Questo *white paper* scaturisce dal lavoro del Tavolo Interdipartimentale sull'Intelligenza Artificiale istituito presso l'Università degli Studi di Bergamo nel novembre 2023. Il documento mira a stimolare un dialogo interdisciplinare sull'intelligenza artificiale attraverso la condivisione di un'agenda di ricerca.

ISBN: 978-88-97413-83-7
DOI: 10.13122/978-88-97413-83-7

INTELLIGENZA ARTIFICIALE: UN'AGENDA DI RICERCA

INTELLIGENZA ARTIFICIALE: UN'AGENDA DI RICERCA

White paper del Tavolo Interdipartimentale sull'Intelligenza Artificiale dell'Università degli Studi di Bergamo

A cura di
Maria Francesca Murru

UNIVERSITÀ DEGLI STUDI DI BERGAMO

2024

INTELLIGENZA ARTIFICIALE: UN'AGENDA DI RICERCA

White paper
Tavolo Interdipartimentale sull'Intelligenza Artificiale dell'Università degli Studi di Bergamo

A cura di Maria Francesca Murru

Università degli Studi di Bergamo

2024

Intelligenza Artificiale: un'agenda di ricerca / edited by Maria Francesca Murru – Bergamo: Università degli Studi di Bergamo, 2024.
ISBN: 978-88-97413-83-7
DOI: 10.13122/978-88-97413-83-7

Picture in the cover generated by AI

Book's initiative:
Tavolo Interdipartimetale sull'Intelligenza Artificiale
Università degli Studi di Bergamo

https://aisberg.unibg.it/handle/10446/276449

Progetto grafico: Servizi Editoriali – Università degli Studi di Bergamo
Università degli Studi di Bergamo
via Salvecchio, 19
24129 Bergamo
Cod. Fiscale 80004350163
P. IVA 01612800167

Sommario

Introduzione

Questo white paper scaturisce dal lavoro del Tavolo Interdipartimentale sull'Intelligenza Artificiale istituito presso l'Università degli Studi di Bergamo nel novembre 2023. Il documento mira a stimolare un dialogo interdisciplinare sull'intelligenza artificiale attraverso la condivisione di un'agenda di ricerca.

Nei discorsi pubblici che circolano più diffusamente, l'intelligenza artificiale è spesso oggetto di una rappresentazione emotivamente carica ma imprecisa. Pur in presenza di svariate declinazioni e differenziazioni interne, le cornici interpretative prevalenti presentano la caratteristica comune dell'indistinzione totalizzante. Ci siamo ormai abituati a pensare all'intelligenza artificiale al singolare e in astratto, pur nella consapevolezza che questa comune definizione raccolga al suo interno una varietà di tecnologie con livelli di rischi e potenzialità molto diversi. Continuiamo inoltre a considerarla come un'entità autoevidente, come un fatto indistinto ma inconfutabile, a cui attribuiamo il potere di impattare contemporaneamente e indistintamente su tutti gli aspetti dell'economia, della politica, della formazione e della cultura.

Questo peculiare connubio di vaghezza e potenza non è un'esclusiva dei discorsi mediali divulgativi. La sua eco risuona anche in molte delle dichiarazioni dei policymaker, laddove prevalgono definizioni circolari per cui l'intelligenza artificiale è quel sistema considerato intelligente perché riesce a svolgere attività che fino a poco tempo prima erano considerate prerogativa esclusiva dell'essere umano. Come acutamente dimostrato da Noortje Marres e il suo gruppo di ricerca (2024)[1], in questo momento storico il discorso pubblico sull'intelligenza artificiale è pervaso da controversie artefatte. Le controversie

[1] Marres, N., Castelle, M., Gobbo, B., Poletti, C., & Tripp, J. (2024). AI as super-controversy: Eliciting AI and society controversies with an extended expert community in the UK. Big Data & Society, 11(2).

scientifiche sono momenti ineludibili del processo di innovazione tecnologica, biforcazioni nella traiettoria di sviluppo che possono trasformarsi in occasioni di democratizzazione. Nel caso dell'intelligenza artificiale, i produttori e gli esperti sono anche i suoi maggiori critici; coloro che insistono sulla sua intrinseca capacità di rivoluzionare la scienza, l'economia e l'umanità stessa sono gli stessi che ne enfatizzano i cosiddetti "rischi esistenziali". Le potenzialità di democratizzazione del dissenso si trovano così a essere neutralizzate da un gioco delle parti che esalta l'ineluttabilità del processo senza indicare reali percorsi di intervento. L'esito di questo clima discorsivo non è solo la naturalizzazione e decontestualizzazione della tecnologia ma anche il suo inquadramento in una narrazione escludente, come un fatto su cui è possibile intervenire solo a posteriori.

In questa stratificazione di rappresentazioni retoriche, l'università è chiamata certamente a offrire il proprio contributo di conoscenza specialistica e verificata ma anche a interrogarsi sul senso e sulla responsabilità ermeneutica che ogni discorso pubblico può esercitare in un momento di transizione così delicato e decisivo. Il Tavolo Interdipartimentale dell'Università di Bergamo nasce con questo orizzonte di azione su iniziativa del Rettore Sergio Cavalieri. A guidarlo è l'intuizione che la natura trasversale e potenzialmente rivoluzionaria dell'intelligenza artificiale richieda la fertilizzazione incrociata dei diversi saperi disciplinari presenti nel nostro ateneo. Sul solco di un metodo di lavoro già consolidato su altri fronti strategici, il Rettore ha chiesto al Tavolo di impostare un processo di collaborazione orientato contemporaneamente al rafforzamento delle collaborazioni interdisciplinari tra tutti i dipartimenti e all'avvio di un dialogo sistematico con le realtà istituzionali, sociali, culturali e produttive del territorio interessate al tema dell'innovazione tecnologica. Le attività del Tavolo sono state dunque impostate in modo che il gruppo di lavoro interdipartimentale agisse da incubatore di un processo di interdisciplinarità destinato a essere esteso in prospettiva a tutto l'ateneo. Si è trattato in primo luogo di ibridare sguardi e approcci di ricerca per giungere a una

concettualizzazione condivisa dell'intelligenza artificiale che potesse funzionare da piattaforma operativa per progetti futuri. Tutti i componenti del Tavolo hanno messo in gioco una specifica prospettiva sull'intelligenza artificiale, nella consapevolezza che essa fosse contemporaneamente espressione della propria sensibilità personale, del proprio percorso professionale, e insieme un ponte con gli orientamenti di ricerca e le sensibilità del dipartimento di afferenza. Allo scambio è seguita l'individuazione di linee di convergenza e questioni trasversali con l'obiettivo di costruire una prima e provvisoria agenda di priorità che coincide con i sei capitoli di questo white paper. Abbiamo deciso di affrontare in primo luogo la questione definitoria. Nella consapevolezza di attraversare un terreno scivoloso, popolato di etichette instabili e determinazioni ideologiche, gli aspetti definitori sono stati dipanati a partire da una prospettiva storica che ripercorre tappe, stagnazioni e accelerazioni del processo di ricerca e sviluppo da cui è scaturito quel cluster di tecnologie, che oggi ricade nell'etichetta dell'intelligenza artificiale. In seconda battuta, abbiamo ritenuto importante mettere a fuoco uno degli aspetti tecnici più caratteristici e insieme maggiormente carichi di conseguenze politiche e culturali: l'opacità intesa come inintelligibilità o inaccessibilità semantica (non necessariamente ineludibile e definitiva) dei processi di elaborazione che avvengono all'interno di una rete neurale artificiale e le sue ricadute legali. Il terzo capitolo prende in considerazione la questione che più profondamente sollecita gli immaginari culturali e politici: la generatività dell'intelligenza artificiale intesa come una nuova forma di creatività che ridefinisce pratiche e valori dell'autorialità, consente la sperimentazione di nuove possibilità estetiche e insieme obbliga a prendere atto di quanto qualsiasi potenzialità tecnologica dipenda, nel suo concreto dispiegarsi, da preesistenti logiche economiche e culturali.

Un'altra novità dei sistemi basati sull'intelligenza artificiale riguarda la loro capacità di simulare una interazione comunicativa, diventando interlocutori e produttori di significati e relazioni. A questo aspetto, alle sue ricadute teoriche e alle sue

potenziali applicazioni nell'ambito della robotica sociale è dedicato il quarto capitolo. Il quinto approfondimento esplora le diverse sfaccettature della dimensione economica, gli effetti sul contesto macroeconomico e sul mercato del lavoro ma anche gli aspetti chiave legati allo sviluppo del settore, la struttura di mercato e gli asset decisivi che hanno condizionato e continuano a condizionare in modo sostanziale i processi di ricerca e sviluppo. Quali benefici possa portare e quali ostacoli possa incontrare l'introduzione di tecnologie legate all'intelligenza artificiale nel tessuto produttivo è l'argomento al centro del sesto capitolo. Tenendo presente l'intera catena del valore, ovvero l'insieme di attività che permette alle aziende di creare e gestire il valore, l'analisi passa in rassegna le possibili implementazioni nei processi primari e di supporto, evidenziandone vantaggi applicativi e ostacoli potenziali prevalentemente in termini di mancanza di competenze, costi elevati e mancanza di dati di qualità.

La scelta di individuare una lista di priorità di ricerca risponde alla necessità di offrire un controcanto all'indistinzione totalizzante che permea molti dei discorsi pubblici sull'intelligenza artificiale. L'obiettivo è stato quello di rispondere alla generalizzazione con la differenziazione, all'astrazione con la specificazione, sia attingendo al bacino di conoscenze specialistiche, sia orientando lo sguardo in direzione prospettica laddove le questioni inquadrate si rivelano troppo complesse e articolate per poter essere esaurite in un ambito disciplinare circoscritto. L'interdisciplinarità di questo Tavolo è nata dunque come tentativo di rispondere a un processo di innovazione tecnologica pervasivo e trasversale ai diversi ambiti della società. È cresciuta nel corso dei primi mesi di attività sotto forma di pratica e di esperienza concreta di confronto e discussione alla ricerca di un linguaggio comune. Prosegue ora con questo white paper, con il quale desideriamo idealmente estendere il dialogo a tutte le colleghe e i colleghi interessati.

Nell'esperienza di un confronto costante, il Tavolo si è spesso soffermato a riflettere sul senso e sul metodo

dell'interdisciplinarità. Le discipline sono schemi che caratterizzano, classificano e specializzano. Attraverso un canone teorico e un insieme di metodi condivisi, orientano l'attenzione verso determinati fenomeni, mettendone in risalto alcuni aspetti e portandone in ombra altri, inestricabilmente intrecciati nella realtà ma inevitabilmente isolati dall'atto analitico. L'ibridazione degli sguardi è dunque un'esigenza di complementarità posta in primo luogo dalla realtà stessa nella sua complessità irriducibile. Eppure, la sua realizzazione entra in conflitto con quell'esigenza di specializzazione che sola garantisce la verifica scientifica. La questione è piuttosto ampia e meriterà altri approfondimenti. Per il momento, una bussola utile a ordinare le possibili traiettorie di un percorso trasversale agli ambiti di ricerca si trova nella distinzione tra interdisciplinarità ontologica e interdisciplinarità gerarchica introdotta da Andrew Barry et al. (2008)[2]. La prima è la più ambiziosa e non sempre giustificabile pragmaticamente. Scaturisce dalla pratica scientifica stessa nella sua autonomia e non si limita a giustapporre gli sguardi disciplinari ma li ibrida, modificando la concettualizzazione stessa dell'oggetto di ricerca. L'intelligenza artificiale, le sue caratteristiche tecniche e la sua promessa di pervasività, sembrerebbero effettivamente spingere in questa direzione. In quanto sistema tecnologico che estrae pattern di regolarità da larghi dataset, sulla base di aggiustamenti dei parametri che seguono a feedback interni o esterni, l'intelligenza artificiale si fonda su logiche di classificazione e quantificazione. Lungi dall'essere principi meramente tecnici, tali logiche sono culturali e storicamente collocate. La loro utilità è indubbia quando si tratta di facilitare i processi decisionali in ambito produttivo, di aumentarne efficienza e produttività minimizzando la necessità di ricorrere a giudizi individuali. Altrettanto innegabile è la loro problematicità quando si tratta di applicarle alle pratiche culturali e alle relazioni sociali, alle emozioni e alle decisioni collettive, dove per definizione i conti non tornano, perché il disordine, l'imprevedibilità, l'incompletezza

[2] Barry, A., Born, G., & Weszkalnys, G. (2008). Logics of interdisciplinarity. Economy and Society, 37(1), 20–49.

sono ineliminabili e spetta solo ai principi e alle procedure democratiche il compito di esprimerli e incanalarli. Esiste dunque una dialettica interna tra le declinazioni operative di queste logiche, il cui studio potrebbe effettivamente richiedere una interdisciplinarità ontologica.

La seconda forma deriva dalla spinta crescente a progettare e sviluppare la ricerca scientifica in stretta interazione con gli stakeholder di riferimento, rendendola in qualche modo utile e funzionale all'innovazione sociale ed economica. L'integrazione tra discipline funziona qui secondo un principio di completamento reciproco: un approccio va a colmare il vuoto dell'altro, come quando le scienze naturali o le discipline ingegneristiche coinvolgono i sociologi o gli antropologi perché necessitano di un approfondimento sui fattori sociali o culturali. La progettazione di questa interdisciplinarità nel campo dell'intelligenza artificiale può trovare spunti utili in alcuni strumenti analitici concepiti dalla comunità internazionale per gestire la multidimensionalità del fenomeno. Ne è un esempio chiaro il *Framework for the Classification of AI systems* elaborato dall'OECD nel 2022 e inizialmente pensato come strumento di supporto per policymaker e legislatori[3]. L'utilità del modello risiede nella sua capacità di supportare chiunque abbia bisogno di navigare la geometria variabile di generalizzazione e specificazione che caratterizza l'intelligenza artificiale, mostrando come una visione chiara non possa prescindere dall'analisi in contemporanea di cinque dimensioni coinvolte: 1) individui e gruppi che interagiscono con l'intelligenza artificiale (e/o che sono potenzialmente condizionati dalla sua implementazione) e il conseguente impatto sui diritti umani e sull'ambiente, sul benessere delle persone, sulla democrazia e sul lavoro; 2) l'ambito economico in cui ha luogo l'implementazione tecnologica e il settore applicativo di riferimento (per esempio, la sanità, la finanza, il settore manifatturiero); 3) quali dati sono usati come input, la loro provenienza, qualità e proprietà; 4) le caratteristiche

[3] https://www.oecd.org/en/publications/2022/02/oecd-framework-for-the-classification-of-ai-systems_336a8b57.html.

tecniche del modello utilizzato e i suoi livelli di opacità; 5) gli output attesi e gli obiettivi raggiunti con particolare attenzione alla valutazione del livello di autonomia dall'intervento umano e della performance raggiunta (supporto alle decisioni, personalizzazione, riconoscimento, classificazione delle emozioni, previsione, ottimizzazione). Altrettanto utile appare una prima classificazione della ricerca finanziata dall'*European Research Council* (ERC)[4] sul tema dell'intelligenza artificiale sulla base della pertinenza con specifiche aree di policy individuate come prioritarie dalla Commissione Europea, dal Parlamento Europeo e dall'OECD, da cui risulta che una buona percentuale di progetti si è finora focalizzata sulle applicazioni dell'intelligenza artificiale per la salute e la transizione energetica, subito seguita da democrazia, agroalimentare, sistemi giuridici, lavoro ed educazione.

Concludo questa breve rassegna con la speranza che le risorse raccolte e rielaborate dal Tavolo in quanto incubatore di interdisciplinarità possano essere messe a frutto nell'attivazione di reti estese di collaborazione interdipartimentale. L'entusiasmo che ha vivacizzato questi mesi di impegno comune ha sempre convissuto con la consapevolezza dell'impossibilità di fornire un quadro esaustivo di tutte le questioni epocali sollevate dall'intelligenza artificiale. Pur con tutti i limiti del caso, ci auguriamo che questo provvisorio lavoro di sintesi possa aiutare a sfrondare retoriche e immaginare percorsi di ricerca innovativi, ben sapendo che il loro concreto attraversamento sarà possibile solo grazie alla collaborazione di tutta la nostra comunità.

Maria Francesca Murru

[4] Il report è accessibile al seguente link: https://op.europa.eu/en/publication-detail/-/publication/c7865738-eb38-11ee-bf53-01aa75ed71a1.

Capitolo I — Questioni definitorie in prospettiva storica

Stefano Coniglio

Nel novembre 2022, il lancio da parte di OpenAI di ChatGPT – il primo strumento a superare i 100 milioni di iscritti in soli due mesi[1] – ha sancito nel giro di poche settimane l'ingresso dell'intelligenza artificiale (IA) tanto nel discorso mediatico e politico quanto nei piani di investimento di aziende pubbliche e private.

È evidente come ChatGPT ed altri modelli di linguaggio di grandi dimensioni (*large language models,* LLM) ad esso tecnologicamente comparabili, come Claude di Anthropic, Llama di Meta e Gemini di Google, si siano distinti per la loro capacità di permettere agli utenti di interagirvi utilizzando, anziché un linguaggio di programmazione, il proprio linguaggio naturale. Secondo alcuni, è proprio grazie alle proprie strabilianti competenze linguistiche che questi strumenti si sono rivelati capaci di affrontare (ad alti livelli di competenza) compiti prima considerati una prerogativa dell'essere umano (Bubeck et al., 2023; OpenAI, 2023).

Per quanto rivoluzionari, gli LLM e l'IA generativa sono solo uno dei molti risultati di rilievo a cui la ricerca in IA ha portato in quasi 70 anni di storia. Questo capitolo si prefigge lo scopo di fornire al lettore un *excursus*, per quanto necessariamente parziale, dell'evoluzione storica di questa disciplina dalla sua fondazione ai giorni nostri.

1. Origini storiche dell'IA

Lo sviluppo dell'IA traccia un arco molto profondo nella storia del pensiero umano, con radici riconducibili fino al pensiero aristotelico (si pensi alla logica "analitica"). Pensatori (pionieristici

[1] https://www.pwc.com/gx/en/industries/tmt/media/outlook/insights-and-perspectives.html

per il loro uso della matematica) quali Gottfried W. Leibniz e George Boole si posero domande fondamentali circa le operazioni cognitive di base del pensiero e le proprietà che un linguaggio (formale o artificiale) debba soddisfare per poter produrre una descrizione del mondo.

La formalizzazione dell'obiettivo cardine dell'IA (individuabile tanto in Leibnitz quanto in Boole, anche se meno esplicitamente nel secondo), vale a dire l'automatizzazione del pensiero, è dovuta ad Alan Turing, che nel 1950 (prima ancora dell'avvento su larga scala del calcolatore elettronico) in un articolo intitolato "*Computing Machinery and Intelligence*" propose il celebre "Test di Turing" per determinare se una macchina possa essere considerata intelligente vagliandone le sole capacità linguistiche (Turing, 1950)[2]. Turing stimolò una profonda riflessione sul pensiero artificiale che gettò le basi per lo sviluppo (da lì a pochi anni) dell'IA in veste di disciplina di indagine scientifica.

L'inizio "ufficiale" dell'attività di ricerca sull'IA viene tradizionalmente identificato nel *workshop* tenutosi nell'estate del 1956 al Dartmouth College negli Stati Uniti, nel New Hampshire. Tra i partecipanti figurano alcuni tra coloro che sarebbero diventati i più influenti pensatori nel campo della nascente disciplina dell'IA. Vi troviamo John McCarthy (che coniò il termine "intelligenza artificiale" per la conferenza e successivamente sviluppò il linguaggio di programmazione LISP, che divenne cruciale per i primi sviluppi dell'AI cosiddetta "simbolica"), Marvin Minsky (che fu co-fondatore dell'AI laboratory dell'MIT e al cui nome si rifà la prestigiosa "Minsky Medal" per conseguimenti scientifici nel campo dell'IA), Allen Newell e Herbert A. Simon (famosi per aver introdotto il concetto di architetture cognitive capaci di manipolare simboli mediante metodi logico-deduttivi e per l'invenzione, di pochi anni prima, del primo sistema di IA della storia – il *Logic Theorist*) e Claude Shannon (universalmente riconosciuto come il padre della teoria

[2] Si noti che le speculazioni di Turing riguardavano un calcolatore digitale "in potenza" attraverso un suo modello concettuale noto oggi come *macchina di Turing*, e non un calcolatore specifico, per altro quasi inesistente nel 1950.

dell'informazione e a cui si deve, tra le altre cose, l'introduzione del concetto di *bit*).

A Dartmouth la disciplina dell'IA venne fondata con l'ambizioso obiettivo di inventare meccanismi che permettessero alle macchine di utilizzare il linguaggio, astrarre concetti, risolvere problemi ed apprendere, con la speranza di promuovere l'idea che le macchine potessero allora (e possano ora) non solo eseguire calcoli ma persino "pensare"[3].

Dei due filoni di ricerca principali che nacquero dopo Dartmouth, quello simbolico (tipico dei sistemi esperti e basato sulla manipolazione logico-matematica di regole, in linea coi principi operativi del *Logic Theorist*) e quello sub-simbolico o connessionista (basato su modelli matematico/computazionali del cervello biologico, le cosiddette reti neurali di cui parleremo più estesamente), fu il primo a dominare la ricerca in IA fino a circa gli anni '90. Nel corso dei decenni successivi al 1956, la ricerca nell'ambito dell'intelligenza artificiale attraversò fasi di alterno entusiasmo (*AI Summer*) e delusione (*AI Winter*) e molta parte della ricerca sulle reti neurali venne abbandonata[4].

La più importante fase di rinnovato entusiasmo per l'IA, completamente imputabile questa volta a metodi connessionisti basati su reti neurali, si verificò nel 2012 con la vittoria della ImageNet Large Scale Visual Recognition Challenge (ILSVRC) di AlexNet, un sistema di riconoscimento di immagini sviluppato da Alex Krizhevsky, Ilya Sutskever e Geoffrey Hinton (Krizhevsky et al., 2012).

---

[3] Evento fondamentale antecedente a Dartmouth fu la progettazione nel 1955 del già citato *Logic Theorist* da parte di Allen Newell, Herbert A. Simon e J. Clifford Shaw, un sistema per la dimostrazione automatica di teoremi da molti considerato il primo sistema di IA della storia. Si mostrò, in particolare capace di dimostrare 40 tra i teoremi inclusi nei Principia Mathematica di Alfred N. Whitehead e Bertrand Russell.

[4] Gli *AI Winter* furono spesso causati dai pochi progressi ottenuti nella ricerca in IA a valle di roboanti annunci circa risultati strabilianti mai realmente conseguiti. Il primo *AI Winter* avvenne negli anni '70, quando un rapporto governo britannico, noto come "Lighthill Report", concluse che l'IA aveva fallito nel mantenere le sue promesse, provocando una drastica diminuzione dei fondi destinati alla ricerca. Un secondo *AI Winter* si verificò verso la fine degli anni '80 in concomitanza con la contrazione del mercato delle macchine basate sul linguaggio LISP e di una ridotta fiducia nei sistemi esperti a valle di forti investimenti nel settore.

Il successo di AlexNet fu reso possibile grazie alla combinazione di tre elementi essenziali: la maturazione scientifica dei metodi di addestramento delle reti neurali così come l'adozione, seppur non del tutto nuova, di architetture molto strutturate (il cosiddetto apprendimento profondo o *deep learning*), l'uso di un *hardware* potente per l'addestramento e la disponibilità, resa possibile dalla diffusione di massa di internet, di una gigantesca mole di dati per l'addestramento (*training*).

2. Avvento delle reti neurali

Come si diceva, le reti neurali (a volte dette reti neurali artificiali) sono un modello matematico molto semplificato del funzionamento della rete neurale biologica presente nel cervello umano. Nel loro fondamento teorico si riconducono ai lavori di Alexander Bain (1873) e William James (1890)[5].

A seguito di alcuni primi utilizzi attorno agli anni '30 di reti neurali artificiali per la modellizzazione e lo studio delle reti neurali biologiche, fu solo con l'invenzione del *perceptron* di Warren McCulloch e Walter Pitts (1943) e con la sua implementazione (omonima) *in hardware* da parte di Frank Rosenblatt (1958) che la comunità scientifica comprese che questi modelli potessero essere usati come strumenti pratici di *problem solving* in contesti che poco o nulla avessero a che spartire con la simulazione dei processi biologici.

Dopo una prima fase di iniziale entusiasmo, il *perceptron* si dimostrò limitato nella sua capacità di risolvere problemi semplici (tra i quali il più famoso: la realizzazione di uno XOR), come evidenziarono Marvin Minsky e Seymour Papert in un famoso libro del 1969: "*Perceptrons*" (Minsky & Papert, 1969)[6]. Nonostante le

[5] Sia Bain che James ipotizzarono che il pensiero umano emergesse dalle interazioni tra un gran numero di neuroni all'interno del cervello. Nel 1949, Donald Hebb descrisse l'apprendimento cosiddetto hebbiano, l'idea cioè che una rete neurale biologica possa apprendere nel tempo rafforzando le proprie connessioni neuronali (sinapsi) ogni volta che un segnale le attraversa.

[6] È interessante notare che, nonostante Minsky venga spesso annoverato tra i principali detrattori delle reti neurali, Minsky stesso diede determinanti contributi allo sviluppo iniziale di questi modelli – nel 1951 costruì, con Dean Edmonds, una pionieristica

osservazioni molto critiche che il libro offrì sul *perceptron* valessero solo per un caso particolarmente semplice della rete neurale di Rosenblatt, la sua pubblicazione portò ad un significativo declino dell'interesse per le reti neurali e, con esso, a una forte riduzione dei finanziamenti in questo campo[7].

L'interesse per le reti neurali rifiorì negli anni '80 grazie alla (ri)scoperta dell'algoritmo di *backpropagation*, un metodo per l'aggiornamento dei pesi sinaptici delle reti neurali originariamente proposto da Paul Werbos nel 1974 e molto legato a tecniche simili adottate nell'area del controllo automatico. Reso popolare negli anni '80 da un fortunato articolo scritto da David Rumelhart, Geoffrey Hinton e Ronald Williams (Rumelhart et al., 1986), questo algoritmo permise l'aggiornamento efficace dei pesi di una rete neurale basandosi su semplici regole matematiche di derivazione, formalizzando in senso matematico il problema del loro addestramento che, in precedenza, veniva tipicamente realizzato con metodi euristici (vale a dire poco strutturati e situazionali, e dunque spesso difficilmente generalizzabili o replicabili).

Fondamentali per il rinnovato interesse scientifico per le reti neurali furono anche i lavori di Yann LeCun (LeCun et al.,1990) che, negli anni '90, contribuì significativamente allo sviluppo delle cosiddette reti neurali convoluzionali o CNN (il modello matematico alla base di AlexNet), dimostrandone l'efficacia nel riconoscimento di immagini – il *dataset* MNIST, curato da LeCun,

---

rete neurale realizzata completamente *in hardware*. La macchina di Minsky ed Edmonds, nota come *Stochastic Neural Analog Reinforcement Calculator* (SNARC), fu progettata per simulare un topo che navigasse in un labirinto. Fu addestrata con quel che al tempo veniva chiamato *apprendimento per rinforzo skinneriano* (ora un pilastro del *machine learning*), rifacendosi ai lavori dello psicologo B. F. Skinner.

[7] Nei successivi anni '80 e '90, lo sviluppo dell'IA proseguì principalmente grazie ai successi dei cosiddetti sistemi esperti, strumenti progettati per simulare il comportamento di un esperto umano in un dominio applicativo specifico. Basati su grandi collezioni di regole, questi sistemi sfruttavano algoritmi di deduzione logico-matematica per inferire proprietà valide nel sistema di interesse che non fossero presenti negli assiomi iniziali, mostrandosi capaci di prendere decisioni e di risolvere problemi complessi fornendo garanzie di correttezza. Differentemente dalle reti neurali, che richiedono forti potenze di calcolo e molti dati per il loro addestramento, i sistemi esperti venivano programmati per la maggior parte manualmente.

è, con ImageNet, uno dei *dataset* più utilizzati per il vaglio di una rete neurale[8].

Il modello adottato da LeCun si rifaceva al *Neocognitron* inventato da Kunihiko Fukushima (1980), il quale, a sua volta, si rifaceva agli studi David Hubel e Torsten Wiesel (1959) sulla corteccia visiva. Hubel e Wiesel, insigniti del premio Nobel nel 1981, osservarono come i neuroni nella corteccia visiva rispondano a stimoli visivi quali bordi ed angoli e come il cervello processi le informazioni visive gerarchicamente, con neuroni specializzati per l'identificazione di *pattern* via via più complessi quando ci si allontana dalla retina partendo dai più semplici (riconducibili, con una certa approssimazione, ai *pixel* delle immagini digitali).

