT Technology Review», 4-12-2020, <https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/>.

HARARI Yuval Noah, *You Can Have the Blue Pill or the Red Pill, and We're Out of Blue Pills*, «The New York Times», 24-3-2023, <https://www.nytimes.com/2023/03/24/opinion/yuval-harari-ai-chatgpt.html>.

KAPOOR Sayash, NARAYANAN Arvind, *GPT-4 and professional benchmarks: the wrong answer to the wrong question*, «AI Snake Oil», 20-3-2023, <https://www.aisnakeoil.com/p/gpt-4-and-professional-benchmarks>.

KAPOOR Sayash, NARAYANAN Arvind, *A misleading open letter about sci-fi AI dangers ignores the real risks*, «AI Snake Oil», 29-3-2023, <https://www.aisnakeoil.com/p/a-misleading-open-letter-about-sci>.

LYNDON Cory, *Wysa AI-chatbot app to be rolled out to teenagers across west London*, «Digital Health», 21-9-2023, <https://www.digitalhealth.net/2022/09/wysa-ai-chatbot-teens-west-london/>.

METZ Cade, *A Second Google AI researcher says the company fired her*, «The New York Times», 19-2-2021, <https://www.nytimes.com/2021/02/19/technology/google-ethical-artificial-intelligence-team.html>.

PERRIGO Billy, *OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic*, «Time», 18-1-2023, <https://time.com/6247678/openai-chatgpt-kenya-workers/?fbclid=PAAabiRhajL9uq0H8WJVr38iZdo0UmZ7DrRPjbWq4lu8Oyhvj9CNtu5mG9uJs>.

QUARANTA Domenico, *Demistificare l'Intelligenza Artificiale*, «Flash Art», 3-7-2023, <https://flash---art.it/article/intelligenza-artificiale/>.

QUINTARELLI Stefano, *Let's forget the term AI. Let's call them Systematic Approaches to Learning Algorithms and Machine Inferences (SALAMI)*, «Quinta's weblog», 24-11-2019, <https://blog.quintarelli.it/2019/11/lets-forget-the-term-ai-lets-call-them-systematic-approaches-to-learning-algorithms-and-machine-inferences-salami/>.

SARRO Doug, *Using gaming tactics in apps raises new legal issues*, «The Conversation», 1-5-2022, <https://theconversation.com/using-gaming-tactics-in-apps-raises-new-legal-issues-181575>.

SCORZA Guido, *AI, è ora di difendere i nostri dati dalla raccolta massiva: ecco come*, «Agenda Digitale», 18-6-2024, <https://www.agendadigitale.eu/sicurezza/privacy/linsostenibile-scorrettezza-del-training-degli-algoritmi-come-difendere-i-nostri-dati/>.

SIMONITE Tom, *What Really Happened When Google Ousted*

Timnit Gebru, «Wired», 8-6-2021, <https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/>.

SMUHA Nathalie A., DE KETELAERE Mieke, COECKELBERGH Mark, DEWITTE Pierre, POULLET Yves, *Open Letter: We are not ready for manipulative AI – urgent need for action*, «Ku Leuven AI Summer School», 31-3-2023, <https://www.law.kuleuven.be/ai-summer-school/open-brief/open-letter-manipulative-ai>.

STEIN-PERLMAN Zach, WEINSTEIN-RAUN Benjamin, GRACE Katja, *2022 Expert Survey on Progress in AI*, «AI Impacts», 3-8-2022, <https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/>.

THE REPLIKA TEAM, *Replika Quest*, «Replika», 4-4-2023, <https://blog.replika.com/posts/replika-quests>.

VINCENT James, *AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit*, «The Verge», 16-1-2023, <https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart>.

VINSEL Lee, *You're Doing It Wrong: Notes on Criticism and Technology Hype*, «Medium», 1-2-2021, <https://sts-news.medium.com/youre-doing-it-wrong-notes-on-criticism-and-technology-hype-18b08b4307e5>.

WALKER Lauren, *Belgian man dies by suicide following exchanges with chatbot*, «The Brussels Times», 28-3-2023, <https://www.brusselstimes.com/belgium/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt>.

WESTKAMP Guido, *Borrowed Plumes: Taking Artists' Interests Seriously in Artificial Intelligence Regulation*, 18-4-2024, <https://ssrn.com/abstract=4799179>.

WYSA, «Wysa.com», <https://www.wysa.com/>.

finito di stampare nel mese di gennaio 2025
presso Printì, Manocalzati (AV)
per conto di elèuthera, via Jean Jaurès 9, Milano