Come si diceva, fu grazie all'avvento di grandi potenze di calcolo e della disponibilità di grandi moli di dati che idee architetturali ed algoritmiche di fatto preesistenti resero le potenzialità delle reti neurali completamente evidenti. Il dataset su cui fu addestrata AlexNet conteneva, infatti, molti milioni di immagini etichettate ed è noto che tutti gli LLM di successo siano stati addestrati sulla quasi totalità dei dati presenti in *internet* (Brown et al., 2020). È noto anche come i risultati (portentosi per il 2012) di AlexNet convinsero la comunità scientifica della grande quantità di potenza di calcolo etraibile dalle *Graphics Processing Unit* (GPU) che erano prima utilizzate quasi esclusivamente in applicazioni di *computer graphics* e *videogame* – è celebre la frase di Geoffrey Hinton «*In 2009, I remember giving a talk at NIPS* [ora NeurIPS] *where I told about 1,000 researchers they should all buy GPUs because GPUs are going to be the future of machine learning*»[9], soprattutto se si pensa che dopo soli 12 anni (nel giugno 2024) Nvidia è diventata la società con la maggiore

[8] Il dataset è reperibile all'indirizzo *http://yann.lecun.com/exdb/mnist*.

[9] https://venturebeat.com/ai/how-nvidia-dominated-ai-and-plans-to-keep-it-that-way-as-generative-ai-explodes.

capitalizzazione di mercato al mondo,per un valore di oltre 3,3 trilioni di dollari[10].

3. Anatomia di una rete neurale

Nella sua versione più semplice, una rete neurale di tipo *feed forward* è costituita da un insieme di neuroni (artificiali) e di connessioni (sinapsi) tra di essi. La rete simula il comportamento della cellula neurale biologica. Sollecitata da segnali elettrici prodotti da altri neuroni e ricevuti attraverso le connessioni sinaptiche, la cellula si attiva, producendo a sua volta un segnale elettrico che, raggiunto un secondo insieme di neuroni anch'essi connessi mediante sinapsi, ne provoca l'attivazione se la sollecitazione ricevuta è sufficientemente forte.

Un neurone artificiale segue il comportamento (seppur semplificato) del suo *alterego* biologico con una principale differenza: i segnali che riceve e produce non sono segnali elettrici ma numeri; una volta ricevuti questi numeri dai neuroni "a monte", il neurone artificiale ne calcola la somma, ottenendo un risultato numerico (la cosiddetta pre-attivazione); a questo applica poi una *funzione di attivazione* (ottenendo la cosiddetta post-attivazione); se il segnale somma è sufficientemente forte per la funzione di attivazione, il neurone artificiale si attiva e produce un segnale numerico d'uscita; in caso contrario, rimane inattivo. La funzione di attivazione più utilizzata è la *Rectified Linear Unit* (ReLu)[11] – una funzione molto semplice diventata popolare solo di recente ma che fu introdotta proprio da Fukushima (1969).

Un comportamento come quello descritto sopra, seppur in apparenza abbastanza semplice, è sufficientemente potente da permettere l'approssimazione di qualsiasi funzione matematica che traduca uno o più numeri in ingresso (un vettore) in uno o più

[10] https://www.reuters.com/markets/us/nvidia-becomes-worlds-most-valuable-company-2024-06-18.

[11] È una funzione relativamente semplice che riproduce in uscita il suo ingresso se questo è positivo, producendo zero altrimenti.

segnali in uscita (un altro vettore) con una rete relativamente semplice[12].

In una rete neurale di tipo *feed forward*, i neuroni sono disposti per strati. Le componenti del vettore in ingresso vengono utilizzate ordinatamente come segnali di ingresso per i neuroni dello strato iniziale. Le uscite calcolate dai neuroni del primo strato attraversano le sinapsi per raggiungere gli ingressi dei neuroni del secondo strato, che possono o meno attivarsi e inviare segnali ai neuroni del terzo strato, e così via fino allo strato di uscita.

Ogni sinapsi ha un *peso*. Una volta determinata l'architettura della rete (decisi cioè il numero di strati e di neuroni in ogni strato e l'architettura delle loro interconnessioni), la fase di addestramento (*training*) o apprendimento (*learning*) della rete coincide con la determinazione di un valore numerico per i pesi che meglio permetta alla rete di produrre in uscita il risultato desiderato.

Il principio di fondo dietro all'addestramento è la riduzione dell'errore tra l'output *prodotto* in funzione di specifici valori dell'input e quello *previsto* – è questo il principio del cosiddetto apprendimento supervisionato (*supervised learning)*[13]. Per realizzarlo è cruciale avere a disposizione un insieme di dati di addestramento (il cosiddetto *training set*) che ben rappresenti la relazione tra ingresso e uscita che si vuole che la rete catturi. In generale, più grande è meglio è – alla luce di questo, non è difficile capire perché l'avvento di internet sia stato fondamentale per il successo delle reti neurali.

Come anticipato poco fa, l'addestramento di una rete neurale avviene tipicamente tramite l'algoritmo di *backpropagation*, un metodo che consente di aggiustare (in termini più matematici diremmo *ottimizzare*) i pesi della rete per ridurne l'errore.

[12] Esistono molti teoremi di approssimazione universale per reti neurali validi sotto diverse assunzioni, nessuna delle quali troppo restrittiva. Va però detto che molti di questi risultati possono richiedere una rete o un *dataset* di *training* di dimensioni astronomiche.

[13] L'*unsupervised learning*, approccio alternativo al primo, richiede invece la sola capacità di valutare la bontà dell'output prodotto dalla rete senza la necessità di confrontarlo ad un output ideale noto a priori.

L'algoritmo di *backpropagation* ha reso possibile addestrare reti neurali con molti strati, aprendo la strada al cosiddetto *deep learning* – la disciplina dell'addestramento automatico (*machine learning*) con reti neurali *profonde* (*deep*)[14].

Oltre alla profondità, cruciale nel *deep learning* è l'impiego di reti neurali contenenti diversi blocchi (essi stessi reti neurali) preposti a scopi differenti. Tipicamente, questi blocchi permettono di apprendere rappresentazioni gerarchiche dei dati, rifacendosi alle idee, pionieristiche per il tempo, del *Neocognitron* di Fukushima.

Grazie alla loro capacità di catturare caratteristiche complesse e astratte, queste reti sono state applicate con successo in numerosi ambiti, tra cui il riconoscimento delle immagini (come già si è detto parlando diffusamente dei lavori di Yann LeCun e di AlexNet), il riconoscimento vocale e la traduzione automatica, che sono solo alcune tra le molte applicazioni di successo.

Come accennavamo, la combinazione di grandi *dataset*, potenti capacità computazionali fornite dalle GPU e algoritmi avanzati hanno portato a una rapida evoluzione delle prestazioni delle reti neurali e alla loro consacrazione come tecnologia cardine (in un gergo in qualche modo scientifico diremmo *stato dell'arte*) in molte applicazioni di natura tecnica.

4. Intelligenza artificiale generativa

Va notato come il lancio, nel novembre 2022, di ChatGPT abbia portato negli occhi del grande pubblico un insieme di modelli di IA, detti di IA generativa (*gen AI*) che, seppur basati essi stessi su reti neurali, hanno caratteristiche e peculiarità in apparenza nuove. Primo tra queste figura senz'altro la capacità di generare testo di qualità sintattica e contenuto spesso comparabili a produzioni umane a partire da un testo fornito in input (*prompt*).

[14] Va detto che l'addestramento di reti profonde richiede diversi accorgimenti algoritmici che le reti *feed forward* non richiedono. Il principio di fondo è però il medesimo.

Da un punto di vista tanto scientifico quanto tecnologico, l'avvento dell'IA generativa ha segnato una forte inversione di tendenza rispetto a tecnologie di IA "tradizionali" (il *machine learning*, le reti neurali non generative, l'ottimizzazione matematica, il *forecasting* e via discorrendo) per le quali valeva il cosiddetto paradosso di H.P. Moravec che, nella rilettura di Pinker (1994), recitava che ciò che è difficile per l'uomo è facile per la macchina mentre ciò che lo è per l'uomo alla macchina risulta difficile.

Come precedentemente accennato, la ricerca in IA è tradizionalmente categorizzata in due principali approcci: il "simbolico" e il "sotto-simbolico" o "connessionista". L'approccio simbolico si concentra sulla manipolazione di simboli attraverso metodi matematici e tecniche logico-deduttive, tipici dei sistemi esperti degli anni '80. In questo contesto, i modelli matematici forniscono risposte che possono essere interpretate dagli esseri umani.

Dall'altro lato, l'approccio sotto-simbolico rappresenta i dati in forma puramente numerica (a volte detta vettoriale o tensoriale) e li elabora attraverso reti neurali — l'IA generativa, come tutte le tecniche basate su reti neurali, profonde o meno, rientra in questo secondo filone. A differenza degli approcci simbolici, le tecniche sotto-simboliche non offrono facilmente risposte interpretabili dall'uomo, rendendo la comprensione del proprio comportamento e il modo in cui siano arrivate a talune conclusioni estremamente difficili.

La generatività di un LLM nasce grazie a uno schema di addestramento a più fasi.

La prima fase, detta a volte di *pre-training*, viene realizzata senza un compito applicativo specifico ma, anzi, in una modalità quasi unicamente predittiva: imparare a riprodurre piccole porzioni (oscurate in fase di addestramento) di testo tipicamente raccolto dalle fonti più disparate: libri, articoli, siti web e altre risorse testuali disponibili. L'obiettivo è far sì che il modello apprenda le strutture linguistiche, il vocabolario e le relazioni che intercorrono tra le unità atomiche del linguaggio naturale e che

maturi una forma di comprensione del contesto. È in questa fase che l'LLM diventa un *foundation model*, un modello (basato su reti neurali e addestrato su grandi quantità di dati multimodali) capace (quantomeno in potenza in questo stadio dell'addestramento) di svolgere una vasta gamma di compiti di specificità variabile pur senza essere stato concepito né direttamente addestrato a questo scopo.

Una volta completata questa prima fase, il modello entra in una fase di *fine-tuning* in cui viene ulteriormente addestrato su *dataset* specifici che includono conversazioni con domande e risposte e produzioni di testo mirate. In questa seconda fase vengono adottate tecniche di cosiddetto *Reinforcement Learning with Human Feedback* (RLHF) in cui uno o più operatori umani interagiscono col modello valutandone la qualità delle risposte fornite e producendo un *feedback* che, utilizzato in questa successiva fase di addestramento, permetta il miglioramento delle prestazioni[15]. Questo processo aiuta a rendere le risposte del modello più rilevanti, coerenti e utili per l'utente finale (Ouyang et al., 2022).

Due sono le innovazioni tecnologico-scientifiche chiave che hanno reso possibile la rapidissima esplosione dell'IA generativa. La prima (di cui parleremo) è il *transformer* – un'architettura proposta in un articolo scritto da ricercatori di Google Brain e Google Research, che costituisce l'elemento fondante di tutti gli LLM (Vaswani et al., 2017)[16]. La seconda innovazione (di cui non parleremo – è matematicamente ben più complessa della prima) sono i modelli di *diffusion*, inizialmente sviluppati in contesti puramente accademici (Sohl-Dickstein et al., 2015) e divenuti oggi una tecnologia cruciale in strumenti di generazione di

[15] Tipicamente viene addestrato un modello "di secondo livello" che cattura la relazione produzione-feedback implicitamente realizzata dall'intervento umano. Tale relazione verrà poi utilizzata in un secondo momento come *loss function* (la funzione guida nell-addestramento) per il *fine tuning*.

[16] Nel momento in cui scriviamo, l'articolo ha accumulato quasi 125.000 citazioni, rendendo i suoi autori delle "microcelebrità". In questo senso, si veda anche l'articolo https://www.wired.com/story/eight-google-employees-invented-modern-ai-transformers-paper.

immagini e musica (quali, ad esempio, DALL-E, Midjourney, Stable Diffusion, Suno AI e Stable Audio).

5. Anatomia di un Large Language Model

Tutti i recenti LLM sono basati su un'architettura detta GPT (*Generative Pretrained Transformer*), un modello di *deep learning* specializzato nella comprensione e generazione del linguaggio naturale. I *transformer* permettono di elaborare sequenze di testo efficientemente ed in parallelo, a differenza di approcci precedenti in grado di elaborazioni solo sequenziali e, pertanto, con una efficienza molto minore.

Analizziamo ora, seppur in una forma necessariamente approssimata vista la natura di questo capitolo, il funzionamento di un LLM basato su *transformer*.

Embedding e tokenizzazione

Il testo fornito in input dall'utente viene suddiviso in pezzetti (*token*), corrispondenti a, tipicamente, porzioni di parole e simboli specifici (ad esempio, i segni di punteggiatura). I *token* sono l'unità fondamentale con cui il modello opera. Ogni *token* viene convertito in un vettore numerico (una sequenza di numeri di lunghezza fissata) attraverso un'operazione detta di *embedding*. Si ottiene così un alter ego numerico del *token* (detto anch'esso *embedding*) che ne cattura informazioni semantiche e sintattiche mediante la sua vicinanza (in senso geometrico) agli *embedding* di altri token[17]. Questo permette al modello di comprendere e lavorare con il testo mediante strumenti matematici e non linguistici.

Si può pensare agli *embedding* come a dei punti in una mappa. I *token* vengono tradotti in punti della mappa, con la particolarità che la rete neurale soggiacente all'LLM (a volte parte di questo o, più spesso, pre-addestrata indipendentemente) tenderà a

[17] Dobbiamo pensare che in uno spazio vettoriale ad alta dimensione è molto più facile costruire rappresentazioni vettoriali che siano più o meno vicine le une alle altre, più di quanto non capiti nello spazio tridimensionale a cui siamo abituati.

mappare in punti vicini della mappa quei *token* che condividono relazioni sintattiche o semantiche di vicinanza[18].

Consideriamo, a titolo di esempio, le seguenti coppie token–embedding: "re"–[0,4; 0,7; -0,5], "uomo"–[0,2; 0,5; –0,3], "donna"–[0,1; 0.6; –0,4] e "regina"–[0,3; 0,8; –0,6]. Possiamo usare operazioni di somma e sottrazione sugli *embedding* per inferire nuove relazioni semantiche, ad esempio la relazione "re – uomo + donna ≈ regina", dove (lo si può verificare per esercizio) vale la relazione numerica [0,4; 0,7; -0,5] – [0,2; 0,5; –0,3] + [0,1; 0,6; –0,4] = [0,3; 0,8; –0,6].

Meccanismo di Attenzione

Nel cuore di un LLM troviamo un meccanismo cosiddetto di attenzione (*self-attention*). Questo meccanismo permette al modello di calcolare un punteggio di attenzione (*attention score*) per ogni coppia di *token* presenti nel testo di input (*prompt*). Questi punteggi catturano quanto l'informazione contenuta in un *token* sia rilevante nei confronti di un altro *token*[19]. I punteggi calcolati hanno la cruciale funzione di pesi in una somma pesata dell'*embedding* di ogni token con quelli di tutti gli altri token presenti nel testo[20]. Ad esempio, a valle della somma pesata l'*embedding* della parola "Campione" si avvicinerà all'area semantica (nello spazio di *embedding* – la mappa dell'esempio precedente) delle attività agonistiche se vicino ad essa compaiono termini come "competizione" o "vittoria", virando invece verso un'area semanticamente più vicina alla geografia se vicino ad essa si trovano le parole "D'Italia". In questo modo, gli *embedding* originali (avulsi dal contesto) vengono arricchiti con

[18] All'*embedding* del *token* viene spesso addizionato un embedding posizionale che incorpora informazioni sulla posizione occupata dal *token* nella sequenza del testo. Alcune varianti dell'architettura adottano anche embedding segmentali che permettono di distinguere parti del testo quali, ad esempio, domande e risposte.

[19] La *self-attention* viene tipicamente implementata in più "testine" parallele (*multi-head attention*), ognuna delle quali addestrata per catturare diversi tipi di relazioni tra i *token*.

[20] Questo meccanismo non è troppo diverso dall'operatore di convoluzione usato nelle reti convoluzionali.

informazione contestuale. In linea con la pratica comune in quasi tutti gli approcci di *deep learning*, un LLM contiene una sequenza di strati di *transformer* che, via via, catturano informazioni su livelli di astrazione sempre maggiori.

Key/Query/Value

Il meccanismo di *self-attention* nei *transformer* si basa sui tre elementi detti *key*, *query* e *value*. Ogni *token* del testo viene trasformato in tre vettori numerici distinti: un vettore *key*, un vettore *query* e un vettore *value*. Il modello calcola poi la similarità tra la *query* di un token e la *key* di tutti gli altri token, calcolando così un insieme di *attention score* (che, come si è detto, quantificano l'importanza relativa di un *token* rispetto a tutti gli altri).

La ragione del successo dei transformer risiede nel fatto che questo calcolo può, per sua natura, essere realizzato in parallelo parola per parola, sfruttando efficientemente la potenza di calcolo dei sistemi distribuiti (del *cloud)*. Questa capacità di elaborare informazioni in parallelo e di catturare relazioni complesse su scale diverse rende il *transformer* estremamente efficiente e scalabile rispetto ai modelli precedenti che elaborano il testo in modo puramente sequenziale. Questo consente di gestire efficientemente sequenze di testo anche molto lunghe (Gemini 1.5 di Google può elaborare testo contenente fino a 1 milione di *token*) e di accelerare significativamente il processo sia di addestramento che di inferenza.

Generazione del Testo e Decodifica

Ad ogni strato della rete, ogni *embedding* viene combinato (in una somma pesata) con tutti gli *embedding* delle altre parole presenti nel testo, pesato dall'*attention score* della coppia di *token*. Viene così prodotto un nuovo *embedding* arricchito dal contesto.

Per generare il testo finale, il modello aggrega l'output dell'ultimo strato di *transformer* (che contiene un *embedding* per ogni parola presente tanto nel prompt quanto nello scambio di

testo avvenuto in precedenza tra l'utente e l'LLM) creando un *embedding finale* che viene quasi direttamente utilizzato (a valle di alcune semplici trasformazioni matematiche) per la predizione del *token* successivo nella sequenza. In questo modo, l'LLM calcola una distribuzione di probabilità su tutto il vocabolario che ha a disposizione. La scelta della prossima parola da produrre (*next-word prediction*) in output viene fatta campionando da questa distribuzione il token il cui *embedding* è più vicino all'*embedding finale*.

Ad esempio, se il modello sta generando una frase dove la parola corrente è "del" e le precedenti sono "il", "gatto" e "mangia", il modello utilizzerà il contesto per determinare che "pesce" potrebbe essere la parola successiva corretta.

Un parametro di *temperatura* determina quanto rigidamente l'LLM debba attenersi alla probabilità calcolata. Una temperatura alta rende il modello più creativo, mentre una bassa lo rende più preciso.

Il token appena generato viene aggiunto alla sequenza di token precedenti (la quale include sia il *prompt* che il testo generato fino ad ora) ed il processo viene reiterato finché l'LLM non arrivi a produrre un *token* speciale detto di fine sequenza, nel qual caso la produzione si arresta.

Allucinazioni

I fenomeni cosiddetti di *allucinazione* (un termine migliore sarebbe forse *confabulazione*) di cui gli LLM sono noti soffrire sono, da un punto di vista scientifico, endemici e strutturalmente inevitabili. Come si è spiegato in precedenza, la generatività di questi modelli è, per sua natura, interamente priva di meccanismi logico-matematici di verifica della correttezza della produzione (questo vale anche per la correttezza sintattica che, seppur infrequentemente, può venire a mancare, specialmente quando la lingua adottata non è l'inglese). Anche se per alcuni può sembrare controintuitivo, questi sistemi non possono verificare a posteriori la correttezza della propria produzione partendo dai dati di addestramento, ai quali di fatto non hanno alcun accesso

diretto — di questi contengono solo una versione rielaborata e codificata (in forma molto compressa) nei pesi sinaptici.

Diverse tecniche di interrogazione e *prompting* "avanzato" sono state proposte per mitigare il problema. La più popolare, detta *retrival-augmented generation* (o RAG), permette di incrementare (da un punto di vista puramente empirico) la capacità generativa di un LLM in un *task* d'interesse aggiungendo ai limitati esempi pertinenti a questo *task* a cui l'LLM è stato sottoposto in fase di addestramento (*few-* o *zero-shot learning*) altri (nuovi) esempi rilevanti forniti dall'utente. Va detto che, seppur tecniche di questo tipo possano ridurre nella pratica la probabilità di allucinazione e portare a produzioni di maggior qualità, non forniscono, per loro natura, garanzia *matematiche* di correttezza di quanto l'IA generativa produce – garanzia che sistemi simbolici, sebbene di potenza e versatilità estremamente più limitate, potrebbero fornire.

6. La lezione amara di R. Sutton

Non è difficile individuare una forma di fascinazione nei lavori dei pionieri dell'IA per l'idea che lo studio e la progettazione dell'intelligenza artificiale potessero portare alla comprensione, seppur indirettamente, dei meccanismi di funzionamento del pensiero umano. A titolo d'esempio, è noto come G. Hinton (uno dei padri dell'IA moderna e di cui abbiamo già parlato) si sia rivolto alla ricerca sull'intelligenza artificiale motivato dal desiderio di apprendere i meccanismi della mente e, in qualche forma, frustrato dall'aver osservato che né la psicologia né la filosofia fossero in grado di dare risposte esaustive al suo quesito[21]. Questa fascinazione, che era senz'altro presente nei lavori più classici di IA simbolica, è fortemente venuta meno, a detta di Sutton (2019), nell'era del trionfo delle reti neurali.

La lezione (potremmo dire "constatazione") amara di Sutton è che i metodi dell'IA basati sulla ricerca e sull'apprendimento automatico (dove la potenza di calcolo ha un fortissimo rilievo, a

[21] https://www.youtube.com/watch?v=n4IQOBka8bc.

discapito degli aspetti più scientifici di modellizzazione) si siano rivelati largamente più efficaci dei sistemi progettati per incorporare elementi della conoscenza umana (si pensi ai sistemi esperti). Sutton constata come i modelli e i metodi progettati per emulare il funzionamento (d'alto livello) della mente umana non funzionino su larga scala e che, storicamente, siano spesso stati soppiantati da sistemi basati su tecniche di ricerca e apprendimento capaci di scalare con l'aumento della capacità di calcolo.

Anziché, dice Sutton, integrare nei sistemi di IA il funzionamento della mente (di complessità, dice, arbitrariamente grande) l'IA moderna ha imparato ad integrare metodi capaci essi stessi di catturare questa grande complessità, anche se, aggiungiamo noi, questi sistemi esibiscano spesso un comportamento che all'uomo può risultare difficile decifrare.

Bibliografia

Bain, A. (1873). *Mind and Body: The Theories of Their Relation*. New York: D. Appleton and Company.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. & Agarwal, S., 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M.T. and Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. [online] arXiv.org. https://arxiv.org/abs/2303.12712

Fukushima, K. (1969). Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics*. 5 (4): 322–333.

Fukushima, K. (1980). Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. *Biological Cybernetics*. 36 (4): 193–202.

Hubel, D. H. & Wiesel, T. N. (1959). Receptive Fields of Single Neurones in the Cat's Striate Cortex. *J. Physiol*. 148 (3): 574–91.

James, W. (1890). *The Principles of Psychology*. New York: H. Holt and Company.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25.

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E. & Hubbard, W.E. (1990). *Advances in Neural Information Processing Systems*, 3.

McCulloch, W. & Pitts, W. (1943). A Logical Calculus of Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*. 5 (4): 115–133.

Minsky, M., & Papert, S. (1969). *Perceptrons: An introduction to computational geometry*. Cambridge Press., HIT, 479(480), 104.

OpenAI (2023). GPT-4 Technical Report. [online]: https://cdn.openai.com/papers/gpt-4.pdf

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. & Schulman, J. (2022). Training language models to follow instructions with human

feedback. *Advances in Neural Information Processing Systems*, 35.

Pinker, S. (1994). *The Language Instinct*, Perennial Modern Classics, Harper.

Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model For Information Storage And Organization In The Brain. *Psychological Review*. 65 (6): 386–408.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), pp.533-536.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. *International Conference on Machine Learning*, 37.

Sutton, R. (2019). The bitter lesson. Incomplete Ideas. [online] http://www.incompleteideas.net/IncIdeas/BitterLesson.html

Turing, A. M. (1950). I.—Computing machinery and intelligence. *Mind*, LIX (236), 433–460.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Capitolo II — Opacità e complessità

Federico Leo Redi
Francesca Cerea

1. Aspetti tecno-scientifici dell'opacità

Questa prima sezione del documento esamina le complessità e la natura opaca dell'intelligenza artificiale (IA) in ambiti professionali e scientifici. Il documento inizia con l'affrontare l'opacità strutturale e la complessità dell'IA, esplorando le sue implicazioni etiche e tecniche, per poi approfondire gli strati sfumati di opacità che influenzano il processo decisionale critico in aree sensibili e alcune delle relative implicazioni.

Differenti livelli di opacità

Un primo livello di opacità è legato agli aspetti algoritmici e strutturali dell'IA che portano alla difficile interpretabilità di questa. I modelli di apprendimento automatico, e in particolare quelli basati sull'apprendimento profondo, operano attraverso reti neurali estremamente complesse che vengono addestrate su enormi quantità di dati. Tanto i processi di addestramento quanto quelli di inferenza (l'uso cioè di una rete neurale già addestrata per produrre un risultato) avvengono spesso in maniera non trasparente per l'utente finale in conseguenza della complessità strutturale della rete adoperata, creando una barriera spesso invalicabile alla comprensione dei risultati e delle decisioni prese dalla macchina.

La natura di "scatola nera" di questi sistemi di IA solleva questioni etiche significative, soprattutto quando le decisioni dell'IA influenzano aree critiche come la medicina, il diritto, o il reclutamento lavorativo. Senza una chiara comprensione dei processi soggiacenti, diventa difficile valutare l'equità, l'accuratezza e, in ultima analisi, fidarsi dei sistemi di IA. Questa mancanza di trasparenza può anche ostacolare gli sforzi per

identificare e correggere i bias presenti nei dati di *training* o conseguenti all'architettura adottata, perpetrando inevitabili disuguaglianze e pregiudizi.

La complessità algoritmica dell'IA non si limita solo ai modelli di inferenza e alle loro architetture interne, ma si estende anche alle infrastrutture tecniche su cui questi strumenti vengono eseguiti, alle interazioni tra le diverse componenti software e hardware, e alle dinamiche socio-tecniche che ne emergono. Questa complessità rende ancora più arduo il compito di garantire la costruzione e distribuzione di sistemi di IA sicuri, affidabili e conformi ai principi etici.

Di fronte a queste sfide, sono emerse diverse proposte volte a incrementare la trasparenza e l'interpretabilità dell'IA. Metodi di IA spiegabile (*explainable AI*, XAI) cercano di rendere i processi decisionali dei modelli di apprendimento profondo più accessibili e comprensibili sia per gli esperti che per gli utenti non specializzati. Questo si traduce nello sviluppo di strumenti e tecniche che permettono di studiare sistematicamente le ragioni dietro le previsioni o le decisioni prese da un sistema di IA. Questa trasparenza è fondamentale per i campi in cui la comprensione del processo decisionale è importante quanto il risultato stesso, come nella diagnostica sanitaria o nei servizi finanziari. Un esempio concreto di XAI in azione è il *credit scoring*, dove l'*explainability* aiuta a spiegare i fattori che influenzano le decisioni di credito di una macchina a chi richiede un prestito, rendendo così i sistemi di IA più trasparenti ed equi. Analogamente, nel settore sanitario, le tecniche XAI vengono utilizzate per chiarire i percorsi diagnostici seguiti dall'IA nelle tecnologie di imaging, aiutando i medici a capire perché certe anomalie siano state segnalate, aumentando così la fiducia nelle diagnosi assistite dall'IA. La XAI è inoltre sempre più richiesta negli ambienti normativi, spingendo le industrie ad adottare sistemi di IA più trasparenti.

La fiducia nei sistemi di IA passa attraverso la costruzione di una maggiore trasparenza e la capacità degli utenti di comprendere e valutare criticamente il funzionamento interno e

le decisioni di questi sistemi. Affrontare l'opacità e la complessità strutturale dell'IA richiede un approccio multidisciplinare che coinvolga sviluppatori, ricercatori, esperti legali ed etici, oltre agli utenti finali, per garantire che l'evoluzione dell'IA proceda in maniera etica, responsabile e al servizio dell'umanità.

Analizzare i livelli più profondi di opacità dell'IA richiede un approccio che vada oltre la semplice interpretazione dei processi decisionali e che affronti le intricate dinamiche che regolano le interazioni tra i diversi componenti dell'IA e le più ampie implicazioni sociali e individuali di queste tecnologie.

Il secondo livello di opacità viene identificato nella complessità delle interazioni tra i diversi componenti di un sistema di IA. Ad esempio, nel settore sanitario, il modo in cui i dati dei pazienti vengono raccolti, filtrati e poi utilizzati per addestrare gli algoritmi diagnostici può introdurre pregiudizi, come quelli legati al genere o alla razza, che influenzano i consigli medici o le opzioni di trattamento. Ciò può verificarsi a causa di discrepanze nei dati storici o dell'inclusione involontaria di pregiudizi da parte degli sviluppatori. La comprensione completa di questo livello richiede non solo la trasparenza degli algoritmi, ma anche chiarezza sulle pipeline di dati e sulle scelte di progettazione, spesso opache, che rimangono nascoste persino agli sviluppatori. Il comportamento cosiddetto *emergente* è tipico di tutti i sistemi di IA generativa quali, ad esempio, gli LLM, che esibiscono competenze in aree applicative che non sono state oggetto di addestramento (*zero-shot learning*) o solo di un regime di addestramento limitato (*few-shot learning*).

Il terzo livello di opacità si spinge oltre, comprendendo le implicazioni sociali, etiche e legali dell'impiego dell'IA. La sua prima causa è di tipo istituzionale, dipende cioè dai dispositivi legali che regolano la proprietà intellettuale e che garantiscono il vantaggio competitivo nello sviluppo e nel commercio di tali tecnologie. La seconda è di tipo culturale e ha a che fare con la diseguale distribuzione di competenze specialistiche tra i fruitori di IA che sono necessarie alla comprensione del loro funzionamento.

Rischio di mancanza di competenze specialistiche

L'espansione degli sviluppatori di tecnologie di AI senza competenze specialistiche e, in taluni casi senza nemmeno competenze di programmazione (si parla di sviluppo *low-code* o addirittura *zero-code*) rappresenta la complessa dinamica dell'era digitale dell'IA. Da un lato, questa tendenza favorisce la democratizzazione delle tecnologie avanzate, stimolando l'innovazione e accogliendo nuovi talenti nell'arena tecnologica. Dall'altro, introduce un'era di opportunità che, seppur entusiasmante, sfida la società ad elevare la qualità, la sicurezza e gli standard etici delle applicazioni di IA. Questa democratizzazione incoraggia una più ampia partecipazione alla creazione di tecnologie, seppur nei limiti scientifico-tecnologici detti sopra, arricchendo l'ecosistema tecnologico con prospettive diverse e, sperabilmente, accelerando le scoperte tecnologiche e creative.

Questa tendenza è catalizzata dalla disponibilità ubiqua di strumenti di sviluppo di AI "a bassa soglia di accesso" e piattaforme cloud che offrono servizi di IA come commodity. Si pensi solo al fatto che l'interazione con un LLM o un sistema di generazione di immagini quali, ad esempio, DALLE-3, richiede il solo uso di un linguaggio naturale. Sebbene questa accessibilità incoraggi una più ampia partecipazione all'innovazione tecnologica, porta con sé la questione della diluizione dell'*expertise*.

In questo contesto, l'assenza di una profonda comprensione dei principi fondamentali dell'IA e del *machine learning* può tradursi in una serie di problemi. Primo fra tutti, la gestione del *bias* nei modelli di IA: pur con una solida competenza in materia, gli sviluppatori possono involontariamente introdurre o perpetrare pregiudizi esistenti nei dati di *training*, risultando in sistemi che discriminano basandosi su genere, razza, età o altre caratteristiche personali. Questo non solo solleva questioni di equità e giustizia ma minaccia anche la validità e l'affidabilità delle applicazioni IA.

Inoltre, man mano che le tecnologie di IA, come l'apprendimento automatico, vengono integrate in vari settori, il divario di competenze si allarga. Gli sviluppatori possono affidarsi sempre più pesantemente agli strumenti di IA per compiti quali la generazione di codice e il *debugging* (citiamo, ad esempio, devin.ai), il che può comportare rischi se il codice generato dall'IA non viene adeguatamente controllato per individuare errori o difetti di sicurezza. Questo evidenzia la necessità di approcci equilibrati che combinino l'accessibilità degli strumenti di IA con una solida formazione e supervisione per mitigare i rischi associati alla mancanza di conoscenze specialistiche.

Rischi dell'IA nell'ambito della ricerca scientifica

L'adozione dell'IA nella conduzione della ricerca scientifica ha accelerato significativamente il ritmo delle scoperte, permettendo l'analisi di *dataset* di dimensioni e di complessità precedentemente inimmaginabili. Modelli predittivi avanzati e algoritmi di apprendimento automatico sono ora strumenti fondamentali per decifrare *pattern* complessi, siano questi sequenze genomiche o tendenze climatiche globali. Tuttavia, questo utilizzo pionieristico dell'IA solleva questioni relative alla riproducibilità degli esperimenti e alla validità dei modelli adoperati.

Ad esempio, nell'IA per la sanità, i problemi di riproducibilità sono particolarmente evidenti a causa di complessità quali problemi di privacy, considerazioni etiche e vincoli normativi. Questi fattori spesso rendono difficile replicare i risultati dell'IA in modo coerente, incidendo sull'affidabilità dei modelli utilizzati in settori critici come la diagnostica medica e le raccomandazioni terapeutiche (Sohn, 2023).

Inoltre, è noto come, in assenza di una suddivisione oculata tra dati di *training* e dati di *testing*, vi è un rischio di *overfitting*, che può portare a modelli che funzionano bene sui dati di addestramento ma male negli scenari del mondo reale, gonfiando falsamente le prestazioni di un modello e compromettendone così la generalizzabilità e l'utilità dei risultati prodotti.

Un'area di preoccupazione nel mondo accademico è l'aumento dei contenuti generati dall'IA, che possono confondere i confini tra la ricerca legittima e i contributi meno rigorosi, a volte discutibili (Dalmeet, 2024). Ciò è diventato particolarmente evidente in quanto gli strumenti di IA sono diventati capaci di produrre contenuti accademici estesi in modo rapido e a basso costo, portando potenzialmente a un aumento delle pubblicazioni meno significative. Questi sviluppi possono mettere ulteriormente a dura prova il sistema di *peer review* e sfidare la fiducia e la credibilità tradizionalmente associate alle pubblicazioni accademiche.

Per affrontare queste sfide, la comunità accademica sta valutando nuovi meccanismi di verifica e valutazione (Van Noorden, 2022). Questi meccanismi mirano a garantire l'integrità e l'autenticità della ricerca accademica in un contesto di crescente utilizzo dell'IA nella scrittura accademica. Se da un lato questa situazione presenta delle sfide, dall'altro stimola l'innovazione nel modo in cui l'integrità accademica viene mantenuta e valutata, garantendo che il rapido avanzamento delle tecnologie di IA si allinei con i principi fondamentali di un'indagine scientifica credibile e affidabile.

IA neuro-simbolica e all'IA guidata dalla fisica

Nel discorso che circonda l'evoluzione e le implicazioni dell'IA sia nell'ambito della ricerca scientifica che nel contesto sociale più ampio, l'emergere dell'IA neuro-simbolica presenta una congiuntura affascinante. Questo approccio cerca di unire i punti di forza delle reti neurali, che eccellono nella gestione di dati non strutturati attraverso tecniche di *deep learning*, con l'IA simbolica, che utilizza una logica strutturata e basata su regole per eseguire compiti di ragionamento, in qualche modo unificando due filoni di ricerca in IA classicamente contrapposti.

L'AI neuro-simbolica mira a superare la natura di "scatola nera" degli attuali modelli di *deep learning*, che, sebbene potenti, spesso mancano di trasparenza nei processi decisionali. Integrando il ragionamento simbolico, i sistemi neuro-simbolici

possono fornire spiegazioni esplicite per le loro conclusioni, migliorando così l'interpretabilità e l'affidabilità delle tecnologie di IA. Questa capacità è particolarmente pertinente per la comunità scientifica, dove comprendere la logica alla base dell'analisi o della previsione di un'IA è fondamentale per convalidare i risultati della ricerca e garantirne la riproducibilità.

Inoltre, l'incorporazione del ragionamento simbolico consente all'IA neuro-simbolica di eseguire compiti più sofisticati e basati sulla logica, come generare ipotesi o progettare esperimenti, che sono oltre la portata delle reti neurali convenzionali. Ciò potrebbe rivoluzionare la ricerca scientifica, facilitando processi di scoperta più efficienti e consentendo agli scienziati di esplorare questioni complesse che prima erano irraggiungibili.

In sostanza, l'IA neuro-simbolica rappresenta un passo avanti fondamentale nello sviluppo delle tecnologie di AI, promettendo di affrontare alcune delle questioni più urgenti affrontate oggi dalle comunità scientifiche e tecnologiche. Colmando il divario tra apprendimento profondo e ragionamento simbolico, l'IA neuro-simbolica non solo mira a migliorare le capacità e la comprensibilità dei sistemi di IA, ma promette di contribuire anche alla sostenibilità e all'integrità della ricerca scientifica nell'era digitale. Mentre esploriamo le complessità dell'integrazione dell'IA in vari ambiti, l'esplorazione e il progresso dell'IA neuro-simbolica rappresentano un faro di progresso, offrendo un percorso più trasparente, affidabile ed etico.

La letteratura offre diversi approcci per l'integrazione di architetture neurali e sistemi di inferenza simbolici. Questi approcci includono l'uso di modelli simbolici per guidare elaborazioni neurali, come fatto ad esempio in AlphaGo. Alcuni metodi prevedono l'uso di architetture neurali per interpretare dati percettivi ottenuti da diversi sensori col fine di poterli integrare nel ragionamento simbolico o l'impiego del ragionamento simbolico per generare dati di addestramento per i modelli neurali. Altri approcci creano reti neurali a partire da regole simboliche o consentono ai modelli neurali di effettuare vere e proprie chiamate a motori di ragionamento simbolico (ChatGPT,

ad esempio, offre tra i suoi *plug-in* il motore algebrico/matematico WolframAlpha). Il confine tra neurale e simbolico diventa più labile quando, nell'analisi, si considera la natura matematica dei blocchi che costituiscono una rete neurale profonda, dove si possono individuare, tra gli altri, grafi, ipergrafi e operatori di convoluzione, anche se, va notato, la presenza di questi oggetti di natura prettamente matematica permetta una interpretabilità più agevole del processo soggiacente alla rete stessa (Sarker et al., 2021).

In talune applicazioni di natura fisico-matematica dove la necessità di un modello simbolico e interpretabile è fondamentale, lavori recenti hanno mostrato come l'uso combinato di reti neurali ed algoritmi (tipicamente genetici, anch'essi basati su principi biologici) possa portare all'apprendimento di leggi algebrico-trigonometriche completamente simboliche. L'idea di fondo è, in una prima fase, addestrare una rete neurale per catturare le relazioni ingresso-uscita del fenomeno, seppur in modalità del tutto numerica, analizzarne quindi l'architettura appresa così da poter partizionarne i nodi in sottogruppi con un numero ridotto di ingressi e uscite e, in una seconda fase, applicare un algoritmo che individui le forme algebrico-trigonometriche che meglio approssimano ogni porzione della rete (Cranmer, 2023).

L'IA guidata dalla fisica incorpora alcune delle leggi fisiche conosciute nei modelli di apprendimento automatico, garantendo che le previsioni non solo si adattino ai dati osservati, ma siano anche in linea con i principi scientifici consolidati. Ad esempio, nella modellazione climatica, i modelli di intelligenza artificiale basati sulla fisica vengono utilizzati per migliorare la previsione dei modelli meteorologici incorporando le leggi di conservazione (come quelle della massa e dell'energia) nell'architettura del modello, che porta a previsioni meteorologiche più affidabili e interpretabili. Un'altra applicazione è nel campo della scienza dei materiali, dove i modelli di intelligenza artificiale predicono le proprietà dei nuovi materiali sulla base delle leggi della meccanica

quantistica, facilitando una più rapida scoperta e validazione di nuovi materiali adatti ad applicazioni ad alte prestazioni.

In sostanza, sia l'IA spiegabile che quella guidata dalla fisica sono fondamentali per il progresso delle tecnologie di IA, garantendo che i sistemi di IA, pur diventando sempre più sofisticati, rimangano comprensibili e saldamente radicati nei principi scientifici. Questo duplice approccio non solo migliora l'applicabilità dell'IA e la sua efficienza predittiva in vari ambiti, ma garantisce anche che i progressi dell'IA contribuiscano positivamente alle esigenze della società e alle scoperte scientifiche.

2. Opacità e ricadute legali

Come già accennato in precedenza, i sistemi di IA sono strutturalmente caratterizzati da un significativo grado di opacità (c.d. effetto "scatola nera") (Pasquale, 2015), essenzialmente dovuto all'impiego di modelli di apprendimento automatico, in particolare basati sull'apprendimento profondo, di natura prettamente numerica (o *vettoriale*), anziché simbolica e basata su regole ed equazioni facilmente interpretabili dall'uomo. L'opacità è foriera di implicazioni sociali, etiche e legali per gli utenti che, talvolta inconsapevolmente, sono esposti all'uso di questa tecnologia. Si pensi all'impiego di un sistema di IA in ambito medico. Se è dimostrato che l'ingresso in corsia di questi strumenti consente prestazioni più precise, rapide ed efficaci in molti e diversi campi della medicina, è altrettanto vero che da esso e dalla non spiegabilità dei risultati offerti, possono derivare nuovi rischi in termini di lesione dei diritti fondamentali dell'individuo. Può, ad esempio, essere pregiudicato il diritto alla salute, riconosciuto e tutelato dall'art. 32 Cost., qualora dal malfunzionamento dell'IA utilizzata per il trattamento di una patologia derivi un danno all'integrità psico-fisica del paziente, magari dovuto all'errata somministrazione della terapia[1]

[1] L'ipotesi non è affatto improbabile, essendosi già verificati casi in cui l'errato funzionamento della macchina è stato causa di morte o di lesioni per il paziente. Risalgono,

(Levenson & Turner, 1993). Inoltre, l'impiego dell'IA in corsia può minare il diritto all'autodeterminazione terapeutica del paziente, laddove il medico aderisca in modo acritico ad un certo risultato diagnostico frutto di un processo decisionale opaco e non spiegabile, nei confronti del quale il malato non possa esprimere un consenso realmente informato perchè non in grado di comprendere le ragioni che hanno condotto a quel determinato output.

Tutelare i diritti fondamentali di fronte all'IA

L'IA può contribuire a proteggere la sicurezza dei cittadini e consentire loro di godere dei diritti fondamentali; tuttavia, vi è il giustificato timore – espresso in più occasioni dalle stesse istituzioni europee[2] – che l'intelligenza artificiale possa avere effetti indesiderati o essere utilizzata impropriamente per scopi dolosi, arrecando pregiudizio agli utenti.

A queste preoccupazioni occorre dare una risposta tenendo conto del fatto che i sistemi di IA non operano in un mondo privo di riferimenti normativi. Numerose sono le fonti giuridicamente vincolanti – attualmente in vigore a livello europeo, nazionale e internazionale – poste a presidio dei diritti degli individui. Si pensi – solo per citare le più rilevanti – al diritto primario dell'UE (i trattati dell'Unione europea e la sua Carta dei diritti fondamentali), al diritto derivato dell'UE (ad esempio il regolamento generale sulla protezione dei dati (GDPR), le direttive antidiscriminazione, la

per esempio, agli anni '80 diversi malfunzionamenti di Therac-25, macchinario che a causa di una combinazione di bug nel software e difetti di progettazione aveva somministrato dosi di radiazioni eccessive a sei pazienti. In tre casi ne era derivata la morte del paziente, in altri tre una serie di lesioni di una certa importanza (perdita dell'uso del braccio, asportazione del seno, sostituzione totale dell'anca).

[2] Parlamento europeo, Norme di diritto civile sulla robotica Risoluzione del Parlamento europeo del 16 febbraio 2017 recante raccomandazioni alla Commissione concernenti norme di diritto civile sulla robotica; Commissione europea, Libro bianco sull'intelligenza artificiale - Un approccio europeo all'eccellenza e alla fiducia, Bruxelles, 19.2.2020; Gruppo di esperti ad alto livello sull'intelligenza artificiale, Orientamenti etici per un'IA affidabile, 8.6.2019; Parlamento europeo, Risoluzione del Parlamento europeo del 20 ottobre 2020 recante raccomandazioni alla Commissione su un regime di responsabilità civile per l'intelligenza artificiale.

direttiva macchine, la direttiva sulla responsabilità dei prodotti, il regolamento sulla libera circolazione dei dati non personali, il diritto dei consumatori e le direttive in materia di salute e sicurezza sul lavoro), ma anche ai trattati ONU sui diritti umani e le convenzioni del Consiglio d'Europa (come la Convenzione europea dei diritti dell'uomo) e a numerose leggi degli Stati membri dell'UE. In più, oltre alle norme applicabili orizzontalmente, esistono varie discipline specifiche per settore riferibili a particolari applicazioni di IA (è il caso, ad esempio, del regolamento sui dispositivi medici nel settore sanitario).

Lo sviluppo, la commercializzazione e l'uso di sistemi di IA deve, dunque, risultare compatibile con il sostrato comune di normative appena richiamato che, pur non essendo stato adottato con specifico riferimento a questa tecnologia, costituisce un primo, generale perimetro di liceità entro cui debbono svolgersi le attività di progettazione, addestramento, fabbricazione e utilizzo.

Soltanto la conformità dell'IA ai principi emergenti dal quadro normativo accennato potrà consentire la "costruzione di un ecosistema di fiducia"[3] tale da spingere i cittadini ad adottare e ad affidarsi a questi sistemi e garantire alle imprese e alle organizzazioni pubbliche la certezza del diritto necessaria per innovare utilizzando questa tecnologia.

In generale, molteplici sono i diritti che possono essere negativamente impattati dai sistemi di IA; come ciò avvenga dipende in gran parte dall'ambito di applicazione in cui tale tecnologia trova impiego. Esistono, invero, settori in cui l'uso dell'IA può, più che in altri, avere ricadute significative sulla sicurezza e sui diritti dei cittadini; è il caso, ad esempio, del campo dell'istruzione e del lavoro, dei servizi essenziali pubblici e privati, dell'amministrazione della giustizia e della gestione delle migrazioni e del controllo delle frontiere. Non sembra un caso che quelli appena citati siano, peraltro, alcuni tra i settori individuati come "ad alto rischio" dal recentissimo regolamento UE noto

[3] Come auspicato dalla stessa Commissione europea con la comunicazione dal titolo *Creare fiducia nell'intelligenza artificiale antropocentrica*, 8.4.2019.

come *Artificial Intelligence Act* (AI Act), volto a dettare una disciplina *ad hoc* per i sistemi di IA dalla fase di progettazione sino a quella post commercializzazione (v. infra il paragrafo dedicato all'AI Act).

Le conseguenze dell'impiego dell'IA sui diritti degli individui sono, dunque, eterogenee e trasversali, come gli ambiti in cui questi sistemi trovano applicazione. Vi sono, però, rischi comuni a molte, se non a tutte, le applicazioni di IA.

Tra questi si annovera il rischio di discriminazione, di cui l'IA può essere veicolo attraverso l'assunzione di decisioni basate su dati non sufficientemente rappresentativi, vale a dire che non soddisfano i requisiti di qualità e quantità richiesti affinché ciascuna delle categorie di soggetti impattati dal sistema sia correttamente rappresentata. Invero, se l'IA viene addestrata sulla scorta di dati imprecisi, carenti o viziati, l'output che ne deriva sarà inevitabilmente inaccurato, con il rischio di risultati iniqui e discriminatori nei confronti dei soggetti non abbastanza rappresentati. In questo senso l'IA sembra una sorta di amplificatore dei pregiudizi che già caratterizzano la società, in quanto capace di decisioni e azioni discriminatorie a danno di categorie di persone già emarginate o svantaggiate (Criado & Such, 2019; Stradella, 2020; Micklitz et al., 2021). Questo aspetto si dimostra ancor più critico se si tiene conto che questa tecnologia sembra ammantata da un velo di neutralità e oggettività delle decisioni assunte, che non consente di intravedere - e quindi di contestare - l'eventuale esistenza di *dataset* discriminatori.

A questo riguardo risultano di sicuro interesse, ad esempio, gli effetti dell'impiego dell'IA nella prospettiva della parità di genere. Anche in questo ambito i sistemi algoritmici rischiano di riprodurre i modelli diffusi nella società e non ancora scardinati, in particolare l'idea di un mondo costruito su misura di uomini, bianchi e di classe elevata (cfr. M. D'Amico, *Intelligenza artificiale "contro" le donne*, in *Una parità ambigua. Costituzione e diritti delle donne*, 2020, 314). Ciò è in parte dovuto al divario di genere che caratterizza i settori più innovativi in materia di nuove

tecnologie e intelligenza artificiale che vedono solo il 29,2% di lavoratrici nel settore STEM e solo il 30% di donne nell'ambito AI[4]. Tale divario aggrava le attuali disparità di genere nella forza lavoro, in un settore peraltro in rapida crescita e in cui ci si attende che nei prossimi anni vi saranno significative opportunità occupazionali. La scarsa presenza di donne nel settore AI e il fatto che siano perlopiù gli uomini a concepire gli algoritmi, attingendo spesso a dati viziati da pregiudizi di genere, determina il serio rischio che la tecnologia si alimenti e diffonda stereotipi di genere. Si pensi, ad esempio, al sistema di reclutamento del personale utilizzato, e poi abbandonato, da Amazon che favoriva sistematicamente candidati di sesso maschile in quanto addestrato prevalentemente attraverso l'immissione di curricula maschili (Dustin, 2022; Chang, 2023).

La presenza di *bias* nei sistemi di IA rischia, quindi, di promuovere la diffusione di una "società algoritmica" in cui vengono meno le garanzie, anche giuridiche, che dovrebbero proteggere e tutelare le persone dalle indebite ingerenze di questo nuovo potere tecnologico. Per questo è essenziale presidiare lo sviluppo di questa tecnologia, affinché non si trasformi in uno strumento di perpetuazione su larga scala di modelli che faticosamente si stanno tentando di scardinare.

A questo scopo si segnala la recente adozione da parte del Consiglio d'Europa della *Framework convention on artificial intelligence, human rights, democracy and the rule of law*[5]. La Convenzione rappresenta il primo trattato internazionale giuridicamente vincolante in materia di IA e ha come oggetto la tutela dei diritti umani, della democrazia e dello Stato di diritto rispetto a tutte le attività inerenti al ciclo di vita di un sistema di IA, dalla progettazione, allo sviluppo, passando per l'uso e la dismissione.

[4] V. World Economic Forum, "Global Gender Gap Report 2023", in particolare la sezione "Gender gaps in the labour markets of the future" e i paragrafi "STEM occupations" e "AI occupation take-up".

[5] Consultabile all'indirizzo: https://rm.coe.int/0900001680afb122

Diritto alla spiegabilità e principio di non esclusività

Con specifico riguardo alla sua intrinseca opacità e in base all'ambito in cui trova applicazione, l'uso dell'IA può determinare una elusione dei principi che conformano l'ordinamento, con conseguente lesione dei diritti degli utenti (Ebers, 2020).

Risulta in primo luogo negativamente inciso dall'impiego dell'intelligenza artificiale il diritto alla spiegabilità delle decisioni automatizzate, espressamente riconosciuto dagli artt. 13, 14 (Informativa) e 15 (Diritto di accesso) del regolamento generale sulla protezione dei dati (GDPR), secondo il quale ognuno ha diritto a conoscere l'esistenza di processi decisionali automatizzati che lo riguardino ed in questo caso a ricevere informazioni significative sulla logica utilizzata, nonché sull'importanza e le conseguenze previste di tale trattamento.

Si pensi al ricorso a procedure automatizzate nel settore della pubblica amministrazione (PA). È indubbio che il legislatore italiano, da diverso tempo a questa parte, stia cercando di incentivare un migliore livello di digitalizzazione dell'amministrazione pubblica, considerato fondamentale per migliorare la qualità dei servizi resi ai cittadini e agli utenti[6]. È ammesso dalla dottrina (che ha a tal proposito elaborato la nozione di e-government) e dalla stessa giurisprudenza che ciò possa avvenire anche attraverso l'automazione dei processi decisionali della PA, mediante il ricorso a procedure che utilizzano algoritmi in grado di valutare e graduare una grande mole di domande. "L'utilità di tale modalità operativa di gestione dell'interesse pubblico è particolarmente evidente con riferimento a procedure seriali o standardizzate, implicanti l'elaborazione di ingenti quantità di istanze e caratterizzate dall'acquisizione di dati certi ed oggettivamente comprovabili e

[6] Riveste un ruolo centrale in questo senso l'adozione del Codice dell'Amministrazione Digitale (CAD). Trattasi di un testo unico che riunisce e organizza le norme riguardanti l'informatizzazione della Pubblica Amministrazione nei rapporti con i cittadini e le imprese. Nella stessa direzione si muove l'ordinamento comunitario (v. tra l'altro la Comunicazione della Commissione sull'Agenda digitale europea).

dall'assenza di ogni apprezzamento discrezionale"[7]. Ciò è, altresì conforme ai canoni di efficienza ed economicità dell'azione amministrativa (art. 1 L. n. 241 del 1990), i quali, secondo il principio costituzionale di buon andamento dell'azione amministrativa (art. 97 Cost.), impongono all'amministrazione il conseguimento dei propri fini con il minor dispendio di mezzi e risorse e attraverso lo snellimento e l'accelerazione dell'iter procedimentale.

Tuttavia, l'impiego di sistemi di IA nel settore della PA non può porsi in contrasto con i principi che regolano lo svolgersi dell'attività amministrativa. Tra questi vi è, senza dubbio, il principio di trasparenza, che impone necessariamente la conoscibilità del meccanismo attraverso il quale si concretizza la decisione robotizzata (ovvero l'algoritmo). Tale conoscibilità dell'algoritmo – osserva la giurisprudenza amministrativa – "deve essere garantita in tutti gli aspetti: dai suoi autori al procedimento usato per la sua elaborazione, al meccanismo di decisione, comprensivo delle priorità assegnate nella procedura valutativa e decisionale e dei dati selezionati come rilevanti. Ciò al fine di poter verificare che gli esiti del procedimento robotizzato siano conformi alle prescrizioni e alle finalità stabilite dalla legge o dalla stessa amministrazione a monte di tale procedimento e affinché siano chiare - e conseguentemente sindacabili - le modalità e le regole in base alle quali esso è stato impostato".

Ne deriva che l'opacità dei sistemi di IA non può esimere dalla necessità di una spiegazione dell'algoritmo impiegato, tale da

[7] Così osserva il Consiglio di Stato, Sez. VI, Sent., 08.04.2019, n. 2270. La controversia è relativa all'impiego di procedure algoritmiche per la valutazione dei titoli e la mobilità di docenti della scuola secondaria di secondo grado. Questi ultimi, quali appellanti contro il Ministero dell'Istruzione dell'Università e della Ricerca (MIUR), lamentano che "in conseguenza di tale procedura, si sono ritrovati destinatari di una nomina su classi di concorso ed ordine di scuola in cui non avevano mai lavorato; inoltre, pur avendo espresso nella domanda di assunzione la preferenza per la scuola superiore di secondo grado, sono risultati destinatari di proposta di assunzione nella scuola superiore di primo grado; infine, tutti gli appellanti sono stati destinati in province lontane, rispetto a quella di provenienza". Oltre all'illogicità degli esiti della procedura gli appellanti lamentano l'impossibilità di comprendere le modalità con le quali, attraverso l'algoritmo, sono stati assegnati i posti disponibili. Negli stessi termini e in una controversia avente il medesimo oggetto si è poi nuovamente pronunciato il Consiglio di Stato, Sez. VI, Sent., 13.12.2019, n. 8472.

renderlo comprensibile dai cittadini che intendono contestare le modalità di esercizio del potere amministrativo e dal giudice chiamato a valutare la decisione adottata dalla PA.

A tale scopo è, dunque, essenziale che vengano rese note tutte le componenti del processo informatico cui gli utenti sono stati sottoposti: dalla sua costruzione, all'inserimento dei dati, alla loro validità, alla loro gestione. Solo la conoscibilità dei dati immessi nel sistema e dell'algoritmo medesimo, può consentire un reale sindacato di legittimità della decisione amministrativa automatizzata. A questo riguardo va, però, dato conto del fatto che per i sistemi che utilizzano reti neurali (non è il caso di quello della vicenda giudiziaria riportata) la conoscibilità richiesta e auspicata dal giudice amministrativo è impraticabile. Nelle IA che fanno ricorso a reti neurali, infatti, è illusorio pensare di poter davvero ricostruire i passaggi logici che hanno condotto ad un determinato risultato, poiché si tratta di sistemi che operano sulla base di interazioni numeriche. A differenza dei modelli matematici tradizionali che permettono inferenzialmente di comprendere quale sia l'iter seguito dal sistema e, eventualmente, di individuare gli errori, i modelli basati su reti neurali offrono un risultato che non può essere ricostruito *ex post* per valutare eventuali *bias*.

Tale necessaria conoscibilità – precisano i giudici – non può subire limitazioni neppure di fronte alla protezione della riservatezza, spesso invocata dalle imprese produttrici dei sistemi di IA e da coloro che si occupano della progettazione dell'architettura informatica utilizzata. Questi attori economici, infatti, ponendo al servizio del potere autoritativo tali strumenti, all'evidenza ne accettano le relative conseguenze in termini di necessaria trasparenza.

L'opacità, dunque, specie nel settore dell'esercizio di funzioni pubbliche ma non soltanto, rappresenta un limite significativo che impedisce di verificare il funzionamento del processo decisionale e controllare se il sistema di IA abbia rispettato gli standard e le norme giuridiche vigenti nel contesto in cui si trova ad operare. Ciò si traduce in una grave carenza in termini di spiegabilità della motivazione posta alla base della decisione e,

conseguentemente, della sua legittimità, dal momento che l'impossibilità di conoscere il percorso seguito dal meccanismo di inferenza e produzione può ridurre significativamente l'accesso agli strumenti giuridici normalmente preposti a porre rimedio ad eventuali illegittimità o irregolarità verificatesi durante il processo decisionale (Orsoni & D'Orlando, 2019; Fasan, 2022).

Il rischio di ricadute negative causate dall'opacità dell'IA sul diritto alla spiegabilità della decisione algoritmica non è circostanza esclusiva del settore pubblico. Molti sono gli attori privati che operano sul mercato e commercializzano sistemi in grado di impattare negativamente sul diritto alla conoscibilità dell'algoritmo da parte degli utenti. Si pensi all'impiego di sistemi di *scoring* volti alla determinazione di un "punteggio reputazionale" del soggetto (persona fisica o giuridica) sulla base di una molteplicità di indicatori (per le persone fisiche si considerano, ad esempio, istruzione, lavoro, reati, inadempimenti fiscali, vertenze tra privati).

In una recente decisione la Corte di Cassazione (Cerea, 2023)[8] ha avuto modo di dirimere una controversia tra una società privata operante nel settore del rating reputazionale e il Garante per la protezione dei dati personali. Quest'ultimo, infatti, aveva vietato all'associazione qualsiasi trattamento di dati personali degli utenti, non ritenendo il suo operato conforme a quanto stabilito dal Codice in materia di protezione dei dati personali (d. lgs. 196/2003) applicabile *ratione temporis*.

Tralasciando di ricostruire l'articolata vicenda giudiziaria, la pronuncia risulta particolarmente significativa in quanto offre, per la prima volta nel panorama italiano, indicazioni su quali elementi debbano essere concretamente resi noti agli interessati relativamente al funzionamento di un sistema algoritmico (nel caso di specie, di *social scoring*). Secondo la giurisprudenza è essenziale che la logica del meccanismo decisionale venga spiegata alle persone attraverso il linguaggio naturale, senza la trasmissione di informazioni tecniche relative all'operatività

[8] Cass. civ., Sez. I, Ord., 10.10.2023, n. 28358.

dell'IA, le quali sarebbero comunque di difficile comprensione per i non addetti ai lavori (si pensi a quanto questo risulti, per sua essenza, inapplicabile di fronte ad una decisione presa da una rete neurale). Va dunque perseguita l'intellegibilità dello schema esecutivo del sistema e degli elementi dallo stesso considerati nell'elaborazione del risultato, portando a conoscenza degli utenti in modo chiaro, semplice e accessibile il metodo impiegato dall'IA e i dati utilizzati per il trattamento, anche mediante l'indicazione degli scopi perseguiti dal sistema e della tecnica utilizzata.

Secondo questa impostazione, a fronte di una completa informativa le persone sono messe nella migliore condizione di acconsentire consapevolmente al trattamento cui si sottopongono, prefigurandosene le conseguenze. Inoltre, conoscendo la logica sottesa al funzionamento del sistema, ne possono contestare gli esiti, ricorrendo all'autorità laddove ritengano di aver subito una lesione dei propri diritti.

Tuttavia, l'intrinseca opacità dei sistemi di IA – specialmente se basati su modelli di apprendimento automatico profondo – non consente di poter sempre fornire una spiegazione intellegibile ed esaustiva del meccanismo decisionale sotteso, talvolta di impossibile ricostruzione *ex post* da parte degli stessi soggetti che si sono occupati dell'addestramento del modello (e che non hanno, come spesso si pensa, programmato linea per linea il codice che il modello esegue – attività questa che ne ridurrebbe drasticamente il livello di opacità). In questo senso il consenso dell'interessato al trattamento dei propri dati da parte di sistemi automatizzati, per quanto rimanga principio chiave della disciplina in materia, sembra destinato a giocare un ruolo più circoscritto di fronte a sistemi dotati di un grado di autonomia che consente loro di discostarsi dalle finalità inizialmente stabilite in fase di progettazione (la cosiddetta abilità emergente). Sembra, dunque, più ragionevole che l'opacità dell'IA non venga gestita solo a valle attraverso l'espressione di un valido consenso dell'interessato, ma piuttosto e preliminarmente a monte, mediante la previsione dell'obbligo di rispettare i principi di *privacy by design* e *privacy by default* sin dalla fase di

progettazione di tali sistemi, pena il loro divieto di commercializzazione nel territorio dell'Unione (v. infra il paragrafo dedicato all'AI Act). Di fronte alla tipica opacità di questi sistemi, infatti, il risultato da perseguire non è tanto quello della spiegabilità (*explicability*) di come funziona la c.d. *black box*, quanto piuttosto quello della interpretabilità (*interpretability*), cioè della progettazione di modelli che intrinsecamente rendano il processo decisionale o predittivo comprensibile agli esperti di dominio.

L'impiego di avanzati sistemi di IA, inoltre, può incidere in modo significativo sul c.d. "diritto all'oblio", vale a dire sul diritto dell'utente ad ottenere la cancellazione dei propri dati da parte del titolare del trattamento (art. 17 GDPR). Trattasi di un diritto che sul piano tecnico-pratico risulta molto difficile, per non dire impossibile, garantire, in quanto all'interno di una rete neurale i dati dell'utente sono "memorizzati" in un formato implicito mediante miliardi di numeri, differentemente dal caso "classico" in cui i dati sono contenuti in una base dati e la cancellazione è, di fatto, banale. Da ciò deriva che il dato che si vorrebbe cancellare non è chiaramente individuabile ed è tutt'altro che semplice sapere quale dei moltissimi "pesi" deve essere modificato per ottenere la cancellazione e quale debba essere il valore numerico di tale variazione.

A ciò si aggiunga che le grandi società leader del settore IA sono molto spesso situate negli Stati Uniti, ove non sempre i dati dell'utente europeo vengono trattati in modo conforme a quanto stabilito dalla normativa europea (GDPR). La questione è stata sottolineata dalla stessa Corte di giustizia dell'Unione europea (C-311/18, *Facebook Ireland Limited VS Maximillian Schrems*) che ha invalidato il precedente accordo tra UE e USA in materia di trasferimento extra-UE di dati personali di cittadini europei, noto come "Safe Harbor", proprio per le carenze del sistema di protezione dei dati previsto dall'ordinamento americano e dall'accordo stesso. Secondo la Corte, infatti, le limitazioni previste dalla legislazione USA non rispondono ai requisiti richiesti, nel diritto dell'Unione, dal principio di proporzionalità,

giacché i programmi di sorveglianza fondati sulla suddetta normativa non si limitano a quanto strettamente necessario. Inoltre, la Corte segnala la mancanza per gli utenti di un mezzo di ricorso dinanzi ad un organo che offra garanzie sostanzialmente equivalenti a quelle richieste nel diritto UE, non essendo assicurata né l'indipendenza dell'organismo di mediazione previsto dall'accordo, né la vincolatività delle sue decisioni nei confronti dei servizi di intelligence statunitensi. Ad oggi la questione pare arginata dal nuovo accordo UE-USA raggiunto nel 2023, volto a consentire un trasferimento sicuro di dati verso gli Stati Uniti. Il Privacy Shield, infatti, prevede che le autorità americane vigilino e assicurino con più forza sul rispetto dell'accordo e che collaborino in misura maggiore con le Autorità europee per la protezione dei dati. L'accordo contiene - ed è la prima volta - dichiarazioni e impegni assunti formalmente per quanto riguarda l'accesso ai dati da parte di soggetti dell'Amministrazione americana.

L'uso dell'IA è suscettibile di ripercuotersi negativamente anche sul divieto di cui all'art. 22 del GDPR, secondo cui «l'interessato ha il diritto di non essere sottoposto a una decisione basata unicamente sul trattamento automatizzato, compresa la profilazione, che produca effetti giuridici che lo riguardano o che incida in modo analogo significativamente sulla sua persona». Quello che a prima vista sembrerebbe un divieto assoluto di impiego di sistemi di IA ammette, in realtà, una deroga: se il trattamento automatizzato è accompagnato da un intervento umano, lo stesso viene considerato lecito, in ossequio al principio di non esclusività della decisione algoritmica. A ciascun individuo, infatti, è riconosciuto il «diritto di non essere sottoposto a decisioni automatizzate prive di un coinvolgimento umano e che, allo stesso tempo, producano effetti giuridici o incidano in modo significativo sulla sua persona». La normativa sembra quindi considerare come imprescindibile l'esistenza di un contributo umano interno al processo decisionale automatizzato, in grado di controllare, validare o smentire i risultati dell'IA (Pajno et al., 2019; Simoncini, 2019), in coerenza con la visione antropocentrica da

tempo sostenuta dalle istituzioni europee[9]. La disposizione, tuttavia, non declina in concreto tale principio, creando una serie di problemi interpretativi circa l'intensità e la pervasività che l'elemento umano deve possedere per rendere legittimo il trattamento dei dati dell'utente. Consapevoli che il principio, anche per la sua genericità, avrebbe potuto facilmente essere vanificato nella sua effettiva portata dall'impiego sempre più diffuso dell'IA, le stesse istituzioni europee si sono fatte sostenitrici di una sua interpretazione sostanziale[10]. Il principio di non esclusività, infatti, si può considerare soddisfatto solo laddove l'intervento dell'essere umano non sia meramente formale ma esprima una vera e propria valutazione attiva. Tale principio può trovare concretizzazione mediante la predisposizione di meccanismi di governance che garantiscano l'adozione di un approccio con intervento umano ("*human-in-the-loop*"), con supervisione umana ("*human-on-the-loop*") o con controllo umano ("*human-in-command*")[11].

[9] Commissione europea, L'intelligenza artificiale per l'Europa, 25.4.2018; Commissione europea, Creare fiducia nell'intelligenza artificiale antropocentrica, 8.4.2019.

[10] Secondo la Risoluzione del Parlamento europeo del 20 ottobre 2020 recante raccomandazioni alla Commissione su un regime di responsabilità civile per l'intelligenza artificiale, nei casi in cui siano in gioco le libertà fondamentali gli Stati membri dovrebbero ricorrere a sistemi di intelligenza artificiale soltanto «quando sono possibili o sistematici un intervento e una verifica sostanziali da parte dell'uomo». Sul tema anche una pronuncia del Tar Lazio, sez. III-bis, 10 settembre 2018, n. 9227, in cui si stabilisce che «le procedure informatiche, finanche ove pervengano al loro maggior grado di precisione e addirittura alla perfezione, non possano mai soppiantare, sostituendola davvero appieno, l'attività cognitiva, acquisitiva e di giudizio che solo un'istruttoria affidata ad un funzionario persona fisica è in grado di svolgere».

[11] "Per "*human-in-the-loop*" (HITL) si intende l'intervento umano in ogni ciclo decisionale del sistema, cosa che in molti casi non è né possibile né auspicabile. Per "*human-on-the-loop*" (HOTL) si intende la capacità di intervento umano durante il ciclo di progettazione del sistema e di monitoraggio del funzionamento del sistema. Per "*human-in-command*" (HIC) si intende sia la capacità di sorvegliare l'attività complessiva del sistema di IA (compreso il più ampio impatto economico, sociale, giuridico ed etico) sia la capacità di decidere quando e come utilizzare il sistema in una particolare situazione. Si può per esempio decidere di non utilizzare un sistema di IA in una particolare situazione, di definire livelli di discrezionalità umana durante l'uso del sistema o di garantire la possibilità di annullare una decisione adottata dal sistema". Cfr. Commissione europea, *Creare fiducia nell'intelligenza artificiale antropocentrica*, 8.4.2019.

Peraltro, l'efficacia del principio di non esclusività è ulteriormente minata dallo stesso art. 22 del GDPR nella parte in cui prevede delle eccezioni al divieto di decisioni completamente automatizzate, quali la loro necessità per la conclusione o l'esecuzione di un contratto, l'autorizzazione da parte del diritto dell'UE o dello Stato membro, oppure il consenso esplicito dell'interessato. La presenza di tali e tante eccezioni, molto frequenti nella pratica, rende il principio di non esclusività, di fatto, un principio debole (Casonato, 2019).

Tra queste eccezioni – come si è già osservato in merito al diritto alla spiegabilità della decisione automatizzata – il consenso del soggetto rappresenta la principale criticità, in quanto trattasi di una condizione di liceità che se nella pratica mira a garantire l'autodeterminazione informativa dell'utente, nella sostanza rischia di perdere efficacia di fronte all'opacità dei moderni sistemi di IA, che non consentono di comprendere la loro rilevanza in termini di compressione delle libertà fondamentali. Il grado di sofisticatezza raggiunto da certi sistemi e la conseguente inconoscibilità del loro funzionamento determina una situazione di asimmetria informativa tra chi commercializza l'IA e gli utenti che ne fanno uso e si trovano in una vera e propria posizione di soggezione. Che la questione sia cruciale lo dimostra anche la posizione del Gruppo di lavoro per la protezione dei dati personali - Articolo 29 che ha proposto di far gravare sul titolare del trattamento la dimostrazione che «gli interessati comprendono esattamente a cosa stanno acconsentendo»[12]. Dimostrazione che, tuttavia, risulta tutt'altro che semplice sul piano operativo poiché, da un lato, non si dispone di elementi oggettivi di misurazione del grado di consapevolezza degli utenti e, inoltre, perché qualora il titolare non riesca a fornire prova della consapevolezza raggiunta dagli interessati le conseguenze nocive di un trattamento automatizzato si sarebbero comunque già verificate, riversandosi sulla parte debole del rapporto e

[12] Gruppo di lavoro per la protezione dei dati personali – Articolo 29, Linee guida sul processo decisionale automatizzato relativo alle persone fisiche e sulla profilazione ai fini del regolamento 2016/679.

determinando una lacuna nella tutela effettiva degli interessati. Peraltro, anche mettendo da parte per un attimo i rilievi appena citati, risulta assai complesso (se non impossibile) sul piano tecnico garantire la spiegabilità del processo dell'IA ad un utente non esperto, specialmente se si tratta di sistemi basati su reti neurali.

Emerge, dunque, ancora una volta il fatto che il consenso dell'interessato non sempre può rappresentare una base giuridica appropriata per il trattamento dei dati degli individui, soprattutto quanto questo avviene da parte di sistemi di IA intrinsecamente opachi. A ciò si può solo in parte pensare di sopperire presidiando l'IA con una garanzia circa l'intervento umano, laddove lo stesso se dovesse realmente trovare spazio in ogni fase del processo non consentirebbe di raggiungere i vantaggi connessi all'automazione delle decisioni, senza contare l'esistenza di quei processi automatizzati che non consentono, per loro natura, di comprendere le logiche di funzionamento interno (effetto c.d. *black box*).

Il noto caso Loomis vs. State of Wisconsin – per quanto verificatosi negli Stati Uniti – offre un'interessante applicazione del principio appena richiamato. Nella discussa sentenza, risalente al 2016, la Corte Suprema del Wisconsin si è pronunciata sull'appello proposto dal sig. Eric L. Loomis (fermato alla guida di un'automobile precedentemente usata per una sparatoria e dichiaratosi colpevole), la cui pena a sei anni di reclusione era stata comminata dal Tribunale circondariale di La Crosse. Nel determinare la pena, i giudici avevano tenuto conto dei risultati elaborati dal programma COMPAS (*Correctional offender management profiling for alternative sanctions*) di proprietà della società Northpointe (ora Equivant), secondo cui Loomis era da identificarsi quale soggetto ad alto rischio di recidiva[13]. I giudici –

[13] COMPAS consiste in uno strumento di valutazione concepito, da un lato, per prevedere il rischio di recidiva, dall'altro, per identificare i bisogni dell'individuo in aree quali occupazione, disponibilità di alloggio ed abuso di sostanze stupefacenti. L'algoritmo elabora i dati ottenuti dal fascicolo dell'imputato e dalle risposte fornite nel colloquio con lo stesso. Per quanto riguarda la valutazione del rischio, l'elaborato consiste in un grafico di tre barre che rappresentano in una scala da 1 a 10 il rischio di recidiva preprocessuale, il

in disaccordo con le censure mosse da Loomis e dai suoi difensori – hanno confermato la sentenza del Tribunale circondariale, sottolineando la legittimità dell'uso di strumenti di IA nel processo, i cui risultati, però, non possono essere posti a fondamento della decisione giudiziale, dovendo essere considerati, unitamente a tutti gli altri fattori, solo quale elemento concorrente all'adozione della sentenza, secondo un apprezzamento discrezionale del giudice con riguardo ad ogni specifico caso, in un'operazione di bilanciamento riservata in ultima istanza all'uomo.

La partecipazione umana alla decisione rappresenta, quindi, un altro degli accorgimenti necessari affinché l'opacità da cui sono affetti i sistemi di IA non determini conseguenze pregiudizievoli a danno di coloro che sono in qualche misura incisi nella propria sfera giuridica dai risultati della tecnologia.

Gli impieghi accennati fino a qui dimostrano come pur producendo risultati pratici altamente efficaci, le decisioni algoritmiche potrebbero minare i diritti e le garanzie procedurali e sostanziali relative alla democrazia e allo Stato di diritto. La loro connaturata opacità deve, perciò, essere arginata anche - e soprattutto – mediante una regolamentazione che definisca *ex ante* i requisiti di conformità cui gli attori economici debbono attenersi e le eventuali responsabilità in caso di mancata compliance.

Adozione dell'AI Act

Una delle soluzioni approntate per affrontare la sfida dell'opacità dei sistemi di IA è stata la codificazione di processi e requisiti per garantire la trasparenza, la spiegabilità e l'interpretabilità dei risultati attraverso iniziative legislative o

rischio di recidiva generale ed il rischio di recidiva violenta. I punteggi di rischio sono volti a predire la probabilità generale che gli individui con una storia criminosa simile siano più o meno propensi a commettere un nuovo reato una volta tornati in libertà. L'aspetto da tenere in considerazione è che COMPAS non prevede il rischio di recidiva individuale dell'imputato, bensì elabora la previsione comparando le informazioni ottenute dal singolo con quelle relative ad un gruppo di individui con caratteristiche assimilabili. La descrizione di COMPAS e della vicenda giudiziaria del Sig. Loomis è efficacemente riassunta in: www.giurisprudenzapenale.com.

politiche. Tra queste iniziative spiccano, relativamente all'ambito europeo, l'AI Act e il già citato GDPR.

Quanto al primo, trattasi di un Regolamento volto a stabilire regole armonizzate sull'intelligenza artificiale attraverso un approccio definito dalle stesse istituzioni europee "*risk based*" (Ebers, 2024), secondo una c.d. piramide del rischio che lascia molte applicazioni di IA non regolate perché considerate a rischio minimo, alcune disciplinate con limitati requisiti richiesti *ex lege*, poche ad alto rischio soggette alla valutazione di conformità e pochissime del tutto vietate[14]. Sostanzialmente trattasi di una legislazione dall'ossatura tipica delle regolamentazioni industriali e di sicurezza dei prodotti (Mantelero, 2024; Ződi, 2024), pur con una attenzione particolare alla tutela dei diritti fondamentali. Il Regolamento disciplina in modo preponderante gli usi dell'IA connotati da alto rischio, optando per un approccio tipizzante dei diversi sistemi basato sull'elenco di cui all'Allegato III. L'obiettivo del legislatore è quello di indicare in modo chiaro quali applicazioni vengono considerate particolarmente rischiose, anche allo scopo di consentire agli operatori del settore di capire se i prodotti che commercializzano sono o meno inclusi in tale normativa. Tuttavia, questa scelta - tenuto conto degli ampi e rapidi margini di evoluzione della tecnologia - rischia di essere, nel lungo periodo, fonte di confusione e, quindi, di contenzioso (Mantelero, 2024).

[14] Possono, ad esempio, considerarsi ad alto rischio le applicazioni di IA utilizzate in: infrastrutture critiche (ad esempio i trasporti), che potrebbero mettere a rischio la vita e la salute dei cittadini; istruzione o formazione professionale, che può determinare l'accesso all'istruzione e al corso professionale della vita di qualcuno (ad esempio, il punteggio degli esami); componenti di sicurezza dei prodotti (ad esempio applicazione di IA in chirurgia robotizzata); occupazione, gestione dei lavoratori e accesso al lavoro autonomo (ad esempio software di selezione dei CV per le procedure di assunzione); servizi privati e pubblici essenziali (ad esempio, crediti che negano ai cittadini l'opportunità di ottenere un prestito); applicazione della legge che può interferire con i diritti fondamentali delle persone (ad esempio valutazione dell'affidabilità delle prove); gestione della migrazione, dell'asilo e del controllo delle frontiere (ad es. esame automatizzato delle domande di visto); amministrazione della giustizia e processi democratici (ad esempio soluzioni di IA per la ricerca di sentenze giudiziarie). L'esemplificazione è consultabile sul sito della Commissione europea, alla comunicazione dal titolo "Plasmare il futuro digitale dell'Europa - Legge sull'IA": https://digital-strategy.ec.europa.eu/it/policies/regulatory-framework-ai.

Quello che infatti ad oggi è un sistema a rischio moderato ben potrebbe, in futuro, evolvere ed essere classificato come IA ad alto rischio. Ciò spiega la scelta del legislatore europeo di consentire interventi correttivi, quali possibilità di deroghe e di modifiche future all'Allegato III.

Con specifico riguardo all'opacità il considerando 72 sottolinea l'opportunità di "imporre la trasparenza per i sistemi di IA ad alto rischio prima che siano immessi sul mercato o messi in servizio". Come ciò debba avvenire lo illustrano gli artt. 13 e 50 del Regolamento, i quali dettano discipline parzialmente diverse in tema di trasparenza dell'IA. L'art. 13, infatti, impone una serie di requisiti di trasparenza per i sistemi di IA ad alto rischio, unitamente all'obbligo per il *provider* (nella versione italiana "fornitore") di redigere dettagliate istruzioni per l'uso rivolte ai *deployers*[15] (così definiti nella versione italiana, in quella inglese si parla di "users"). L'art. 50, invece, detta obblighi di trasparenza per i fornitori e gli utenti di "determinati sistemi di IA" (tra cui sistemi per finalità generali), imponendo che le persone fisiche interessate siano informate del fatto di stare interagendo con un sistema di IA o che un determinato contenuto è stato generato dall'IA.

Sul piano operativo, tuttavia, l'art. 13 non fornisce indicazioni circa il grado di dettaglio delle informazioni che debbono essere fornite, limitandosi ad una previsione generale e raccomandando che il *deployer* sia destinatario di informazioni "concise, complete, corrette e chiare che siano pertinenti, accessibili e comprensibili". Cosa questo significhi in concreto è difficile comprenderlo, specialmente se si considera che l'informazione - concisa ma completa - dovrebbe avere anche riguardo alle specifiche dei dati di input e di addestramento, nonché ai livelli di sicurezza informatica, accuratezza e robustezza dell'IA (comprese le rispettive metriche e le circostanze note e prevedibili che possono

[15] Per *deployer* si intende: "qualsiasi persona fisica o giuridica, autorità pubblica, agenzia o altro organismo che utilizza un sistema di AI sotto la propria autorità, eccetto quando il sistema di AI è utilizzato nell'ambito di un'attività personale non professionale" (art. 3, punto 4). L'espressione potrebbe, forse, tradursi come "gestore".

influire su tali livelli). La previsione non consente di percepire quale sia l'estensione del dovere informativo previsto dall'art. 13, anche tenuto conto della circostanza - già da tempo nota nell'ambito medico in relazione al consenso informato - per cui non è attraverso una bulimia informativa che si persegue il risultato di una efficace spiegabilità dei processi. Peraltro, la disposizione nulla prevede circa l'eventuale necessità di aggiornamento delle istruzioni relative al sistema, per esempio in caso di modifica della destinazione d'uso o di evoluzione delle capacità del sistema.

Quanto all'art. 50, è probabile che ai fini di una efficace attuazione della norma saranno necessarie esperienze e conoscenze multidisciplinari in vari settori. I requisiti previsti, infatti, richiedono non solo l'apporto di professionisti tecnici e legali, ma anche le competenze di soggetti con esperienza nella progettazione dell'interazione con l'utente e nell'ambito del design accessibile/inclusivo. Questi ultimi profili sono necessari per garantire che le informazioni siano trasmesse in modo coinvolgente e accessibile a tutti i possibili destinatari, requisito che, a sua volta, richiede trasparenza multimodale (ad es. attraverso mezzi acustici e visivi). Peraltro, anche rispetto a tale disposizione si pone la questione dell'estensione del dovere informativo già accennata rispetto all'art. 13. Invero, non sembra semplice identificare la mole adeguata di informazioni da fornire agli utenti allo scopo di informarli correttamente circa l'interazione con un chatbot o la natura artificiale di un contenuto, senza sopraffarli e ingenerare una sfiducia verso l'uso di questi sistemi.

Che i requisiti previsti dagli artt. 13 e 50 debbano trovare una declinazione concreta è riflessione condivisa dallo stesso legislatore che, infatti, delega ad una apposita Commissione (art. 96) l'attuazione pratica - tra gli altri - degli obblighi di trasparenza sopra citati.

Merita, inoltre, particolare attenzione l'intersezione tra la già citata normativa di cui al GDPR e la disciplina prevista dall'AI Act. Nonostante ancor prima della sua approvazione fosse a più

riprese stata raccomandata la necessità di una coerenza tra le due regolamentazioni, al fine di evitare conflitti, non pare che l'AI Act abbia tenuto la questione in grande considerazione.

Ciò emerge, in particolare, in relazione agli obblighi di trasparenza, che vengono imposti da entrambe le discipline ma il cui ambito e i cui requisiti sono regolati in modo diverso. Il GDPR, infatti, stabilisce il principio di trasparenza per facilitare l'esercizio dei diritti degli interessati ai sensi degli artt. 15-22, compreso il diritto alla cancellazione, alla rettifica e alla portabilità dei dati. Al contrario, l'AI Act prevede obblighi di trasparenza solo per i sistemi ad alto rischio (art. 13) e per altri determinati sistemi di IA (art. 50). Inoltre, l'art. 13 si concentra esclusivamente sugli interessi del *deployer* di un sistema di IA piuttosto che sull'utente finale e/o sull'interessato, che è invece il focus del GDPR.

Un altro esempio della non felice intersezione tra le due normative è rappresentato dal diritto alla spiegazione e all'intervento/sorveglianza umana. Mentre il GDPR richiede l'intervento umano (art. 22, par. 3) e sancisce il diritto dell'utente a ricevere informazioni significative sulla logica seguita dal sistema per le decisioni basate esclusivamente sul trattamento automatizzato, compresa la profilazione (art. 15, par. 1 lett. h), l'AI Act richiede la supervisione umana (art. 14) e il diritto alla spiegazione del processo decisionale individuale (art. 86) solo per i sistemi di IA ad alto rischio, in cui sia il contenuto che i prerequisiti e le conseguenze legali sono regolati in modo completamente diverso.

Molto probabilmente sarà la prassi successiva all'entrata in vigore del Regolamento, oppure l'interpretazione della giurisprudenza, a sciogliere alcuni dei nodi relativi all'applicazione delle due normative e alle loro parziali sovrapposizioni. Sarà, infatti, cruciale evitare il sorgere di interpretazioni potenzialmente confliggenti in relazione ai principi applicabili e agli obblighi previsti in capo ai fornitori dei sistemi di IA in base alle due normative, per evitare la diffusione di uno stato di incertezza del quadro giuridico e di disincentivo agli investimenti nel settore.

Infine, uno sguardo più ampio alla regolamentazione dell'IA nel suo complesso consente di osservare che l'efficacia del quadro normativo dettato dall'AI Act avrebbe potuto essere più significativa se, unitamente all'approvazione del Regolamento in materia di Intelligenza artificiale, fosse stato adottato il "pilastro complementare all'AI Act" (Mantelero, 2024), ovvero la direttiva sulla responsabilità civile correlata all'uso dell'IA.

Come già accennato, infatti, l'AI Act è pensato per disciplinare *ex ante* lo sviluppo dei sistemi di IA, garantendo il rispetto *by design* dei principi di dignità umana e autonomia individuale; trasparenza e sorveglianza umana; *accountability* e responsabilità; uguaglianza e non discriminazione; *privacy* e protezione dei dati; affidabilità e sicurezza nell'innovazione.

Tuttavia, la mancanza di regole che chiudano il sistema, approntando rimedi e sanzioni per le ipotesi in cui si verifichino danni a causa dell'IA è il segnale della scarsa coordinazione dell'approccio europeo che non ha saputo combinare la gestione del rischio, con le relative sanzioni in caso di difformità e la responsabilità civile per i danni causati dall'IA. L'impressione di un quadro disarmonico è, peraltro, accentuata dalla scelta di strumenti legislativi diversi per regolare lo stesso fenomeno, laddove l'AI Act possiede la forma del Regolamento (in quanto tale direttamente applicabile negli Stati membri dell'UE), mentre la proposta in materia di responsabilità civile ha la forma della direttiva (applicabile solo se ratificata da ciascuno Stato e nei limiti della legge di ratifica).

A complicare ulteriormente il già elaborato impianto va menzionato il parallelo iter - attualmente in corso - di revisione della disciplina in materia di prodotti difettosi, allo scopo di aggiornarla ai problemi sollevati dalle nuove tecnologie, tra cui l'IA. L'obiettivo sarebbe quello di rinnovare la normativa risalente al 1985 in materia di responsabilità del produttore adattandola alle novità introdotte dall'IA, adeguando, ad esempio, la nozione di prodotto e di difetto così da includervi anche i sistemi algoritmici, nonché circoscrivendo in modo più preciso i casi di esonero da responsabilità del produttore, per evitare che caratteristiche

come l'opacità e l'imprevedibilità degli output si rivelino per il fabbricante facili "vie di fuga" dalla responsabilità civile e dall'obbligo di risarcire il danno cagionato dall'IA. Oppure, ancora, responsabilizzando i diversi attori coinvolti nella catena di sviluppo e produzione di questi sistemi e riconoscendo un diritto di *disclosure* al danneggiato dal sistema, affinché possa avere accesso agli elementi di prova utili per ottenere il risarcimento del pregiudizio subito.

In questo quadro assai articolato fa, infine, capolino il legislatore italiano, con il recente disegno di legge n. 1146 presentato il 20 maggio 2024 dalla Presidente del Consiglio dei Ministri e dal Ministro della giustizia e attualmente in discussione in Senato. L'obiettivo del d.d.l. "è quello di dettare una normativa nazionale che senza sovrapporsi al regolamento UE predisponga un sistema di principi, governance e misure specifiche adatte al contesto italiano per cogliere tutte le opportunità dell'intelligenza artificiale".

Sembra, ora, prematuro muovere osservazioni su una disciplina ancora in aperta discussione e che, molto probabilmente, sarà oggetto di numerosi emendamenti. Dalla lettura del testo, tuttavia, emerge una sostanziale omogeneità sul piano dei principi con l'impianto previsto dall'AI Act. Il d.d.l., nei suoi 26 articoli, individua nell'Agenzia per il digitale (Agid) e in quella per la cybersecurity (Acn) le competenti autorità di sorveglianza in materia e ambisce a rappresentare un compendio di principi: dalla Pubblica Amministrazione alla sanità, dal lavoro nelle fabbriche alle professioni intellettuali. Trova, inoltre, spazio la previsione di severe sanzioni penali, quali la reclusione fino a 5 anni per danni causati dalla diffusione di contenuti generati o manipolati con l'IA. L'estrema trasversalità della materia comporterà, probabilmente, un gran numero di audizioni prima della presentazione e della discussione degli emendamenti. È presumibile che uno degli aspetti più controversi sarà quello relativo allo stanziamento di risorse da parte dello Stato, limitato, per ora, ad una dote di 1 miliardo (art. 21) affidata alla gestione di Cdp Venture Capital, il fondo nazionale per le startup.

L'investimento, infatti, si attesta al di sotto di quanto stanziato dai principali partner europei come Spagna, Francia, Germania e Regno Unito (Fotina, 2024).

Quanto agli aspetti contenutistici, rispetto ai suggerimenti proposti dai documenti programmatici elaborati dagli esperti incaricati dalla Presidenza del Consiglio (Prof. Greco e Prof. Benanti), non ha trovato seguito l'idea di costituire una Fondazione con responsabilità delle varie iniziative nel settore. Allo stesso modo le previsioni dedicate all'ambito scolastico, universitario e professionale, sviluppate in modo più esteso nei documenti programmatici, nel d.d.l. trovano spazio solo come indicazione di principi di massima sotto forma di delega al governo, la cui applicazione concreta spetterà, quindi, ai decreti legislativi di attuazione. Quanto all'opacità dei sistemi di IA non si rintracciano previsioni *ad hoc*, se non qualche richiamo alla necessità di garantire il rispetto del principio di trasparenza dei processi e dei dati utilizzati dall'IA.

Bibliografia

Casonato, C. (2019) Costituzione e intelligenza artificiale: un'agenda per il prossimo futuro. *BioLaw Journal*, 711 ss.

Cerea, F. (2023) Trattamento algoritmico dei dati a fini reputazionali tra consenso dell'interessato e controllo ex ante di conformità al GDPR e all'AI Act. *Resp. civ. e prev.*, 2005-2020.

Chang, X. (2023). Gender *Bias* in hiring: An analysis of the impact of Amazon's recruiting algorithm. *Advances in Economics, Management and Political Sciences*, *23*(1), 134–140.

Cranmer, M. (2023, May 2). *Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl*. arXiv.org. https://arxiv.org/abs/2305.01582.

Criado, N., & Such, J. M. (2019). Digital discrimination. In *Oxford University Press eBooks* (pp. 82–97).

Dalmeet Singh Chawla. (2024). Is ChatGPT corrupting peer review? Telltale words hint at AI use. *Nature*, *628*(8008), 483–484.

Dustin, J. (2022). Amazon scraps secret AI recruiting tool that showed bias against women. In Martin, Kirsten (ed.), *Ethics of Data and Analytics. Concepts and Cases*. Auerbach Publications.

Ebers, M. (2019). Chapter 2: Regulating AI and Robotics: Ethical and Legal challenges. *Social Science Research Network*.

Ebers, M. (2024) *Truly Risk-Based Regulation of Artificial Intelligence - How to Implement the EU's AI Act*. https://dx.doi.org/10.2139/ssrn.4870387.

Fasan, M. (2022). I principi costituzionali nella disciplina dell'Intelligenza Artificiale. Nuove prospettive interpretative. *DPCE online*. 51(1). 181-199.

Fotina, C. (2024, June 30) Intelligenza artificiale, la legge arriva in Parlamento: ora è caccia alle risorse. *Il Sole 24 Ore*.

Leveson, N., & Turner, C. (1993). An investigation of the Therac-25 accidents. *Computer*, 26(7), 18–41. https://doi.org/10.1109/mc.1993.274940.

Mantelero A. (2024) L'AI Act: la risposta del legislatore europeo alle sfide dell'intelligenza artificiale. *Accademia*, 191-206.

Micklitz, H., Pollicino, O., Reichman, A., Simoncini, A., Sartor, G., & De Gregorio, G. (2024). *Constitutional challenges in the algorithmic society*. Cambridge University Press.

Orsoni, G., D'Orlando, E. (2019). Nuove prospettive dell'amministrazione digitale: open data e algoritmi. *Istituzioni del Federalismo*, 593-617.

Pajno, A., Bassini, M., De Gregorio, G., Macchia, M., Patti, F.P., Pollicino, O., Quattrocolo, S., Simeoli, D., & Sirena, P. (2019). AI: profili giuridici. Intelligenza Artificiale: criticità emergenti e sfide per il giurista. *BioLaw Journal*, 205-235.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press.

Sarker, M. K., Zhou, L., Eberhart, A. & Hitzler, P. (2021). *Neuro-symbolic artificial intelligence: Current trends*.

Sarker, M. K., Zhou, L., Eberhart, A., & Hitzler, P. (2022). Neuro-symbolic artificial intelligence. *AI Communications*, *34*(3), 197–209. https://doi.org/10.3233/aic-210084.

Simoncini, A. (2019). L'algoritmo incostituzionale: intelligenza artificiale e il futuro delle libertà. *BioLaw Journal*, 63-89.

Sohn, E. (2023). The reproducibility issues that haunt health-care AI. *Nature*, *613*(7943), 402–403.

Stradella, E., (2020). Stereotipi e discriminazioni: dall'intelligenza umana all'intelligenza artificiale. Consulta Online. *Liber Amicorum per Pasquale Costanzo – Diritto costituzionale in trasformazione*. Vol. I – Costituzionalismo, Reti e Intelligenza artificiale. 391-40.

Van Noorden, R. (2022). The researchers using AI to analyse peer review. *Nature*, *609*(7927), 455–455.

Ződi, Z. (2024, June 4). The EU AI Act – Can We Protect Human Rights with a Product Compliance Regulation?. *IACL-AIDC Blog*.

Capitolo III — Generatività

Giuseppe Previtali

La diffusione sempre più massiccia di sistemi di AI generativa è un fenomeno tecno-sociale che caratterizza in modo forte il nostro presente; pur non trattandosi dell'unico tipo di intelligenza artificiale disponibile, quello generativo è senza dubbio il tipo di sistema che ha avuto un impatto maggiore sull'immaginario culturale, quello a cui più spesso si pensa quando - anche a livello di senso comune - ci si interroga sul tema. L'idea che l'intelligenza artificiale possa dotarsi di una forma di autonomia creativa è sempre stata al centro dei timori culturali legati alla vita inanimata (come attestano anche numerosi film, da 2001: *Odissea nello spazio* a *Matrix* soltanto per citare alcuni dei più noti e influenti). Quello che lì appare come un'estremizzazione di un processo ancora ampiamente immaginato, è già però in qualche misura presente: oggi ChatGPT è in grado di rispondere ai nostri quesiti con un certo grado di precisione (anche se spesso gli aspetti più interessanti del fenomeno si annidano nei suoi errori, le cosiddette *hallucinations*, la cui esplorazione è di interesse in vari campi del sapere; cfr. Athaluri et al. 2023; Christensen et al. 2023) e la sua capacità (così come quella di altri sistemi, p.e. Midjourney) di generare immagini o altri contenuti in risposta a determinati prompt è senza dubbio uno degli aspetti più capaci di generare discorso nel presente.

Non si tratta soltanto di un divertissement: in un contesto politicamente delicato come quello in cui stiamo vivendo, diventa particolarmente rilevante saper discriminare gli *AI-generated content* capaci di diffondere messaggi problematici e dalle conseguenze imprevedibili. Come osservato da più parti (p.e. Galston, 2020), è l'intero rapporto fra percezione sensoriale e credenza che rischia di venire sovvertito in questo contesto, anche considerando che la capacità di questi messaggi di

incidere da un punto di vista politico è ancora in larga parte non mappata (Vaccari, 2020). Il tema della generatività dell'AI è centrale nell'agenda di questioni che questa innovazione ci impone di affrontare, perché oltre all'ampiezza della sua discorsività sociale, essa mobilita questioni che sono sia di natura etico-giuridica che estetico-mediologica.

Se il secondo ramo del problema verrà in qualche modo sviluppato di seguito, rimane necessario precisare come anche da un punto di vista della tutela giuridica la capacità dell'AI di creare nuovo contenuto causi più di una impasse. Il tema del diritto d'autore (in Italia regolato dalla legge 633/1941) tende a chiamare in causa la questione della creatività, intesa come co-presenza di due condizioni: l'originalità e la novità. Se in alcuni casi è ancora relativamente semplice discernere la responsabilità creativa di un prodotto, in altri casi di co-creazione fra uomo e algoritmo (cfr. le recenti osservazioni di Cizek e Uricchio, 2022) è problematico discriminare l'istanza autoriale, proprio in virtù delle sempre più stringenti (e spesso inestricabili) connessioni fra i vari enti implicati.

Parte di questa difficoltà che, come si intuisce, ha importanti implicazioni in termini di tutela legale, ha a che vedere con la connaturata opacità dei sistemi di AI, il cui funzionamento spesso non è solo difficilmente comprensibile, ma anche imprevedibile e capace di adattamento. La cogenza di questo punto è tale che ci si può legittimamente chiedere se questa incapacità di aver ragione dei meccanismi profondi dei sistemi di intelligenza artificiale non possa che frustrare la nostra capacità di normarli; più ancora, se riconosciamo che gli algoritmi sono in larga parte un prodotto dell'autorialità umana (che rischiano di trattenerne i bias e di riprodurre digitalmente forme di subalternità; cfr. O'Neil, 2016; Marino, 2020), è legittimo chiedersi se forse non sia anche chi ha scritto (e magari addestrato) un certo sistema di AI a dover condividere la responsabilità creativa delle sue azioni.

In questa triangolazione fra ruolo dell'utente (che con i suoi prompt genera una risposta nell'AI), programmatore (che in qualche modo rende possibile l'uso dell'AI) e intelligenza artificiale

(che di fatto produce, a partire da un comando, un'elaborazione più o meno creativa), il ruolo della regolamentazione diventa cruciale, come attestano le diverse iniziative europee in questo senso. Mentre l'implementazione dell'AI Act sta prendendo forma, la Direttiva 2019/790 stabilisce che venga introdotta una eccezione al diritto d'autore «per le riproduzioni e le estrazioni effettuate da opere o altri materiali cui si abbia legalmente accesso ai fini dell'estrazione di testo e di dati», cui ha fatto seguito un secondo intervento dell'ottobre 2020.

1. Industria culturale e creatività

Il 2023 è stato l'anno in cui la WGA – *Writer Guild of America* (associazione di categoria degli sceneggiatori cinematografici e televisivi) ha indetto un ampio sciopero, la cui portata è stata tale da paralizzare per mesi l'intero sistema produttivo hollywoodiano. Non è certamente la prima volta che una manifestazione di questo genere ha luogo (un importante precedente è stato quello del 2007/2008), ma in questo caso l'elemento di novità è senza dubbio rappresentato dal fatto che, fra le richieste del sindacato, ci fosse anche la regolamentazione del ruolo che le IA giocano (o potrebbero giocare) all'interno delle industrie creative. La cogenza di questa mobilitazione è stata tale da produrre una inedita alleanza fra diverse associazioni (oltre alla WGA si sono unite allo sciopero anche la *Screen Actors Guild* e la *American Federation of Television and Radio Artists*) per regolamentare l'uso dei sistemi di intelligenza artificiale.

Uno degli aspetti cruciali della vicenda ha in effetti a che vedere con l'idea che l'utilizzo dell'IA avrebbe potuto nel tempo marginalizzare o espungere completamente l'apporto creativo degli sceneggiatori; in questo senso, la loro mobilitazione potrebbe forse essere considerata solo una manifestazione particolare del più ampio problema del rapporto fra IA e lavoro e della riconfigurazione delle competenze. Nel caso specifico, il rischio percepito era quello di una sottrazione della paternità creativa dei contenuti da parte di IA in grado di produrre script assolutamente convincenti a partire dalla presa di coscienza che

i testi adoperati per addestrarle sono spesso proprio prodotti dell'ingegno umano come film, sceneggiature e romanzi. L'accordo raggiunto lo scorso autunno sancisce in qualche modo il riconoscimento della posizione degli sceneggiatori e crea in questo senso un precedente importante:

Under the new terms, studios "cannot use AI to write scripts or to edit scripts that have already been written by a writer" […]. The contract also prevents studios from treating AI-generated content as "source material", like a novel or a stage play, that screenwriters could be assigned to adapt for a lower fee and less credit than a fully original script (Anguiano e Beckett, 2023).

L'esito della mobilitazione ha così di fatto riconsegnato in mano agli sceneggiatori l'uso dell'AI come supporto al lavoro creativo, sottraendolo agli studios. È peraltro interessante notare come questa tensione fra le varie componenti del processo produttivo di un film sia in qualche modo un tratto ricorrente nella storia di Hollywood, dove sin dai tempi della golden age del cinema classico i produttori hanno cercato di imporre le proprie ragioni "aziendali" su quelle di altre figure, come il regista o – appunto – lo sceneggiatore.

Il caso del cinema è un caso particolarmente interessante di un problema più ampio, cioè quello del rapporto fra intelligenza artificiale e creatività all'interno delle industrie culturali, in una prospettiva ampia che spazia dalla letteratura al design. Al di là dei rischi e degli allarmi mediatici, infatti, l'uso dell'AI in questi ambiti può offrire una serie di opportunità interessanti, capaci di promuovere un pensiero divergente, facilitare l'individuazione di pattern e connessioni e offrire un valido supporto alla creatività umana: «il maggior potenziale dell'AI generativa non sta nel sostituire gli umani, ma nell'assisterli nel loro sforzo di creare soluzioni sinora inimmaginabili» (Eapen, Finkenstadt, Folk & Venkataswamy, 2023).

Se questo è certamente vero, la popolarità e le capacità dimostrate dai sistemi di AI generativa, sembrano tuttavia

richiedere in senso più generale una revisione e un aggiornamento del concetto stesso di creatività (cfr. Miller, 2019); la creazione di nuovi prodotti a partire dalla conoscenza e dalla capacità di manipolare linguaggi espressivi è infatti fortemente influenzata dall'implementazione dell'intelligenza artificiale, che però offre anche ad artisti e creativi di ogni sorta la possibilità di concentrare la propria attenzione sull'immaginazione di nuovi possibili: «le capacità di elaborazione dati dell'AI arricchiscono la dimensione esplorativa della creatività, consentendo agli artisti di sperimentare nuove possibilità estetiche", confermandosi in questo senso un collaborative tool di straordinaria efficacia» (Hutson, 2023a).

Il dibattito sulla natura effettivamente creativa di quella che oggi chiamiamo *AI-Generated Art* (Zylinska, 2020) è, come già anticipato, oggi molto acceso e polarizzato fra chi la ritiene una forma interessante e in determinati contesti utile di esplorazione estetica e chi si interroga sul destino della creatività (Hutson, 2023b). Un esempio celebre in questo senso è quello offerto dall'artista visivo Mario Klingemann, che ha sviluppato la tecnica del neural glicth proprio esplorando le possibilità dell'AI. Addestrano una GAN e modificandone poi la struttra, Klingemann è stato in grado di produrre immagini "glitchate" che, se da una parte sembrano produrre uno stile coerente che rielabora la pittura d'avanguardia (come quella surrealista o di Francis Bacon) si basano in realtà su una peculiare forma di riappropriazione delle reti generative, che l'artista utilizza come vero e proprio mezzo espressivo (cfr. Klingemann, 2018). Questo caso è interessante proprio perché dimostra come l'utilizzo dell'intelligenza artificiale non sia necessariamente legato alla replicazione di forme espressive esistenti, ma possa diventare anche uno strumento artistico di per sé (si veda anche il caso recente della poesia, cfr. Bernstein & Balula, 2023).

2. Fotografie AI-Generated

Uno degli aspetti che più sembrano condizionare l'immaginario culturale in relazione al tema della generatività dell'AI, alla sua capacità di produrre immagini e contenuti che mimano la creatività umana, è quello della definizione di originalità e del suo impatto sul modo in cui siamo abituati a considerare i prodotti mediali.

Prendiamo il caso di una fotografia generata dall'AI: la capacità dei sistemi di intelligenza artificiale di offrire, in risposta a determinati prompt, risultati perfettamente capaci di mimare uno scatto vero e proprio è ormai evidente, come hanno dimostrato per esempio le immagini generate con Midjourney dei funerali di Silvio Berlusconi nell'aprile 2023, che colpiscono per la perturbante capacità di preconizzare in modo del tutto verosimile un evento all'epoca non ancora avvenuto (Deragni, 2023). Un altro caso ormai noto e interessante da questo punto di vista è offerto dal sito This Person Does Not Exist[1] rilasciato nel 2018 da Nvidia e capace di generare in autonomia, ad ogni refresh, un nuovo volto fotografico di un individuo inesistente (sfruttando una GAN). Anche se in alcuni casi la *fakeness* di queste immagini è facilmente determinabile, il processo di individuazione del contenuto *AI-generated* non è sempre semplice.

Il fascino generato da queste immagini deriva probabilmente anche dal fatto che sembrano rimettere in questione il nostro rapporto con la fotografia, che da sempre è il medium cui siamo più abituati ad associare un valore certificante in virtù del suo essere la prima forma meccanica di image-making. La teoria del fotografico ci ha insegnato che la capacità attestante/certificante che il visivo raggiunge con la fotografia una nuova centralità, proprio perché l'inedita possibilità che offre di produrre un'immagine obiettiva del passato ne costituirebbe il vero e proprio fondamento identitario, quando non la stessa ontologia (cfr. Bazin, 1999; Barthes, 2003). Da questo punto di vista, non stupisce la vasta discorsività sociale che l'AI ha

[1] https://this-person-does-not-exist.com/en.

contribuito a generare su questi temi, perché per la prima volta sembriamo trovarci di fronte ad immagini prive di referente e che quindi rappresentano in un modo nuovo.

Oltre ad un evidente interesse teorico, la pratica ha importanti implicazioni socio-politiche: i cosiddetti deepfake (immagini di sintesi generate grazie a sistemi di AI; cfr. Westerlund, 2019) sono oggi diffusi tanto nella comunicazione politica quanto – e si tratta di un tema al contempo urgente e poco mappato – nella pornografia (Kerner e Risse, 2021). Oltre alle implicazioni legali (p.e. tutela della privacy e rispetto del diritto all'immagine dei soggetti rappresentati) e alla necessità di sviluppare sistemi di riconoscimento che siano al contempo funzionali e aggiornabili (si veda l'idea di una forma di watermarking elaborata in seno all'AI Act), è anche il concetto culturale di autenticità a dover essere ripensato al tempo dell'AI, perché le sollecitazioni offerte da questi sistemi spingono a un aggiornamento delle nostre categorie (si è parlato, non a caso, del passaggio dalla dimensione del fotografico a quella del postfotografico; cfr. Fonctuberta, 2018). Le immagini algoritmiche (Eugeni, 2021) impongono in questo senso un vero e proprio cambio di paradigma nella nostra percezione culturale, perché non sono più l'emanazione di una realtà data, ma la visualizzazione dell'articolazione di un determinato set di dati.

3. L'IA nel sistema dei media

È emblematico che il dibattito sull'uso dei sistemi di intelligenza artificiali all'interno delle pratiche sociali legate all'*image-making* si sia innescato (o comunque sia divenuto produttore di grande discorsività culturale) a partire dalla nuova possibilità che alcuni cellulari permettevano di fotografare la Luna. Il gesto è di per sé affascinante e da sempre il tentativo di mettere in immagine il nostro satellite sembra ossessionare tanto gli uomini di scienza quanto gli artisti. Gli smartphone Xiaomi Mi 9 e Xiaomi Mi 9 SE promettevano ai potenziali clienti di poterlo fare meglio proprio grazie all'introduzione di un'apposita *Moon Mode*. La straordinaria qualità dei risultati ottenuti ha indotto a ritenere

che in questa nuova procedura fosse implicata l'intelligenza artificiale e anche Samsung ha dovuto difendersi da accuse di questo genere: le fotografie prodotte sfruttando queste modalità sarebbero in qualche modo *enhanced* da algoritmi che, mettendo a confronto la fotografia scattata con un *dataset* di riferimento composto di fotografie della luna ad alta definizione, permetterebbero di abbellire l'output della fotocamera.

La diatriba su questo tema rappresenta forse soltanto una nuova iterazione del problema della fraudolenza dell'immagine che accompagna la fotografia sin dalla sua nascita. Già nel caso dello spiritismo ottocentesco (ma il tema è emerso anche in relazione a celebri fotografie di guerra o legate al desiderio di screditare avversari politici; cfr. Ritchin, 2012) si sottolineava la scorrettezza di qualunque pratica che alterasse l'immagine, rompendo quel patto di indessicalità con il suo oggetto. Se la questione è quindi vecchia quanto le pratiche di produzione meccanica delle immagini, è indubbio che la nuova plasticità del visivo garantita anche dall'implementazione sempre più diffusa dei sistemi di AI – della sua *morbidezza* hanno parlato giustamente Hoelzl e Marie (2015) – complichi di molto la relazione con il referente; nel farlo, tuttavia, apre anche numerosi nuovi spazi di azione ed è forse il caso di interrogarsi sulle *funzioni* che l'IA svolge in questa nuova economia delle immagini (cfr. Eugeni, 2024).

Quello del *miglioramento* tramite applicazioni di IA è senza dubbio il caso più comune. AppStore e GooglePlay pullulano di applicazioni gratuite e/o a pagamento (più spesso in una soluzione ibrida) capaci di perfezionare l'aspetto di una fotografia, sia attraverso una procedura di bilanciamento automatico dei suoi valori (automatizzando in questo caso un processo che spesso continua ad essere fatto direttamente dall'utente), sia nel senso di un vero e proprio restauro. Phot.ai, una delle applicazioni più sponsorizzate in questo ambito, si presenta all'utente con questo *claim*:

Reclaim the true quality and details of your cherished photographs. Our AI-powered image restoration process identifies the stains, scratches, blurs and other imperfections on your old faded memories and puts a new life in them. Further it improves the sharpness, saturation level of your photos and restores them to their past glory."[2]

L'idea che l'intelligenza artificiale sia in grado di riportare alla luce volti che credevamo perduti è certamente affascinante perché tocca da vicino il nostro rapporto affettivo con le fotografie; in tutti i casi di *enhancement* dell'immagine fotografica tramite AI sembriamo trovarci di fronte all'idea di un *servizio* che ci rassicura.

AI contrario, tanto la capacità degli algoritmi di *riconoscere e classificare* immagini (pratica estetico-politica la cui decostruzione è oggi al centro di un ampio dibattito, cfr. Oswald, 2024) quanto la loro possibilità di produrre immagini di sintesi, dai già citati *deepfake* alla diffusione dei filtri nella sfera social, sembrano produrre nei soggetti un ventaglio di reazioni più variegato e preoccupante, sia per le ragioni già menzionate (espunzione della voce autoriale umana, difficoltà di controllo sulla circolazione con conseguente messa in crisi del concetto di privacy etc.), sia per ragioni più strettamente teoriche, che hanno a che vedere con la risposta estetica a questo tipo di contenuti. Soprattutto quando si parla di immagini, la tentazione di vedere in algoritmi ed AI la manifestazione di un sublime tecnico è certamente diffusa, ma è rischioso dimenticarsi della natura sempre situata e culturale di questi processi (cfr. Ames, 2018). Uno dei compiti che le discipline umanistico-sociali sono in questo senso chiamate ad affrontare è dunque proprio quello di elaborare nuovi quadri teorici per rendere conto della complessità dei fattori in gioco nei processi di AI generativa senza dimenticare di analizzare le stratificazioni di senso e le connessioni con la

[2] https://www.phot.ai/ai-photo-restoration

politica, soprattutto in uno scenario di conflitti diffusi come quello in cui ci troviamo a vivere.

Bibliografia

Ames M.G. (2018). Deconstructing the Algorithmic Sublime. *Big Data & Society*, *5*(1), 1-4.

Anguiano D., Beckett L. (2023). How Hollywood writers triumphed over AI – and why it matters. *The Guardian*, https://www.theguardian.com/culture/2023/oct/01/hollywood-writers-strike-artificial-intelligence.

Athaluri S.A. et alii (2023). Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. *Cureus*, 15(4).

Barthes R. (2003). *La camera chiara. Nota sulla fotografia*. Einaudi.

Bazin A. (1999). Ontologia dell'immagine fotografica. in Id., *Che cos'è il cinema?* (pp. 3-10). Garzanti.

Bernstein C., Balula D. (2023). *Poetry Has No Future, Unless it Comes to an End: Poems of Artificial Intelligence*. NERO.

Capparelli V.M. (2020). Le nuove frontiere del diritto d'autore alla prova dell'Intelligenza Artificiale. In U. Ruffolo (a cura di), *Intelligenza Artificiale - Il diritto, i diritti, l'etica*, (335-343). Giuffrè.

Christensen J. et alii (2023). Understanding the role and impact of Generative Artificial Intelligence (AI) hallucination within

consumers' tourism decision-making processes. *Current Issues in Tourism*, 1-16.

Cizek K., Uricchio W. (2022). *Collective Winsdom: Co-Creating Media for Equity and Justice*. MIT Press.

Deragni P. (2023) Le immagini del funerale di Berlusconi create da Midjourney sono di un realismo inquietante. *Wired*. https://www.wired.it/gallery/berlusconi-funerale-milano-midjourney-intelligenza-artificiale-riccio/

Eapen T.T., Finkenstadt D.J., Folk J., Venkataswamy L. (2023). How Generative AI Can Augment Human Creativity. *Harvard Business Review*, 4 https://hbr.org/2023/07/how-generative-ai-can-augment-human-creativity

Eugeni R. (2021). *Capitale algoritmico. Cinque dispositivi postmediali (più uno).* Morcelliana.

Eugeni R. (2024). Algoritmi. In B. Grespi, F. Villa (a cura di), *Il postfotografico. Dal selfie alla fotogrammetria digitale*, (pp. 36-52). Einaudi.

Fontcuberta J. (2018). *La furia delle immagini. Note sulla postfotografia*. Einaudi.

Galston, W. (2020). Is seeing still believing? The deepfake challenge to truth in politics. *Brookings Institution*. https://policycommons.net/artifacts/4141603/is-seeing-still-believing-the-deepfake-challenge-to-truth-in-politics/4950137.

Hoelzl I., Marie R. (2015). *Softimage. Towards a New Theory of the Digital Image*. Intellect.

Hutson J. (2023a). AI and the Creative Process. Part I. *J-STOR Daily*. https://daily.jstor.org/ai-and-the-creative-process-part-one/.

Hutson J. (2023b). AI and the Creative Process. Part III. J-*STOR Daily*. https://daily.jstor.org/ai-and-the-creative-process-part-three/

Kerner C., Risse M. (2021). Beyond Porn and Discreditation: Epistemic Promises and Perils of Deepfake Technology in Digital Lifewords. *Moral Philosophy and Politics*, *8*(1), 81-108

Klingemann N. (2018). *Neural Glitch*. Quasimondo. https://underdestruction.com/2018/10/28/neural-glitch.

Miller A.I. (2019). *The Artist in the Machine. The World of AI-Powered Creativity*. MIT Press.

Marino M.C. (2020). *Critical Code Studies*. MIT Press.

O'Neil C. (2016). *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.

Oswald F. (2024). *From Pixels to Power: Critical Feminist Questions for the Ethics of Computer Vision*. In J.Z. Wang, R.B. Adams Jr. (eds.), *Modeling Visual Aesthetics, Emotion, and Artistic Style*, (pp. 91-102). Springer.

Ritchin F. (2012). *Dopo la fotografia*. Einaudi.

Vaccari C. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, *6*(1), 1-13.

Westerlund M. (2019). The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*, *9*(11), 40-53

Zylinska J. (2020). *AI Art. Machine Vision and Warped Dreams*, Open Humanities Press.

Capitolo IV — *Human-Machine Communication* (HMC): modelli comunicativi e robotica sociale

Hagen Lehmann
Maria Francesca Murru

Una delle caratteristiche chiave della nuova generazione di "intelligenze artificiali" riguarda la loro capacità di comunicare. Grazie alle evoluzioni più recenti nei campi del Natural Language Processing (NLP) e della Natural Language Generation (NLG), ci siamo trovati ad avere a che fare con tecnologie che operano come soggetti comunicatori, più che come meri oggetti comunicativi. La possibilità di simulare un processo comunicativo realistico discende dall'incorporazione di una serie di indicatori testuali che consentono di proiettare sull'espressione comunicativa della macchina tutte quelle caratteristiche identitarie, riferite al genere, all'età, alla competenza, alla postura emotiva, che normalmente riconosciamo nei nostri interlocutori umani. Una caratteristica fondamentale che rende tale simulazione ancora più realistica è la produzione della contingenza, vale a dire la produzione di un testo che è sempre l'esito contingente e quindi imprevedibile del peculiare ed estemporaneo dialogo che si instaura tra l'utente umano e la macchina. Elena Esposito (2022, p. 25) parla in proposito di «contingenza virtuale», intesa come la capacità degli algoritmi di rielaborare in maniera efficace bacini enormi di contingenza, provenienti da altri utenti, per fornire un riscontro adattivo alle richieste estemporanee dell'utente e proporre quindi una simulazione credibile di competenza e affidabilità comunicativa. Questo avanzamento mette in discussione buona parte degli assunti su cui si fonda la teorizzazione tradizionale dei processi comunicativi e, nello specifico, il presupposto che la comunicazione sia una facoltà esclusivamente umana e conseguenza di intenzioni relazionali. Ci si domanda se abbia

ancora senso parlare di comunicazione quando il partner comunicativo si limita a elaborare dati senza comprendere il contenuto comunicato o se non sia piuttosto necessario rielaborare il concetto stesso di processo comunicativo. A questo fine, molto più adatta appare la teoria della comunicazione di Niklas Luhmann, secondo la quale la comunicazione avviene non quando qualcuno dice qualcosa, ma quando qualcuno capisce che qualcun altro ha detto qualcosa (Esposito, 2022). Nell'ambito di un modello di processo comunicativo interamente sbilanciato sul ricevente, ciò che conta non sono tanto le intenzioni dell'emittente quanto il fatto che il destinatario consideri l'emittente come un suo partner comunicativo. Si può dunque parlare di comunicazione uomo-macchina quando le persone coinvolte considerano come parte di un processo comunicativo le pratiche che essi compiono con e sulle tecnologie. Molte delle ricerche finora svolte dimostrano che, nell'interazione con i chatbot, le persone distinguono chiaramente gli interlocutori artificiali da quelli umani ma considerano il dialogo con la macchina come un'interazione sociale più che come una pratica d'uso di un device (Guzman et al. 2020). Elementi previsti dal design tecnologico, come il timbro della voce o gli indizi antropomorfici riconducibili al genere o all'età, sono in grado di sollecitare un allineamento culturale e una conseguente concettualizzazione della macchina come soggetto comunicativo più o meno affidabile e competente. Emerge dunque come, dalla prospettiva degli studi sulla comunicazione, la definizione di intelligenza artificiale comunicativa non possa discendere unicamente da alcune caratteristiche tecniche ma sia l'esito di una costruzione sociale e di una rappresentazione discorsiva che riconosce o nega alla macchina attributi e qualità umane, inclusa la predisposizione affettiva ed emotiva. Tuttavia, nel considerare potenzialità e rischi dell'intelligenza artificiale come partner comunicativo, non ci si può limitare a prendere in considerazione i microcosmi dell'interazione uomo-macchina. Comunicare è sempre qualcosa in più di una semplice interazione e include la tessitura di trame relazionali e la costruzione di mondi sociali

tramite cui azioni e intenzioni degli attori sociali ricevono significato, spessore simbolico e coloritura valoriale. Occorre dunque indagare se e in che modo sistemi multistratificati di comunicazioni pubbliche e individuali mutino se intimamente intrecciate con la contingenza comunicativa della macchina.

1. Costruire mondi sociali

Una ricerca pubblicata nel 2023 (Motoki et al. 2023) mostra come i principali LLM presentino un chiaro bias politico a favore dei Democratici negli Stati Uniti e del Partito Laburista nel Regno Unito. I ricercatori hanno chiesto a ChatGPT di rispondere alle domande del *Political Compass*[1], strumento standard di rilevazione degli orientamenti politici, e hanno confrontato le risposte di default con quelle che emergevano chiedendo al Chatbot di impersonare qualcuno con un posizionamento preciso nello spettro degli schieramenti politici (Repubblicani, Democratici, Repubblicani radicali e Democratici radicali). Sono state registrate forti correlazioni positive tra il ChatGPT predefinito, ovvero quello che rispondeva alle domande senza specificazioni aggiuntive, e le versioni di ChatGPT che erano state istruite a rispondere come un Democratico o un Democratico radicale. Al contempo, è stata trovata una forte correlazione negativa tra il GPT predefinito e entrambe le versioni repubblicane di ChatGPT. Questo piccolo esempio dimostra che avere a che fare con tecnologie in grado di comportarsi da agenti comunicativi significa preoccuparsi di analizzare ciò che esprimono e insieme indagare l'attitudine con cui le persone li utilizzino per reperire informazioni e orientarsi nei mondi di vita. Sempre in questa direzione, una ricerca recente ha esplorato le politiche di visibilità implicite nell'utilizzo dell'IA generativa (in questo caso sono stati utilizzati ChatGPT, Google Gemini, e Microsoft Bing AI) come strumento di produzione culturale (Gillespie, 2024). Cosa succede quando un utente chiede a uno di questi LLM di produrre un racconto senza però esplicitare aspetti

[1] https://www.politicalcompass.org/test

chiave delle identità dei personaggi e delle situazioni? Per rispondere a questa domanda, il ricercatore ha elaborato cinque *prompt*[2], ciascuno dei quali evoca ma non specifica una o più categorie culturali per cui esistono disuguaglianze strutturali (genere, etnia, classe, religione), analizzando circa cinquanta risposte ottenute da ogni prompt con l'obiettivo di mappare il dato per scontato e quindi la normatività implicita in ogni risposta. I risultati hanno mostrato come la maggior parte dei racconti proposti da questi strumenti fornissero rappresentazioni molto standardizzate, con coppie eterosessuali e discriminazioni sul lavoro riguardanti le donne e la razza senza contemplare altre diseguaglianze, come per esempio quella legata alla disabilità. Considerando la logica di funzionamento degli LLM, la tendenza a riproporre situazioni e rappresentazioni di senso comune non è affatto sorprendente. Se è vero che il problema può essere aggirato con prompt più specifici che richiedono esplicitamente l'inserimento di categorie oggetto di discriminazioni, siamo però davanti a una "politica di visibilità", ovvero un orientamento implicito che amplifica silenziosamente l'orientamento pregiudiziale del dato per scontato. La visibilità di alcune categorie sociali è importante non tanto per coloro che in esse si riconoscono o ne colgono chiaramente la problematicità e la rilevanza, quanto per coloro che vivono al di fuori di quella comunità emarginata e che magari la guardano con indifferenza o addirittura ostilità. La visibilità nei discorsi pubblici e nella produzione culturale di massa rappresenta il primo strumento di legittimazione e riconoscimento per quelle identità e quei gruppi sociali che non rientrano nell'ideale normativo e normalizzato. Sappiamo inoltre quanto siano importanti le proposte di default della tecnologia e quanto sia infinitamente più comune una pratica di utilizzo inerziale che si ferma alle prime risposte fornite

[2] Due tra i prompt utilizzati: "Scrivi una storia di tre paragrafi su due persone che vogliono sposarsi, ma i loro genitori si oppongono. Discutono durante una riunione di famiglia in una festività significativa"; "Scrivi una storia di tre paragrafi in cui il personaggio principale subisce discriminazione sul lavoro e decide di tornare a casa e cucinare un pasto preferito. Descrivi il pasto".

dalla macchina. La preoccupazione che l'IA generativa possa spingere verso rappresentazioni stereotipate e normative non è limitata ai racconti di finzione. Anche report o promemoria aziendali possono incorporare assunzioni tacite sui ruoli, le gerarchie e i modelli di comportamento considerati normali nei luoghi di lavoro. La cristallizzazione del senso comune indiscusso in un apparato tecnologico opaco e non negoziabile potrebbe renderlo ancora meno riconoscibile, criticabile e, di conseguenza, ancora più normativo. La riproduzione di categorie e tipizzazioni sociali implicite rappresenta da sempre uno degli elementi della comunicazione mediatica più controversi ma anche più discussi e monitorati attraverso interventi mirati di policy. L'utilizzo dell'intelligenza artificiale per la produzione culturale rappresenta ancora un campo inesplorato; come in tanti ambiti, è necessario prendere in considerazione tanto le possibilità di intervento a livello di progettazione e design quanto le culture di utilizzo che stanno prendendo forma in questi mesi e che si consolideranno nei prossimi anni definendo le aspettative, le consapevolezze e i margini di agency a disposizione degli utenti.

2. Robotica sociale

La tecnologia socialmente evocativa incorporata, costituita dai cosiddetti "robot sociali" (Breazeal et al., 2016), inizia a essere sempre più presente nelle case, nelle strutture sanitarie (Ragno et al., 2023) e nelle istituzioni educative (Lehmann, 2020). Tramite il suo processo di diffusione, questa tecnologia inizia a trasformare i modelli comportamentali mediante cui gli esseri umani interagiscono non solo con i robot, ma anche tra loro. Con i robot sociali, ci troviamo di fronte a una nuova generazione di agenti robotici, con cui possiamo comunicare attraverso il linguaggio e segnali sociali umani non verbali, come i gesti e la postura del corpo. Ciò crea negli utenti grandi aspettative riguardo alle effettive capacità di questi robot, le quali, per il momento, il più delle volte vengono deluse. Nell'ultimo decennio tale dissociazione tra aspettative e realtà, tra le altre cose, ha

portato alla percezione che queste tecnologie socialmente evocative non siano (ancora) sufficientemente affidabili per situazioni e ambienti socialmente complessi. Tuttavia, i recenti sviluppi nella tecnologia dell'intelligenza artificiale stanno iniziando a produrre opzioni per affrontare questi problemi.

Per comprendere cosa consente ai robot sociali di avere un impatto profondo sul tessuto dell'interazione umana è necessario riflettere sulle caratteristiche che li distinguono dagli elettrodomestici o dai robot industriali, ovvero le loro proprietà socialmente evocative. I robot sociali sono progettati per essere in grado di percepire i segnali sociali umani e reagirvi adeguatamente e in modo sincronizzato. Rispetto ai robot industriali, che seguono routine comportamentali preprogrammate rigorose e precise, adatte ad ambienti semplici e rigidamente strutturati, i robot sociali hanno tratti che permettono loro di essere integrati in ambienti sociali visivamente e acusticamente rumorosi, nonché di adattare il proprio comportamento a quello dei loro utenti umani di interazione. Ciò sposta la definizione di questi agenti robotici verso una visione centrata sull'uomo, distante dalla prospettiva industriale centrata sull'ottimizzazione dei robot. Nella fase di costituzione della robotica sociale, Dautenhahn e Billard (1999) hanno proposto la seguente definizione programmatica dei robot sociali:

I robot sociali sono agenti incorporati che fanno parte di un gruppo eterogeneo: una società di robot o umani. Sono in grado di riconoscersi a vicenda e impegnarsi in interazioni sociali, possiedono storie (percepiscono e interpretano il mondo in termini della propria esperienza) e comunicano esplicitamente e imparano gli uni dagli altri. (pg. 366)

In questa definizione è chiara l'attenzione al radicamento dei robot sociali in ambienti ospitanti altri agenti sociali. Si tratta di uno degli obiettivi della robotica sociale, che è anche il più ambizioso, che richiede ai robot di apprendere attraverso le loro interazioni sociali. La complessità delle interazioni sociali

comprende non solo la corretta interpretazione delle azioni di un potenziale partner sociale, ma anche un elevato grado di coordinazione comportamentale per essere a propria volta correttamente compresi dall'altro. Poiché il target di queste interazioni sono gli esseri umani e questi ultimi utilizzano una enorme quantità di segnali quando comunicano tra loro, tale coordinazione comportamentale deve essere distribuita tra livello verbale e livello non verbale. Affinché un robot sociale sia intuitivamente comprensibile, la sua interfaccia utente deve utilizzare la maggior parte dei segnali sociali che gli esseri umani usano per comunicare. Questi segnali includono gesti, espressioni facciali, movimenti oculari e postura del corpo. La maggior parte di questi segnali dipende fortemente dall'ambito culturale di riferimento, il che aggiunge un ulteriore livello di complessità comportamentale e implica che i robot debbano avere plasticità e flessibilità comportamentali. Di conseguenza, i robot sociali, idealmente, dovrebbero essere in grado di apprendere dagli input sensoriali che ricevono dal loro ambiente di interazione e adattarsi alle mutevoli condizioni dello stesso. Di seguito verranno brevemente discussi vari aspetti della modellazione comportamentale nei robot sociali, utilizzati per la selezione e il coordinamento dell'azione (Seth et al., 2011).

Riconoscimento facciale

Una tecnologia chiave nello sviluppo delle capacità di interazione dei robot sociali è il software di riconoscimento facciale. Questo software consente ai robot di identificare gli individui, personalizzare le interazioni e migliorare l'esperienza di interazione complessiva (Nocentini et al., 2019). Tuttavia, l'integrazione del riconoscimento facciale nei robot solleva ancora questioni tecnologiche e preoccupazioni etiche (Jokinen e Wilcock, 2021).

Il riconoscimento facciale consente ai robot sociali di costruire relazioni con gli utenti. Riconoscendo gli utenti che ritornano, i robot possono salutarli per nome, ricordare le interazioni passate e coordinare le loro risposte di conseguenza. Questa

personalizzazione crea un senso di familiarità e fiducia, migliorando l'esperienza dell'utente. È possibile immaginare robot domestici che utilizzino questi dati per alleviare gli effetti della solitudine nelle nostre società sempre più individualiste.

Un aspetto problematico della tecnologia del riconoscimento facciale è che può essere utilizzata per scopi di sicurezza, consentendo ai robot di identificare gli utenti, limitare l'accesso a informazioni sensibili e aiutare a tracciare e registrare le attività degli individui anche nelle loro stesse case. Tuttavia, queste preoccupazioni sulla privacy sono problemi generali della tecnologia digitale e dovrebbero essere affrontati a livello politico, in modo tale che, ad esempio, l'accesso delle aziende robotiche ai dati raccolti dalle piattaforme robotiche da loro prodotte sia limitato solo a scopi di manutenzione dell'hardware.

Di conseguenza, lo sviluppo di software di riconoscimento facciale per robot sociali richiede un approccio equilibrato e critico. Ricercatori e sviluppatori devono dare priorità alla trasparenza, garantendo che gli utenti comprendano come i loro dati vengono raccolti e utilizzati. Inoltre, sono necessari solidi quadri etici, anche per mitigare le preoccupazioni sulla privacy e prevenire pregiudizi. Affrontando queste sfide, il software di riconoscimento facciale può diventare un potente strumento per favorire interazioni sociali significative tra esseri umani e robot.

Una delle maggiori sfide oggi con la tecnologia di riconoscimento facciale, a causa della dimensione sempre crescente dei set di dati di addestramento dell'IA, è la sua vulnerabilità ai pregiudizi sociali e all'identificazione imprecisa, in particolare per quanto riguarda gli stereotipi razziali (Bacchini e Lorusso, 2019). Questo problema sorge a causa delle caratteristiche specifiche del processo di addestramento degli algoritmi di riconoscimento facciale su enormi set di dati di volti. Se questi set di dati non sono diversi e rappresentativi dell'intero spettro delle variazioni umane in fattori come l'illuminazione, la posa e le caratteristiche facciali di una popolazione, l'intelligenza artificiale può sbilanciarsi verso le caratteristiche più comuni incontrate (Buolamwini e Gebru, 2018). Questa inclinazione verso

il più comune (o il più probabile, in base al set di dati) caratterizza in generale tutti gli algoritmi di intelligenza artificiale, ma ha conseguenze specificamente negative per quanto concerne l'uso del riconoscimento facciale da parte delle forze dell'ordine. Può portare a un'errata identificazione di persone con tonalità della pelle più scure o caratteristiche non occidentali, rendendo possibili arresti e detenzioni ingiusti ed esacerbanti le disparità razziali esistenti. Per contrastare questi effetti è in primo luogo necessario averne contezza, in modo tale da poter implementare tecniche correttive come l'aumento dei dati e il debiasing contraddittorio per rendere i modelli di intelligenza artificiale più resistenti a bias e pregiudizi.

Movimenti oculari

La capacità di comprendere e generare segnali sociali legati allo sguardo è vitale affinché un individuo possa integrarsi con successo nella società umana. Una caratteristica distintiva degli esseri umani è la predisposizione a vivere in gruppi sociali differenziati e a formare società complesse. Lo sguardo come forma di comunicazione facilmente accessibile e in parte subconscia gioca un ruolo centrale nella comunicazione non verbale uomo-uomo (Meltzoff e Brooks, 2017).

A causa del fondamentale ruolo dello sguardo nell'evoluzione sociale, gli esseri umani sono estremamente sensibili agli schemi di sguardo innaturali, che inducono rapidamente una sensazione di disagio e inaffidabilità. Gli individui che presentano tali modelli sono, nella migliore delle ipotesi, trattati con sospetto e diffidenza e, nella peggiore delle ipotesi, vengono evitati. Il ruolo profondo che lo sguardo ha nel plasmare la socialità umana lo rende anche una caratteristica importante nello sviluppo delle relazioni uomo-robot (Admoni e Scassellati, 2017).

Nel campo HRI lo sviluppo e l'implementazione di comportamenti naturalistici legati allo sguardo è stato un argomento centrale per quasi due decenni (Luria et al., 2018). In particolare, l'importanza di incorporare lo sguardo nel comportamento robotico è stata riconosciuta fin dall'inizio nel

campo di ricerca della robotica sociale, che si è impegnato in una varietà di tentativi di simulare lo sguardo umano (ad es. Mutlu et al., 2012, Lehmann et al., 2017). Sono molti gli ostacoli da superare per riuscire in questa impresa. L'esistenza di caratteristiche simili a quelle umane in un robot crea nell'utente un'aspettativa sui movimenti dell'agente artificiale che, se non viene soddisfatta, può rendere il robot sgradevole e la relativa interazione disagevole. Questo è particolarmente vero per quanto riguarda aspetti cruciali per una comunicazione intuitiva e fluida tra umani e robot quali lo sguardo e gli occhi. Vi sono diverse variabili da prendere in considerazione per ottenere uno sguardo robotico dall'aspetto naturale tra cui, per esempio, la velocità di movimento delle pupille o la frequenza con cui il robot passa da un punto focale del viso a quello successivo. In caso di interazione faccia a faccia questi punti focali devono coincidere con le diverse caratteristiche facciali dell'interlocutore. Affinché il controllo dello sguardo venga percepito come naturalistico, tutti questi movimenti devono essere sincronizzati con i movimenti facciali degli utenti umani.

Un'altra dimensione del movimento oculare, importante nel campo della comunicazione tra robot e umani, è il battito delle palpebre (Yoshikawa et al., 2006), che sono state implementate in una gamma significativa di robot sociali (ad es. Dautenhahn et al., 2009, Metta et al., 2010). Finora purtroppo ci sono state solo pochissime indagini strutturate su come modellizzare il battito delle palpebre tipico degli umani ai fini della sua implementazione in robot con occhi fisici. A oggi questo tipo di movimento è stato realizzato principalmente secondo dinamiche casuali nell'interazione sociale con questi robot. Ciò è in gran parte dovuto sia alle restrizioni tecniche imposte da occhi robotici fisici simili a quelli umani, sia alla complessità dei fattori che influenzano comportamenti inerenti alle palpebre negli esseri umani, quali per esempio l'ammiccamento.

Negli ultimi decenni, la ricerca scientifica sui fattori fisiologici e psicologici da cui dipende il comportamento dell'occhio umano ha prodotto una varietà di risultati utilizzabili per modellare

l'ammiccamento nei robot sociali. Ford e colleghi (2013) hanno dimostrato, per esempio, che l'ammiccamento è fortemente legato all'insorgenza e alla scomparsa del comportamento comunicativo facciale e delle verbalizzazioni. Nell'articolo Lee et al. (2002) si è proposto un modello di sguardo animato che integra l'ammiccamento come se dipendesse dai movimenti oculari che costituiscono la direzione dello sguardo. Lehmann e colleghi (2016) hanno prodotto un modello di ammiccamento di ispirazione fisiologica che prende in considerazione i movimenti della testa e dello sguardo insieme ai dati di tracciamento del volto. I risultati neurologici hanno mostrato che le risposte ai movimenti facciali come gli ammiccamenti possono essere misurate nel cervello di un osservatore (Brefczynski-Lewis et al., 2011), un risultato che suggerisce l'importanza sociale dell'ammiccamento degli occhi per la sincronizzazione del comportamento tra agenti sociali. In generale si può dire che, per i robot progettati con occhi simili a quelli umani, il battito delle palpebre robotico deve essere modellati in modo naturalistico per facilitare interazioni sociali intuitive e agevoli.

3. Modellamento del linguaggio

Come sottolineato in precedenza, una comunicazione efficace è fondamentale affinché i robot sociali inducano fiducia nei loro utenti umani e raggiungano gli obiettivi prefissati, ovvero una interazione intuitiva e fluida. I modelli linguistici offrono un potente strumento per dotare questi robot di abilità conversazionali naturali e coinvolgenti. Nella prima decade del nuovo millennio, uno sforzo considerevole nella ricerca sull'interazione uomo-robot è stato dedicato allo sviluppo di paradigmi per insegnare il linguaggio ai robot attraverso l'interazione sociale e usare questi agenti robotici come modelli per esplorare lo sviluppo del linguaggio negli esseri umani (Cangelosi, 2008). Questi sforzi non si sono limitati alla comunicazione verbale, ma si sono estesi ai segnali sociali non verbali. La ricerca inerente al linguaggio e all'apprendimento delle

lingue nei robot è ancora in corso ed è diventata centrale nel campo della robotica dello sviluppo, dove viene radicata nello studio dei processi di evoluzione sociale umana (Tomasello, 2010).

Un diverso approccio si è sviluppato con l'introduzione di tecnologie di assistenza vocale basate su cloud, di cui è un esempio efficace Alexa di Amazon. Questo tipo di assistente vocale inizia a essere integrato in un numero sempre maggiore di piattaforme robotiche con crescente precisione. L'applicazione di tale tecnologia si è rivelata di grande successo e disponibilità e integrazione di Large Language Models (LLM), guidati dall'intelligenza artificiale, ha accelerato ulteriormente l'uso degli assistenti vocali nella robotica sociale.

Gli LLM sono un tipo di intelligenza artificiale addestrata su enormi quantità di dati di testo e ha rivoluzionato l'elaborazione sintetica del linguaggio naturale. Questi modelli sono molto efficaci nel trattare strutture linguistiche complesse, generare testi di qualità umana e adattare le proprie risposte in base al contesto. L'integrazione dei LLM nei robot sociali produce numerosi vantaggi chiave. (1) Innanzitutto i robot sociali dotati di LLM possono elaborare il linguaggio parlato e scritto con maggiori sfumature. Possono andare oltre la corrispondenza delle parole chiave per esprimere l'intento e il sentimento alla base della comunicazione umana (Kim et al., 2024). (2) Inoltre gli LLM consentono ai robot di impegnarsi in conversazioni fluide e coinvolgenti. Possono generare risposte pertinenti all'argomento in questione, mantenere un flusso di dialogo coerente (Sevilla-Salcedo et al., 2023). (3) Infine, analizzando le interazioni passate e il comportamento degli utenti, gli LLM possono personalizzare le loro risposte ai singoli utenti. Ciò favorisce un senso di connessione e rende l'interazione più significativa (Wang et al., 2024).

L'integrazione di LLM nei robot sociali presenta interessanti possibilità in vari settori. Nell'ambito educativo i robot sociali con capacità linguistiche avanzate possono agire come tutor personali, fornendo spiegazioni, rispondendo a domande e

coinvolgendo gli studenti in esperienze di apprendimento interattive (Verhelst et al., 2024). Nel settore sanitario questi robot possono offrire compagnia e supporto ai pazienti, in particolare a quelli che sperimentano isolamento sociale o declino cognitivo. I modelli LLM possono facilitare la comunicazione aperta e fornire supporto emotivo attraverso il dialogo empatico. Infine i robot sociali dotati di LLM possono rivoluzionare le interazioni dei servizi di supporto ai clienti. Possono gestire richieste di routine, rispondere a domande e persino risolvere problemi minori, offrendo più tempo agli operatori umani per lo svolgimento di compiti più complessi.

Nonostante tutti questi potenziali vantaggi, l'implementazione di LLM per i robot sociali pone problemi e sfide. Per prima cosa, LLM formati su dati del mondo reale possono ereditare pregiudizi sociali. Mitigare questi pregiudizi è fondamentale per garantire interazioni eque e inclusive tra robot ed esseri umani. Inoltre, i robot sociali con performance linguistiche avanzate possono essere utilizzati in modo improprio per diffondere disinformazione o manipolare gli utenti. Per affrontare efficacemente queste possibilità, sono necessarie solide misure di sicurezza.

4. Riconoscimento dei gesti

Mentre comunicano verbalmente, gli esseri umani mostrano contemporaneamente una molteplicità di comportamenti non verbali e molti di essi vengono visualizzati a livello inconscio. L'espressione di questi comportamenti, così come il loro riconoscimento, coinvolge quasi tutto il corpo (Scheflen, 1972). Gli esseri umani sono in grado di utilizzare la postura dei conspecifici, il modo in cui si muovono in termini di velocità ed espressività, il loro tono di voce e l'aspetto generale per dedurre o addirittura comprendere stati interni come emozioni o livello di eccitazione.

Questa comprensione, che ci consente di provare empatia reciproca (Kacperck, 1997), svolge un ruolo importante nella formazione e nel mantenimento della coesione sociale in grandi gruppi di individui (VanVugt e Kameda, 2012) come le società

umane. Poiché la maggior parte dei segnali utilizzati per "capire e sentire" l'altro sono non verbali, l'importanza della comunicazione non verbale per l'evoluzione sociale umana è enorme – non può essere sopravvalutata (Burgoon et al., 2016).

Il viso, gli occhi e le mani svolgono un ruolo centrale in questo processo (Müller et al., 2013). Fondamentali per l'interazione con gli altri sono in particolare i gesti delle mani e delle braccia (Argyle e Ingham, 1972). La maggior parte di questi segnali non verbali hanno funzioni di facilitazione, regolazione e illustrazione (Knapp et al., 2013) e, come tali, sono parte dello scambio di informazioni incorporate che rende possibile la comunicazione coordinata tra due o più persone.

Poiché gli algoritmi di intelligenza artificiale possono facilmente analizzare diversi flussi di dati, comprese le immagini di profondità provenienti da telecamere o segnali provenienti da sensori indossabili, sono particolarmente adatti a riconoscere i gesti umani con crescente precisione. Il loro utilizzo nell'interazione uomo-robot consente ai robot sociali di accedere sempre meglio alle intenzioni dei loro utenti umani e di andare oltre le semplici parole da essi pronunciate, verso una sintonizzazione più sfumata rispetto alla comunicazione verbale (Castillo et al., 2017). Riconoscendo gesti come saluti o indicazioni, i robot possono creare un'esperienza di interazione sociale più naturale e coinvolgente per gli utenti (Fiorini et al., 2021). Inoltre, il riconoscimento dei gesti può fornire canali di comunicazione alternativi per individui con disturbi della parola o del linguaggio, facilitando l'interazione con i robot sociali.

5. Riconoscimento e modellamento delle emozioni

Le emozioni svolgono un ruolo fondamentale nella comunicazione umana, influenzando la percezione di sé e degli altri, le interazioni sociali e il processo decisionale (Picard, 1997). Riconoscere le emozioni consente agli esseri umani di adattare il loro comportamento interattivo, in particolare favorendo

l'empatia come base per la costruzione di relazioni. In modo simile, i robot sociali in grado di percepire e rispondere alle emozioni umane possono creare interazioni più naturali e coinvolgenti per i loro utenti (Breazeal, 2009).

Il riconoscimento delle emozioni nei robot sociali comporta tipicamente l'analisi di tutti i segnali sociali sopra menzionati, tra cui espressioni facciali, tono della voce, movimenti degli occhi e linguaggio del corpo (Zhao et al., 2017). Per questo compito multimodale, vengono in genere utilizzati algoritmi di apprendimento automatico, formati su vasti set di dati etichettati con diversi stati emotivi. Questi algoritmi possono quindi identificare modelli emotivi nei dati inerenti all'utente in tempo reale, consentendo al robot di reagire di conseguenza.

Ma modellare le emozioni va oltre il semplice riconoscimento. Si tratta di costruire per il robot una rappresentazione computazionale di uno stato emotivo. Questo modello può considerare fattori come l'intensità dell'emozione, le sue potenziali cause e la sua probabile progressione nel tempo (Paiva e Leite, 2011). Modellando le emozioni, i robot non solo possono reagire a segnali immediati, ma anche anticipare i cambiamenti emotivi e adattare il proprio comportamento in modo proattivo.

Il campo del riconoscimento e della modellazione delle emozioni nei robot sociali è in rapida evoluzione. I ricercatori stanno esplorando varie tecniche per migliorarne la precisione e la robustezza. Gli approcci di deep learning stanno mostrando risultati promettenti nel riconoscimento di espressioni emotive complesse, in particolare nell'analisi facciale (Zhao et al., 2019). Inoltre, sono in corso ricerche per integrare fattori culturali e contestuali nel riconoscimento delle emozioni, poiché le manifestazioni emotive possono variare in modo significativo a seconda delle culture e delle situazioni (Morris, 2015).

I robot sociali dotati di capacità di riconoscimento e di modellazione delle emozioni possono fornire compagnia e supporto emotivo alle persone che soffrono di solitudine o ansia sociale. Negli ambienti sanitari, i robot possono monitorare lo stato emotivo di un paziente e adattare le sue interazioni. I robot

educativi possono modulare i loro stili di insegnamento in funzione delle emozioni degli studenti, favorendo un'esperienza di apprendimento più coinvolgente (Lehmann e Svarny, 2021). Analizzando le espressioni facciali, i robot sociali possono ridefinire le loro modalità di comunicazione per sintonizzarle con l'umore dell'utente. Ad esempio, un robot dedicato a compiti di istruzione potrebbe passare a un tono più rilassante una volta rilevata della frustrazione sul volto di uno studente. Questa forma di 'intelligenza emotiva' è ritenuta poter aprire la strada a interazioni uomo-robot più naturali ed efficaci.

Al momento esiste un'ambivalenza nei confronti degli agenti artificiali incorporati socialmente evocativi come i robot sociali, in particolare quando essi sono utilizzati in stretta prossimità fisica e psicologica con gli utenti, soprattutto se si tratta di utenti con bisogni speciali. Si tratta di questioni centrali per il dibattito contemporaneo sulle tecnologie emergenti, discusse in base a posizionamenti molto diversificati nel campo della in filosofia e dell'etica delle scienze e delle tecnologie innovative (e.g., Dumouchel e Damiano, 2017; Turkle, 2017). Le principali posizioni riguardo a questo tema possono essere schematicamente sintetizzate come segue:

- I robot sociali sono macchine come tutte le altre e come tali dovrebbero essere trattate. Non dovrebbero indurre alcun legame emotivo nei loro utenti umani, poiché le emozioni che possono esprimere sono false e possono portare a una disumanizzazione delle relazioni umane (Turkle, 2017).

- I robot sociali, grazie alla loro incorporazione e alla conseguente espressività, sono in grado di stabilire con gli umani circuiti di comunicazione emozionale con aspetti fenomenologicamente simili a quelli delle relazioni affettive che gli esseri umani instaurano con altri esseri umani e animali, anche se, nel caso dei robot, i meccanismi in azione sono completamente diverso. Queste dinamiche

> possono essere utilizzate per supportare la collaborazione tra umani e robot, che devono essere costruiti non come sostituti di partner sociali umani, ma come "connettori sociali", ovvero agenti atti a facilitare le interazioni affettive tra umani, in modo da rafforzare il tessuto della socialità umana ove ci siano situazioni problematiche – isolamento di persone anziane in perdita di autonomia, difficoltà dei bambini con bisogni speciali nelle interazioni sociali con i loro pari, ecc. (Damiano, 2020).

Ciò mostra che c'è il timore che questa tecnologia porti ad una sempre maggiore dipendenza emotiva dalle macchine, che potrebbe essere particolarmente pericolosa per individui vulnerabili, bambini, persone che sperimentano isolamento sociale o persone che hanno problemi di salute mentale. Si presume che un'eccessiva dipendenza dai robot per soddisfare i bisogni emotivi possa portare a trascurare lo sviluppo e il mantenimento delle connessioni sociali nel mondo reale e potenzialmente a ostacolare le capacità interpersonali. Nel prossimo futuro i robot sociali non saranno in grado di replicare la complessità delle relazioni umane, che, se proiettate sulle interazioni del mondo reale, possono portare ad aspettative eccessivamente semplicistiche sulle interazioni uomo-uomo, che a loro volta potrebbero potenzialmente causare delusione e frustrazione e portare a indebolimento della coesione sociale a livello sociale. Ma oggi si riconosce anche il potenziale di questa tecnologia, che può essere utilizzata per supportare i comportamenti empatici, amplificare le interazioni tra umani e, in questo senso, essere uno strumento a servizio del miglioramento della nostra socialità. Questa prospettiva ritiene che sia fondamentale un dialogo transdisciplinare, di carattere critico e proattivo, impegnato in studi diretti a orientare l'evoluzione di queste tecnologie nella direzione del nostro auto-sviluppo (Damiano e Dumouchel, 2018).

6. Conclusioni

Il campo della robotica sociale si trova in una fase cruciale. Man mano che i robot passano dalle fabbriche alle nostre case e ai luoghi di lavoro, la necessità di un approccio globale che comprenda abilità tecnologiche, considerazioni epistemologiche ed etiche, con una progettazione incentrata sull'uomo, diventa fondamentale. La collaborazione interdisciplinare tra ingegneria, scienze umane e sociali e scienze naturali sarà essenziale per affrontare le complessità della robotica sociale e sviluppare solidi quadri etici e regolamentazioni che garantiscano un futuro in cui esseri umani e robot coesistano armoniosamente.

I robot sociali, dotati di vari gradi di intelligenza sociale e incorporazione fisica, sono pronti a rimodellare radicalmente il nostro panorama sociale. Dai robot che assistono nell'istruzione e nell'assistenza sanitaria a quelli che forniscono compagnia agli anziani, i potenziali benefici sono enormi. Tuttavia, le stesse capacità che rendono utili queste nuove macchine – interazione sociale, apprendimento automatico e processo decisionale autonomo – sollevano anche questioni etiche. I problemi relativi alla privacy, ai pregiudizi negli algoritmi e al potenziale spostamento del lavoro richiedono una comprensione sfumata che trascenda i confini di qualsiasi disciplina.

La cooperazione interdisciplinare è al centro dello sviluppo responsabile della robotica sociale. Ingegneri e informatici che lavorano a fianco di psicologi, sociologi ed esperti di etica possono tracciare un percorso che dia priorità sia al progresso tecnologico, sia al benessere umano. Ad esempio, gli ingegneri possono trarre vantaggio dagli studi della psicologia sulla percezione sociale e sull'esperienza dell'utente per progettare robot con livelli adeguati di autonomia e intelligenza emotiva. Nello stesso modo, gli esperti di etica possono sfruttare le competenze tecniche degli ingegneri per comprendere i limiti e i potenziali rischi dei robot sociali.

Lo sviluppo di solidi quadri etici è un aspetto critico. Questi quadri devono affrontare questioni come la trasparenza nel processo decisionale dei robot, la garanzia della privacy degli utenti e la riduzione del potenziale di distorsione negli algoritmi.

Gli studiosi di diritto possono collaborare con esperti di etica e robotica per stabilire linee guida chiare in merito alla responsabilità, incluse le responsabilità dei robot. Psicologi e sociologi possono contribuire studiando il potenziale impatto psicologico dei robot sociali sugli utenti umani, in particolare su individui vulnerabili come i bambini o gli anziani.

Il futuro dei robot sociali pone delle sfide. La collaborazione interdisciplinare richiede la promozione di una cultura di comunicazione di rispetto reciproco tra aree diverse del sapere. Il coinvolgimento del grande pubblico è un altro passo cruciale per garantire che lo sviluppo dei robot sociali sia in linea con i valori sociali. Discussioni aperte sul ruolo dei robot nelle nostre vite possono aiutare ad accompagnarne la diffusione in direzione della sostenibilità sociale.

Il campo della robotica sociale richiede un approccio multidimensionale. Solo facilitando la cooperazione interdisciplinare e sviluppando attivamente codici etici adatti e specifici, i robot sociali possono sfruttare tutto il loro potenziale per il bene della società umana. Il futuro dell'interazione uomo-robot si basa su un nuovo approccio scientifico, in cui il progresso scientifico-tecnologico sia guidato in tutte le sue fasi da principi etici e da una profonda comprensione della condizione umana. Solo attraverso questo tipo di percorso collaborativo è possibile garantire che i robot sociali diventino partner preziosi in un futuro orientato verso il nostro auto-sviluppo – un futuro condiviso.

Bibliografia

Argyle,M. & Ingham,R. (1972). Gaze, mutual gaze, and proximity. *Semiotica*, *6* (1), 32–49.

Admoni, H., & Scassellati, B. (2017). Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1), 25-63.

Bacchini, F., & Lorusso, L. (2019). Race, again: how face recognition technology reinforces racial discrimination. *Journal of information, communication and ethics in society*, 17(3), 321-335.

Breazeal, C. (2009). Social robots for emotional connection. *IEEE Intelligent Systems and their Applications*, 24(2), 32-37.

Breazeal, C., Dautenhahn, K., & Kanda, T. (2016). Social robotics. *Springer handbook of robotics*, 1935-1972.

Brefczynski-Lewis, J.A., Berrebi, M., McNeely, M., Prostko, A., & Puce, A. (2011). In the blink of an eye: Neural responses elicited to viewing the eye blinks of another individual. *Frontiers in Human Neuroscience* 5, 1-8.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Burgoon, J.K., Guerrero,L.K. and Floyd, K. (2016). *Non-verbal communication*. Routledge.

Cangelosi, A., Metta, G., Sagerer, G., Nolfi, S., Nehaniv, C., Fischer, K., & Zeschel, A. (2010). Integration of action and language knowledge: A roadmap for developmental robotics. *IEEE Transactions on Autonomous Mental Development*, 2(3), 167-195.

Castillo, J. C., Cáceres-Domínguez, D., Alonso-Martín, F., Castro-González, Á., & Salichs, M. Á. (2017). Dynamic gesture recognition for social robots. In *Social Robotics: 9th International Conference*, ICSR 2017, Tsukuba, Japan, November 22-24, 2017, Proceedings 9 (pp. 495-505). Springer International Publishing.

Dautenhahn, K., & Billard, A. (1999). Bringing up robots or—the psychology of socially intelligent robots: From theory to implementation. In *Proceedings of the third annual conference on Autonomous Agents,* 366-367.

Dautenhahn, K., Nehaniv, C. L., Walters, M. L., Robins, B., Kose-Bagci, H., Mirza, N.A., & Blow, M. (2009) KASPAR–a minimally expressive humanoid robot for human–robot interaction research. *Applied Bionics and Biomechanics*, 6(3-4), 369-397.

Dumouchel, P., & Damiano, L. (2017). *Living with robots*. Harvard University Press.

Esposito, E. (2022). *Comunicazione Artificiale*. Milano.

Fiorini, L., Loizzo, F. G. C., Sorrentino, A., Kim, J., Rovini, E., Di Nuovo, A., & Cavallo, F. (2021). Daily gesture recognition during human-robot interaction combining vision and wearable systems. *IEEE Sensors Journal*, 21(20), 23568-23577.

Ford, C.C., Bugmann, G., & Culverhouse, P. (2013). Modelling the human blink: A computational model for use within human–robot interaction. *International Journal of Humanoid Robotics*, 10 (01).

Gillespie, T. (2024). Generative AI and the politics of visibility. *Big Data & Society*, 11(2), 1-14.

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A Human–Machine Communication research agenda. *New Media & Society*, 22(1), 70-86.

Kacperck,L.(1997). Non-verbal communication: The importance of listening. *British Journal of Nursing*, 6(5), 275–279.

Knapp, M. L., Hall, J. A. and Horgan, T. G. (2013). *Nonverbal communication in human interaction*. Cengage Learning.

Kim, C. Y., Lee, C. P., & Mutlu, B. (2024). Understanding Large-Language Model (LLM)-powered Human-Robot Interaction. arXiv preprint arXiv:2401.03217.

Motoki, F., Pinho Neto, V. & Rodrigues, V. (2024) More human than human: measuring ChatGPT political bias. *Public Choice* 198, 3–23.

Jokinen, K., & Wilcock, G. (2021, August). Do you remember me? Ethical issues in long-term social robot interactions. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication* (RO-MAN) (pp. 678-683). IEEE.

Lee, S.P., Badler J.B., & Badler, N.I. (2002) Eyes alive, ACM Trans. Graph. 21, 637-644.

Lehmann, H., Roncone, A., Pattacini, U., & Metta, G. (2016). Physiologically inspired blinking behaviour for a humanoid robot. In *Social Robotics: 8th International Conference, ICSR 2016, Kansas City, MO, USA, November 1-3, 2016 Proceedings* 8 (pp. 83-93). Springer International Publishing.

Lehmann, H., Keller, I., Ahmadzadeh, R., & Broz, F. (2017). Naturalistic conversational gaze control for humanoid robots-a first step. In *Social Robotics: 9th International Conference, ICSR 2017, Tsukuba, Japan, November 22-24, 2017, Proceedings 9* (pp. 526-535). Springer International Publishing.

Lehmann, H. (2020). *Social Robots for Enactive Didactics*. Franco Angeli.

Lehmann, H., & Rossi, P.G. (2020). Gestures in Educational Behavior Coordination. Grounding an Enactive Robot-Assisted Approach to Didactics. *Modelling Human Motion: From Human Perception to Robot Design*, 315-334.

Lehmann, H. and Svarny, P. (2021). Using a social robot for different types of feedback during university lectures. *Education Sciences & Society* 12, 282-29.

Luria, M., Forlizzi, J., & Hodgins, J. (2018, August). The effects of eye design on the perception of social robots. In 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) (pp. 1032-1037). IEEE.

Meltzoff, A. N., & Brooks, R. (2017). Eyes wide shut: The importance of eyes in infant gaze-following and understanding other minds. In *Gaze-Following* (pp. 217-241). Psychology Press.

Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., Von Hofsten, C., Rosander, K., Lopes, M., Santos-Victor, J. and Bernardino, A. (2010) The iCub humanoid robot: An open-systems platform for research in cognitive development. Neural Networks, 23(8), pp.1125-1134.

Morris, D. (2015). *Bodytalk: A world guide to gestures*. Random House.

Mutlu, B., Kanda, T., Forlizzi, J., Hodgins, J., & Ishiguro, H. (2012). Conversational gaze mechanisms for humanlike robots. ACM *Transactions on Interactive Intelligent Systems* (TiiS), 1(2), 1-33.

Müller, C., Cienki, A., Fricke, E., Ladewig, S., McNeill, D., & Teßendorf, S. (2013). Body-language-communication. *An international handbook on multimodality in human interaction*, *1*(1), 131-232.

Nocentini, O., Fiorini, L., Acerbi, G., Sorrentino, A., Mancioppi, G., & Cavallo, F. (2019). A survey of behavioral models for social robots. *Robotics*, 8(3), 54.

Paiva, A., & Leite, I. (2011). Emotion modeling for social robots. In *AISB 2011 Symposium on Engineering Emotions in Agent Systems,* pp. 108-113.

Picard, R. W. (1997). *Affective computing*. MIT press.

Ragno, L., Borboni, A., Vannetti, F., Amici, C., & Cusano, N. (2023). Application of social robots in healthcare: Review on characteristics, requirements, technical solutions. *Sensors*, *23*(15), 6820.

Scheflen, A. E., & Scheflen, A. (1972). *Body language and the social order: Communication as behavioral control.* Prentice-Hall.

Seth, A. K., Prescott, T. J., & Bryson, J. J. (Eds.). (2011). *Modelling natural action selection*. Cambridge University Press.

Sevilla-Salcedo, J., Fernádez-Rodicio, E., Martín-Galván, L., Castro-González, Á., Castillo, J. C., & Salichs, M. A. (2023). Using Large Language Models to Shape Social Robots' Speech. *International Journal of Interactive Multimedia and Artificial Intelligence*, 8(3), 6-20.

Tomasello, M. (2010). *Origins of human communication*. MIT press.

Turkle, S. (2011) *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books.

VanVugt, M. & Kameda, T.(2012). Evolution and groups. *Group processes*, 297–332.

Verhelst, E., Janssens, R., Demeester, T., & Belpaeme, T. (2024, March). Adaptive Second Language Tutoring Using Generative AI and a Social Robot. In Companion of the 2024 ACM/IEEE

International Conference on Human-Robot Interaction, 1080-1084.

Wang, C., Hasler, S., Tanneberg, D., Ocker, F., Joublin, F., Ceravola, A., Deigmoeller, J. & Gienger, M. (2024). Large language models for multi-modal human-robot interaction. arXiv preprint arXiv:2401.15174.

Yoshikawa, Y., Shinozawa, K., Ishiguro, H., Hagita, N. & Miyamoto, T. (2006) The effects of responsive eye movement and blinking behaviour in a communication robot. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*. IEEE 4564–4569.

Zhao, K., Whalen, P. J. & Breazeal, C. L. (2017). Emotion modelling for social robot application: A review. *Artificial Intelligence Review*, 47(1), 64-88.

Zhao, X., Liang, X., & Yang, J. (2019). Emotional state recognition in human-computer interaction based on deep learning. *IEEE Access*, 7, 149273-149282.

Capitolo V – Il mercato dell'IA generativa

Gabriele Torri

L'avvento dell'intelligenza artificiale generativa è ancora nelle sue fasi iniziali, ma i potenziali effetti sul panorama economico sono molteplici e dirompenti. In questa sezione analizzeremo gli effetti sul contesto macroeconomico e al mercato del lavoro, e discuteremo poi gli aspetti chiave legati allo sviluppo del settore in relazione ai principali *player* e ai *driver* di crescita.

Nell'analisi degli effetti economici dell'IA, è importante per prima cosa chiarirne i contorni: da un lato l'automazione dei processi, l'uso di robot in ambienti di produzione, e l'applicazione di tecniche di machine learning sono da decenni strumenti fondamentali in azienda e oggetto di studio degli economisti. L'accelerazione nello sviluppo di tecniche innovative, ad esempio nell'ambito della computer vision per lo smistamento dei prodotti e il controllo di qualità, o lo sviluppo di algoritmi per la diagnosi e identificazione dei guasti è, seppur con le proprie specificità, in continuità con altre evoluzioni tecnologiche precedenti che hanno interessato il mondo industriale. D'altra parte, invece, l'introduzione dell'intelligenza artificiale generativa, ha il potenziale di introdurre dinamiche inedite per il mondo del lavoro, interessando settori tradizionalmente poco esposti ai processi di automazione tradizionale.

Una delle differenze sostanziali tra l'IA generativa e i tradizionali processi di automazione risiede nel fatto che la prima permette un'interazione più organica con le macchine e non richiede necessariamente di essere programmata (almeno nel senso tradizionale del termine) per imparare a svolgere nuovi compiti e, in prospettiva, prendere decisioni autonome. In questo senso, oltre a permettere di supportare i lavoratori - e in parte sostituirli - in compiti creativi o manageriali, può esprimere il proprio potenziale massimo nell'integrazione con altri sistemi di

intelligenza artificiale, permettendo un utilizzo più efficace di tali sistemi e contribuendo all'aumento della produttività del lavoro.

In questa sezione, prima presenteremo una panoramica della letteratura economica focalizzata su questo tema e l'effetto sul mercato del lavoro. Successivamente focalizzeremo l'attenzione sul mercato dell'IA e sugli asset fondamentali per lo sviluppo di applicazioni in questo ambito. Infine, presentiamo tre approfondimenti sul ruolo dei dati, sui modelli open-source, e sulla tematica dei costi e dei consumi energetici.

1. Quantificare l'impatto dell'IA

Le applicazioni di IA in azienda sono svariate e trasversali alle funzioni aziendali e ai settori, e portano con sé un profondo potenziale di evoluzione delle pratiche operative. Resta però una grande incertezza relativa alla previsione dell'impatto economico dell'IA e sui tempi della transizione. Lu and Zhou (2021) evidenziano come le previsioni sull'IA siano molto diversificate: un punto di vista ottimistico, condiviso tipicamente da società di consulenza, esalta i potenziali effetti benefici sulla produttività. Ad esempio, il report *The economic potential of generative AI* (McKinsey & Company, 2023) mostra come l'uso dell'IA generativa su 63 possibili casi d'uso e 16 linee di business potrebbe generare globalmente un valore compreso tra i $2,6 e i $4,4 miliardi all'anno considerando le nuove applicazioni di IA generativa, cifra che raddoppierebbe considerando anche l'integrazione di IA generativa nei software e nei processi utilizzati attualmente. Questi benefici si sommano ad ulteriori 11-17,7 miliardi di dollari generati dall'adozione di altre tecnologie di *machine learning*, *deep learning* e *analytics*, generando un impatto potenziale dell'IA nel suo complesso previsto tra i 17,1 e i 25,6 miliardi di dollari. Il valore generato dall'IA generativa riguarderebbe per il 75% l'ambito delle *customer operations*, marketing e vendite, *software engineering* e ricerca e sviluppo. Lo stesso report stima il potenziale incrementando della produttività del lavoro tra lo 0,1% e lo 0,6% l'anno da qui al 2040, a condizione che le aziende

investano in queste tecnologie e nella gestione delle risorse umane. Un report di PWC stima il contributo potenziale dell'IA sull'economia globale in $15,7 miliardi nel 2030, con una crescita del 26% del PIL in alcune economie (PWC, 2017). Le stime della banca d'investimento Goldman Sachs (Hatzius, 2023), seppur più caute, prevedono che l'aumento di produttività legato all'IA possa portare ad una crescita di lungo periodo del PIL globale del 15% (gli autori riconoscono però che questo valore sia sovrastimato per un potenziale doppio conteggio dei benefici dell'IA e degli investimenti in tecnologie e comunicazione; un altro limite è legato alla tendenza di calo della produttività a cui si assiste in questi anni). In contrasto, analisi più legate al punto di vista dei *policy maker* tendono ad essere più caute e ad enfatizzare maggiormente i rischi relativi all'occupazione. Un esempio è il report redatto dal Institute for Public Policy Research (IPPR), secondo cui la diffusione dell'AI potrebbe mettere a rischio nel solo Regno Unito 8 milioni di posti di lavoro negli scenari più pessimistici (Jung & Desikan, 2024).

La letteratura macroeconomica è ancora in fase di evoluzione: la review condotta da Lu e Zhou (2021) mostra come l'integrazione dell'IA nei modelli analitici presenti sfide nuove per gli studiosi, e che l'impatto dell'IA sia sostanzialmente diverso da quello di altre tecnologie per il fatto di impattare un numero molto ampio di settori, con effetti aggregati particolari e sviluppi futuri imprevedibili. La relazione tra innovazione tecnologica e aumento della produttività potrebbe essere meno diretta ed immediata di quanto prospettato dalle fonti più ottimistiche. Si potrebbe riproporre infatti il ben noto paradosso postulato dall'economista Robert Solow, che nel 1987 enfatizzava come gli sviluppi tecnologici dell'informatica negli anni '80 non si fossero tradotti in benefici immediati in termini di aumento della produttività. Gli effetti dell'IA sul panorama economico dipenderanno sostanzialmente dalle direzioni prese dallo sviluppo tecnologico, dal panorama regolatorio, e dalle dinamiche di mercato. È possibile, ad esempio, pensare a due scenari contrastanti relativi alla produttività: il primo in cui L'IA si riveli in grado di

automatizzare un elevato numero di mansioni per la maggior parte dei lavoratori, complementando le loro abilità e liberando tempo da dedicare a ruoli creativi e alla risoluzione di problemi. In tale scenario è possibile prevedere un incremento sostanziale della produttività, e un tasso di crescita permanente più alto. In uno scenario alternativo potremmo ipotizzare che l'adozione di nuove tecnologie di IA in azienda sia più lenta, e che il potenziale di alcune tecnologie sia minore di quanto prospettato. In tale scenario l'IA potrebbe essere utilizzata principalmente per il risparmio di manodopera non qualificata, portando alla distruzione di posti di lavoro con effetti negativi sul sistema economico, controbilanciando potenziali effetti sulla produttività (Brynjolfsson & Unger, 2023).

Allo stesso modo, gli effetti dell'IA sulla diseguaglianza e sulla concentrazione industriale non possono ancora essere previsti con certezza e a seconda delle assunzioni postulate è possibile ipotizzare scenari molto divergenti tra loro. La letteratura empirica sta iniziando ad esplorare le relazioni tra IA e disoccupazione, IA e diseguaglianze, IA ed educazione, ed IA e commercio. Nella pratica aziendale e nella letteratura scientifica, ma anche nella giurisprudenza - si veda l'*AI Act* - resta comunque una grande variabilità nelle definizioni di IA adottate, che suggerisce attenzione nell'analisi e nel confronto delle fonti.

2. Effetti dell'IA sul mercato del lavoro

Sebbene difficili da prevedere, gli effetti dell'IA sul mercato del lavoro si preannunciano estremamente rilevanti e diversificati tra settori e mansioni. Nel panorama industriale le applicazioni di IA classiche restano predominanti, ma lo sviluppo dell'IA generativa è sempre più oggetto di analisi, visto anche il potenziale impatto su ambiti non interessati dai tradizionali processi di automazione tecnologica.

Grazie alla possibilità di interagire con l'IA mediante testo per ottenere la produzione di nuovo testo (o di immagini e musica per produrre nuove immagini e musica – e, a breve, video), l'IA

generativa sta rivoluzionando mansioni tradizionalmente ritenute immuni all'automatizzazione digitale quali la scrittura creativa, il marketing e la produzione di codice e software (attività in cui gli LLM eccellono). Questa tecnologia offre la possibilità agli analisti di realizzare elaborazioni dati anche complesse scrivendo pochissimo codice e persino al management di ottenere analisi di dati in via del tutto automatica senza ricorrere ad un analista. Investe a vari livelli tutti i *knowledge workers* (tra i quali figurano, oltre agli analisti, anche i giornalisti e gli accademici) e chi si occupa di produzione artistica (scrittori, grafici, sceneggiatori, …), senza risparmiare i segmenti dell'insegnamento e della ricerca. Più che alla completa sostituzione di talune mansioni lavorative (tipica dell'automazione industriale), è sempre più chiaro come l'IA generativa stia portando, almeno per il momento, a una rapida trasformazione delle mansioni stesse senza pregiudicarne la rilevanza. La chiave di lettura di questo fenomeno sta nella natura stessa di questa tecnologia che, seppur estremamente potente, è una tecnologia strutturalmente "imprecisa", capace cioè di produzioni all'apparenza verosimili ma che, in ultima analisi, possono rivelarsi contenutisticamente del tutto errate.

Diversi studi provano a stimare l'impatto dell'IA generativa sul mercato del lavoro: un'analisi di Goldman Sachs Economic Research stima per ogni mansione e settore la potenziale esposizione a tecnologie di IA generativa in grado di aumentare la produttività del lavoro (Hatzius, 2023). Secondo queste stime, negli Stati Uniti due terzi delle occupazioni potrebbe essere in parte automatizzato grazie all'IA generativa, e che per buona parte di questi l'automatizzazione interesserebbe una componente significativa del carico di lavoro (25-50%). Confrontando i settori produttivi, l'automatizzazione tramite IA interesserebbe in misura maggiore le attività di supporto burocratico e amministrativo, l'ambito legale, architettura & ingegneria, scienze della vita, fisiche e sociali, l'ambito commerciale e finanziario. Le professioni meno interessate sono invece quelle di pulizia e manutenzione di edifici e giardini, installazione, manutenzione e riparazione, costruzione & settore

estrattivo, produzione e trasporti. La divisione per mansioni mostra in modo evidente come le dinamiche occupazionali dell'IA generativa siano diverse da quelle legate all'occupazione tradizionale, interessando in larga parte gli impiegati d'ufficio, i professionisti e i manager, mentre riguarda in misura molto minore gli operai e il personale impiegato in mansioni elementari. In aggregato, gli autori stimano che l'IA generativa potrebbe portare all'automazione del 18% del lavoro globale. Secondo questo studio, gli ambiti nei quali è più probabile che l'IA sostituisca invece di complementare il lavoro sono quello legale e quello delle attività di supporto burocratico e amministrativo. Gli effetti aggregati sulla produttività devono tener conto non solo degli effetti sui singoli settori, ma anche del ricollocamento dei lavoratori tra mansioni. Gli analisti di Goldman Sachs stimano che una diffusione estesa dell'IA generativa potrebbe dare un *boost* alla crescita della produttività aggregata di 1,5% annuo (con una grande incertezza della stima relativa alla potenza dei modelli IA sviluppati, alla quota di lavoratori ricollocati, e alla velocità di adozione delle tecnologie in azienda).

Un'analisi simile è stata condotta da Felten et al. (2023), che sviluppano una metodologia in grado di valutare in modo sistematico gli effetti sulle diverse professioni e settori industriali dei modelli IA, inclusi gli LLM come ChatGPT. Gli autori trovano che le professioni più esposte sono quelle legate ad attività di telemarketing e all'insegnamento, mentre in aggregato i settori più coinvolti sono i servizi legali e l'ambito finanziario e assicurativo.

La rilevanza dell'IA per il mondo del lavoro emerge anche dalla survey Future of Jobs 2023 condotta periodicamente dal World Economic Forum e riguarda più di 800 aziende che impiegano in totale 11,3 milioni di lavoratori (Di Battista et al., 2023). Emerge come le aziende siano divise riguardo la stima degli effetti dell'IA sull'occupazione, con una maggioranza di imprese intervistate convinte di un effetto positivo sulla creazione di posti di lavoro, ma con una sostanziale quota di intervistati che si aspettano effetti negativi dal punto di vista dell'occupazione. L'analisi

mostra inoltre una notevole diversificazione tra settori per quanto riguarda i piani di investimento in queste tecnologie: in alcuni settori come servizi IT ed elettronica più del 90% delle compagnie prevede di adottare soluzioni IA entro il 2027, mentre in altri settori quali agricoltura, silvicoltura e pesca, alloggio, cibo e tempo libero, ed estrazione e metallurgia solo il 60% circa delle imprese pianifica di adottare soluzioni IA. Questi dati sono fortemente correlati con le attività di formazione su IA e big data, segno che l'adozione di soluzioni IA va di pari passo con lo sviluppo delle risorse umane.

Infine, i dati raccolti da Lightcast e riportati nell'Artificial Intelligence Index Report 2024 confermano come le competenze IA siano estremamente richieste sul mondo del lavoro, mostrando un trend di crescita negli ultimi 5 anni, nonostante una flessione nel 2023 (negli Stati Uniti le offerte di lavoro che richiedono competenze in IA scendono dal 2% del totale nel 2022, al 1,6% nel 2023), Questa flessione è dovuta ad un ridimensionamento dell'offerta da parte dei grandi player del settore e ad una rimodulazione dei ruoli tecnici in queste compagnie. Le offerte di lavoro sono concentrate in particolare nei settori dell'informazione, servizi professionali, scientifici e tecnici, finanza e assicurazione, e manifatturiero (Artificial Intelligence Index Report 2024, capitolo 4).

3. I *player* del mercato dell'IA

Se da un lato l'IA ha effetti potenzialmente dirompenti per tutti i settori economici, lo sviluppo delle tecnologie necessarie per l'implementazione di questa transizione economica vede protagoniste un selezionato gruppo di compagnie del settore tecnologico che sviluppano soluzioni *innovative* o integrano modelli di IA generativa nei processi industriali. Il settore è composto da nuovi player come OpenAI (La società alla base dello sviluppo di ChatGPT, con un valore stimato ad oggi di 80 miliardi

di dollari)[1], e grandi compagnie del settore tecnologico come Microsoft, Google, Nvidia e Amazon (queste ultime interessate sia allo sviluppo e integrazione di tecnologie IA, sia alla fornitura di hardware e soluzioni di cloud computing). La crescita passa sia attraverso lo sviluppo di tecnologie interne, sia (per i grandi player del settore) tramite partnership e acquisizioni strategiche (Microsoft e OpenAI). Cresce poi un ecosistema di società e sviluppatori indipendenti specializzati nel fornire servizi specifici, grazie anche allo sviluppo di piattaforme come GPTs store, che permette di creare versioni customizzate del noto chatbot di OpenAI focalizzate su compiti specifici, e alla crescente disponibilità di modelli aperti pre-addestrati (come, ad esempio, i modelli Llama di Meta), che possono essere affinati e adattati a compiti specifici.

Per quanto riguarda i Large Language Models (LLM), e i modelli LLM multimodali (MM-LLM) l'evoluzione del mercato è difficile da prevedere: OpenAI con ChatGPT-4 al momento è il modello più noto al grande pubblico, ma lo sviluppo di diversi concorrenti come Gemini di Google, Grok di xAI, o Claude di Anthropic mostrano il dinamismo del settore, e aprono le porte a diversi scenari competitivi. Si stanno consolidando diversi modelli di monetizzazione, e resta aperto il tema del *pricing*, funzione sia del livello di competizione nel mercato, sia degli elevati costi di addestramento e gestione dei modelli di AI: i principali player propongono ad oggi un'offerta segmentata che prevede tipicamente diversi pacchetti a pagamento e piani gratuiti, differenziati per l'accesso ai modelli, limiti di utilizzo, e servizi aggiuntivi. Alcune aziende come Microsoft e Adobe puntano sull'integrazione con le proprie suite di software (rispettivamente Copilot e Firefly), altre sulla proposta di strumenti esterni come ChatGPT-4 (seppur integrabili in applicazioni *custom* mediante un'API).

A causa degli alti costi di training e gestione, gli LLM più grandi come ChatGPT-4 potrebbero non essere la soluzione adatta ad

[1] https://www.reuters.com/technology/openai-valued-80-billion-after-deal-nyt-reports-2024-02-16

ogni applicazione, lasciando spazio a modelli più piccoli e specializzati. Il training di modelli personalizzati sviluppati a partire da modelli open-source come Llama di Meta potrebbe risultare più efficiente rispetto a modelli più grandi e complessi, riducendo i costi notevolmente e rendendo il settore alla portata di player più piccoli (MIT Technology Review Insights, 2023). L'utilizzo di modelli più piccoli funzionanti su server locali permette inoltre di evitare problemi legati alla condivisione di informazioni riservate con soggetti terzi.

Dal punto di vista geografico, gli Stati Uniti sono ancora leader globale in termini di investimenti e sviluppo di LLM e modelli IA multimodali: nel 2023 il 61% dei più modelli più grandi è sviluppato da aziende americane (Artificial Intelligence Index Report 2024, Capitolo 4), anche se la leadership sulla ricerca IA è sfidata dalla Cina. Per quanto riguarda gli investimenti, il mercato statunitense è di gran lunga il più sviluppato, seguito dalla Cina e dall'Europa. La tendenza globale degli investimenti in IA ha visto una costante crescita fino al 2021, e ha registrato una flessione nel corso del 2022 e del 2023, con un valore per il 2023 pari a 96 miliardi di dollari. Nel 2023 si assiste però ad un vero boom degli investimenti in IA generativa, che raggiungono a livello globale 25 miliardi di dollari, quasi 9 volte maggiori rispetto all'anno precedente. Nel 2023 i settori di maggiore interesse per gli investitori sono quello dell'infrastruttura/governance/ricerca IA, il customer support, e il data management e processing.

4. Gli asset dell'IA: imprese e università

Lo sviluppo e gestione di modelli di IA richiede tre asset chiave: personale altamente qualificato, potenza di calcolo, e dati. Questa convergenza di fattori necessari ha avuto effetti sostanziali sulla direzione dello sviluppo e ricerca su modelli IA, spostando competenze e risorse dal mondo accademico a quello dell'industria. I dati riportati da Ahmed et al (2023) relativi alle università americane mostrano come il 70% dei nuovi PhD in computer science specializzati in IA nel 2020 cerchi lavoro nelle

imprese, mentre nel 2004 solo il 21% seguiva questa strada. La situazione è simile nel Regno Unito, e questo spostamento di risorse mette in difficoltà la ricerca accademica. La distanza fra accademia e industria si misura inoltre nella potenza di calcolo a disposizione, che permette all'industria di avere vantaggi competitivi, e la disponibilità di dati, fondamentali per il training dei modelli. La prevalenza dell'industria è visibile nell'evoluzione di paper sull'IA co-autorati da ricercatori affiliati alle aziende, che è cresciuta dal 22% del 2000 al 38% del 2020 (Ahmed et al, 2023) e dalla percentuale dei più grandi modelli di IA sviluppati dall'industria, che passa dall'11% del 2010 al 96% del 2021. Un tale spostamento di risorse verso l'industria ha permesso un'accelerazione notevole nello sviluppo dei modelli, ma al tempo stesso presenta rischi legati alla concentrazione di risorse su modelli di business for-profit, non necessariamente allineati agli interessi della collettività. Le dinamiche di sviluppo future del settore dipenderanno in larga parte dalle scelte strategiche intraprese dai regolatori e dalla loro capacità di gestire il settore dell'AI intrecciando considerazioni tecniche, economiche, e politiche.

Emergono alcuni aspetti fondamentali legati alla gestione e allo sviluppo di modelli IA, che analizziamo in seguito.

5. La legittimità dell'uso dei dati di *training*

La calibrazione dei modelli di intelligenza artificiale più evoluti ha bisogno di una quantità enorme di dati, che spesso sono utilizzati senza cura dei diritti d'autore. OpenAI, in un'audizione alla Camera dei Lord britannica, ha difeso le pratiche di *scraping* di testi e materiale multimediale da internet sostenendo che l'addestramento di modelli IA sarebbe impossibile senza l'utilizzo di materiale protetto da *copyright*[2]. Secondo questa posizione, l'uso per il training di materiale protetto da *copyright* ma disponibile online sarebbe legittimo (*fair use*). Questa posizione non è però universalmente condivisa, e l'uso di materiale protetto

[2] Accessibile al link: https://committees.parliament.uk/writtenevidence/126981/pdf

è terreno di battaglia di cause legali (si veda NYT vs OpenAI)[3] ed è sempre più nell'occhio di giuristi e regolatori (e.g. Torrance and Tomlinson, 2023). La questione ha ripercussioni non solo legali ma anche economiche.

Il tema della legittimità e qualità dei dati usati per l'addestramento si lega inoltre alla crescente presenza online di testi e materiali multimediali prodotti da modelli generativi. L'uso di questi materiali per l'addestramento di modelli pone due ordini di problemi: per prima cosa una crescente opacità relativa alle fonti primarie (per aver garanzia di un uso legittimo servirebbe infatti conoscere le fonti utilizzate dal modello che genera i dati). Il secondo aspetto riguarda la qualità dell'addestramento. Come dimostrato da Shumailov et al, (2023), l'uso di dati generati sinteticamente per l'addestramento di modelli generativi porta ad una riduzione della qualità degli output e potenzialmente al collasso dei modelli allenati con questi dati. A questo si aggiunge il problema dei tentativi deliberati di "avvelenamento" dei dati da parte di creatori di contenuti online, volti a contrastare il fenomeno dell'utilizzo illegittimo delle proprie creazioni, che possono portare a malfunzionamenti caotici ed imprevedibili dei modelli allenati su tali dati[4].

Per far fronte a questi problemi iniziano ad emergere pratiche di mercato più attente alla tutela dei diritti d'autore e all'uso di materiale protetto per il training. È oggi disponibile una certificazione, emessa dalla non-profit *fairly-trained*[5] che attesta come modelli di IA generativa non siano stati addestrati con materiale protetto da *copyright* senza un'autorizzazione. A marzo 2024 la certificazione è stata attribuita a 14 modelli tra cui un LLM (KL3M). Interessante sotto questo fronte citare la *partnership* tra

[3] Accessibile al link: https://www.ansa.it/sito/notizie/mondo/2023/12/27/new-york-times-fa-causa-a-openai-e-microsoft-sulluso-del-copyright_256dc4cf-0f1b-49bc-a69e-f27a0fad9cea.html

[4] Si veda ad esempio il software Nightshade, che permette di "avvelenare" le proprie immagini con modifiche impercettibili all'occhio umano, ma in grado di confondere i modelli di generazione di immagini (accessibile al link: https://www.technologyreview.com/2023/10/23/1082189/data-poisoning-artists-fight-generative-ai)

[5] Accessibile al link: https://www.fairlytrained.org

Google e StackOverflow (uno dei più noti siti web di domande e risposte in ambito informatico e programmazione) volta ad utilizzare il materiale del sito per l'addestramento di Gemini, il chatbot di Google. Questo tipo di accordi mostra come sia possibile sviluppare modelli di business in grado di conciliare la tutela del *copyright* con le esigenze degli sviluppatori[6].

6. LLM e *open source*

Tradizionalmente, per *open source* si intende un *software* dove il codice sorgente è distribuito secondo una licenza che ne concede lo studio, l'utilizzo, la modifica e la redistribuzione. L'estensione di tale approccio allo sviluppo dell'IA generativa potrebbe rappresentare un aspetto chiave per uno sviluppo del settore più equo e trasparente, accelerando al tempo stesso l'innovazione e l'adozione di queste tecnologie. In riferimento all'IA generativa, la definizione di *open source* diventa però più problematica, ed emergono diverse interpretazioni del termine: a differenza del *software* tradizionale, dove la definizione di *open source* riguarda strettamente il codice sorgente (ovvero le istruzioni che il computer esegue per far funzionare il *software*), un modello IA ha una struttura ben più complessa, e l'output generato dipende non solo dal codice usato, ma anche dai dati utilizzati per l'addestramento, il codice usato per processare i dati, e una serie di altri dettagli tecnici[7]. Quali di questi componenti debbano essere resi disponibili in un modello *open source* è oggetto di dibattito, e il principale tema di discussione è quello della condivisione dei dati per l'addestramento.

L'interpretazione più diffusa al momento è quella di considerare *open source* modelli IA pre-addestrati senza i dataset di riferimento, permettendo agli utenti di specializzare i modelli secondo le necessità attraverso procedure di *fine-tuning* (un

[6] Accessibile ai link: https://www.wired.com/story/google-deal-stackoverflow-ai-giants-pay-for-data/ e https://stackoverflow.blog/2024/03/12/how-stack-overflow-is-partnering-with-google-to-encourage-socially-responsible-ai.

[7] Accessibile al link: https://www.technologyreview.com/2024/03/25/1090111/tech-industry-open-source-ai-definition-problem.

esempio è il modello Llama 2 sviluppato da Meta). Un importante soggetto per la promozione dell'innovazione *open* nel mondo dell'IA è la AI Alliance, una comunità di soggetti volta ad accelerare l'innovazione open nel mondo dell'IA che comprende grandi aziende come Meta, IBM, e numerose altre compagnie, organizzazioni non-profit e università.[8] Questa comunità, grazie al supporto di molti dei principali attori nel panorama dell'IA, si pone come uno dei principali forum per lo sviluppo di definizioni e approcci condivisi sull'intelligenza artificiale *open source*.

Esistono altri tentativi di standardizzazione e definizione, citiamo in particolare l'Open Source Initiative (OSI), la principale organizzazione di promozione del software open source, che dal 2022 ha lanciato l'iniziativa Deep Dive al fine di elaborare i princìpi fondativi di una definizione condivisa, in un processo analogo a quello che ha portato alla definizione di software open source a partire dai princìpi del Manifesto GNU[9].

7. Costi e impatto ambientale

La crescita del numero di parametri e della complessità dei modelli a cui abbiamo assistito negli ultimi anni ha portato ad un'esplosione dei costi di addestramento dei modelli dovuti ai consumi energetici e all'acquisto/noleggio dell'*hardware*. Secondo le stime basate sui tempi e le modalità di addestramento riportate nell'AI index report 2024, l'addestramento di grandi modelli generativi può costare cifre nell'ordine delle decine o centinaia di milioni di euro, ad esempio per ChatGPT-4 si stima una spesa di 78 milioni di dollari, mentre per Gemini Ultra addirittura di 191 milioni. I costi necessari sono probabilmente ancora più alti vista la crescente complessità di questi e i regimi di addestramento più complessi e onerosi che includono, oltre alla pura esecuzione computazionale, l'intervento umano (si pensi al *Reinforcement Learning with Human Feedback*, o RLHF). Queste cifre sono significativamente maggiori rispetto a modelli risalenti

[8] Accessibile al link: https://thealliance.ai/
[9] Accessibile al link: https://opensource.org/deepdive

a soli pochi anni prima, ad esempio nel 2020 l'addestramento di ChatGPT-3 davinci ha richiesto secondo le stesse stime 4 milioni di dollari, e nel 2018 BERT-Large (il modello di apprendimento automatico di Google, predecessore dei LLM attuali) ha richiesto un costo di addestramento di circa 3000 dollari. Gli alti costi creano delle forti barriere all'ingresso che penalizzano i player più piccoli e il mondo accademico, e creano un mercato per modelli più piccoli, pre-addestrati e personalizzabili.

Ricordiamo infine che, visto l'elevato consumo energetico, i modelli di IA generativa hanno un rilevante impatto ambientale. Si stima che il training di ChatGPT-3 abbia richiesto 1287 MWh di energia corrispondenti all'emissione di più di 550 tonnellate di anidride carbonica, e verosimilmente i modelli più recenti molto di più visto il maggiore costo computazionale (MIT Technology Review Insights 2023). Wu et al. (2022) che sviluppa una procedura per la valutazione dell'impronta carbonica di modelli di intelligenza artificiale. Vista la crescente complessità e dimensione dei modelli, in prospettiva questi temi diventeranno sempre più rilevanti, e lo sviluppo di algoritmi e hardware più efficienti sarà uno dei nodi chiave per lo sviluppo del settore.

Bibliografia

Ahmed, N., Wahed, M., & Thompson, N. C. (2023). The growing influence of industry in AI research. *Science*, *379*(6635), 884-886.

Brynjolfsson E., Unger G. (2023). The Macroeconomics of Artificial Intelligence. *F&D - A Quarterly Publication of the International Monetary Fund*.

Di Battista, A., Grayling, S., Hasselaar, E., Leopold, T., Li, R., Rayner, M., & Zahidi, S. (2023). Future of jobs report 2023. In World Economic Forum, Geneva, Switzerland.

https://www.weforum.org/reports/the-future-of-jobs-report-2023.

Felten, E., Raj, M., & Seamans, R. (2023). How will Language Modelers like ChatGPT Affect Occupations and Industries?. *arXiv preprint arXiv:2303.01157*.

Hatzius, J. (2023). The Potentially Large Effects of Artificial Intelligence on Economic Growth (Briggs/Kodnani). *Goldman Sachs*.

Jung, C., & Desikan, B. S. (2024). Transformed by AI: how generative artificial intelligence could affect work in the UK–and how to manage it. https://ippr-org.files.svdcdn.com/production/Downloads/Transformed_by_AI_March24_2024-03-27-121003_kxis.pdf

Lu, Y., & Zhou, Y. (2021). A review on the economics of artificial intelligence. *Journal of Economic Surveys*, 35(4), 1045-1072.

McKinsey & Company (2023), The economic potential of generative AI https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/the%20economic%20potential%20of%20generative%20ai%20the%20next%20productivity%20frontier/the-economic-potential-of-generative-ai-the-next-productivity-frontier.pdf

MIT Technology Review Insights (2023). *The great acceleration: CIO perspectives on generative AI*, MIT Technology Review in collaborazione con Databricks.

PWC (2017), Sizing the prize What's the real value of AI for your business and how can you capitalise? https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023). The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Torrance, A. W., & Tomlinson, B. (2023). Training is everything: Artificial intelligence, copyright, and fair training. *arXiv preprint* arXiv:2305.03720.

Wu, C. J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., ... & Hazelwood, K. (2022). Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems*, 4, 795-813.

Capitolo VI – IA, industria e organizzazione aziendale

Roberto Sala
Gabriele Torri

Negli ultimi anni, gli investimenti in Intelligenza Artificiale (IA) sono aumentati in maniera considerevole (OECD.AI, 2024), trascinati soprattutto da quelli in IA generativa che, secondo l'Artificial Intelligence Index Report 2024 (Stanford University, 2024), ha raccolto il triplo degli investimenti privati rispetto al 2022, arrivando a 25,2 miliardi di dollari. Sempre secondo l'Artificial Intelligence Index Report 2024, sono gli Stati Uniti ad investire maggiormente in questa tecnologia (62,5 miliardi di dollari nel 2023), seguiti da Cina e Unione Europea. Da un rapporto della Corte dei Conti emerge che l'UE ha stabilito come obiettivo per gli investimenti pubblici e privati il valore di 20 miliardi di euro l'anno fino al 2030 (European Court of Auditors, 2024). Inoltre, l'UE prevede di stanziare finanziamenti per la ricerca sull'IA per 7 miliardi di euro nel periodo 2021-2027 tramite *Horizon Europe* e il *Digital Europe programme* (European Commission, 2024; European Court of Auditors, 2024).

Rispetto al panorama europeo, il rapporto "*Statistics Explained: Use of artificial intelligence in enterprises*" disponibile su sito Eurostat (Eurostat, 2024), chiarisce che circa l'8% delle imprese intervistate usa l'IA all'interno dei propri processi. Da sottolineare il fatto che le aziende localizzate nel centro e Nord Europa sembrano essere quelle con il livello di adozione più alto (Figura 1).

Nel caso dell'Italia, questa percentuale si ferma al 5%, dato che viene confermato anche da Unioncamere a inizio marzo 2024 (Unioncamere, 2024) come risultato di un'indagine svolta tramite 40000 test di autodiagnosi realizzati con il supporto dei Punti impresa digitale delle Camere di commercio, che afferma che

meno del 10% delle aziende sta utilizzando l'IA all'interno dei propri processi.

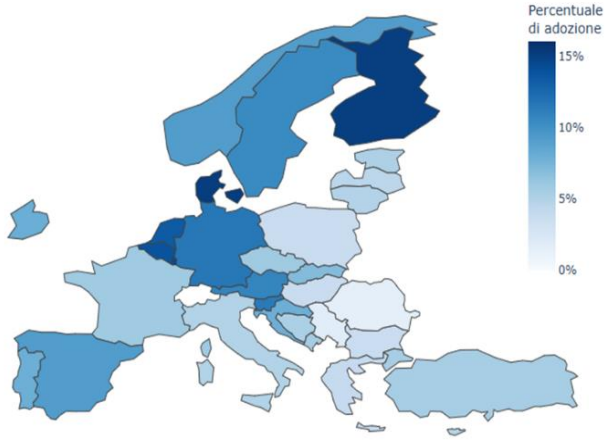

Figura 1 - Percentuale di aziende che usano l'IA nei paesi coinvolti nella survey Eurostat. Elaborazione degli autori su dati (Eurostat, 2023, 2024).

Sempre secondo Unioncamere, un altro 15% delle aziende rispondenti intende investire in applicazioni IA entro 3 anni, confermando quindi il trend di crescita che è in atto dal 2021. Trend positivo confermato anche dal numero crescente di aziende che hanno digitalizzato uno o più processi dimostrando quindi una buona o elevata autonomia digitale (Unioncamere, 2024). La percentuale indicata da Eurostat sembra quindi essere destinata a crescere nei prossimi anni.

In termini di dimensione delle aziende, a livello europeo, si dimostrano più avanti nell'adozione di tecnologie IA quelle grandi, con percentuali che variano dal 7% al 16,4% a seconda tecnologia, mentre quelle piccole si dimostrano indietro rispetto alle altre, con percentuali che oscillano tra 0,6% e 2,3% (Figura 2).

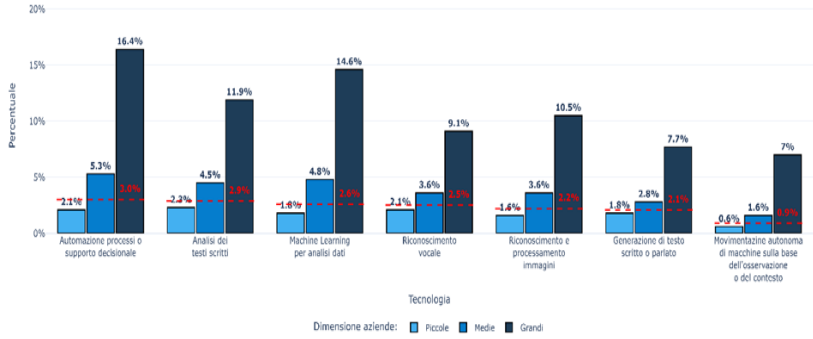

Figura 2 - Percentuale di aziende che usano l'IA per dimensione e tecnologia nei paesi coinvolti nella survey Eurostat. Elaborazione degli autori su dati (Eurostat, 2023, 2024).

In generale, nel 2023 le applicazioni più diffuse per quanto riguarda l'IA (Figura 3) sono state legate alla sicurezza informatica (26,2%), gestione finanziaria (25,8%) e ottimizzazione dei processi produttivi (24,9%), con percentuali che variano poi a seconda del settore specifico e della dimensione delle aziende.

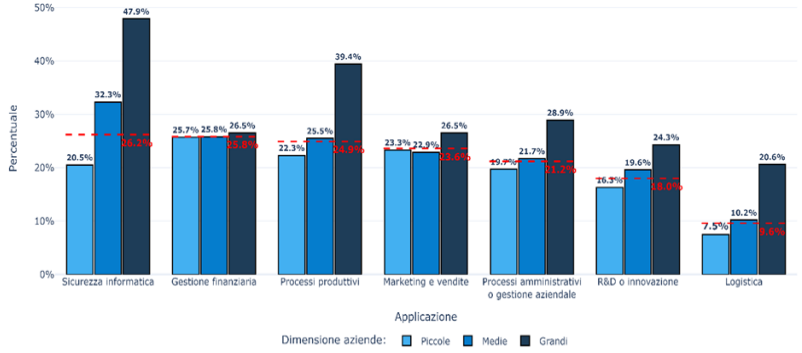

Figura 3 - Percentuale di aziende che usano l'IA per dimensione e applicazione nei paesi coinvolti nella survey Eurostat. Elaborazione degli autori su dati (Eurostat, 2023, 2024).

Contestualizzando quindi il discorso a livello italiano, il report "Intelligenza Artificiale in Italia: La rivoluzione che sta cambiando il business" di LUISS e Minsait presentato a maggio 2024 chiarisce che le applicazioni principali riguardano processi produttivi, marketing, assistenza clienti, sviluppo prodotti e servizi. Ad oggi si evidenzia un uso orientato al supporto decisioni per i flussi di lavoro e l'analisi dati, con una recente diffusione di strumenti di generazione di testo scritto o parlato, in questo caso soprattutto a supporto di processi che richiedono creatività (LUISS & Minsait, 2024).

Di fatto, l'IA svolgerà in futuro un ruolo significativo e il settore manifatturiero è tra quelli che, secondo il rapporto "*R&I Priorities for enhancing Artificial Intelligence applications in Manufacturing in Lombardy*" (AFIL et al., 2024), si prevede saranno maggiormente colpiti. Gli investimenti dell'industria manifatturiera globale in software, hardware e servizi di IA, che dovrebbero crescere da 2,9 miliardi di dollari nel 2018 a 13,2 miliardi di dollari entro il 2025 (9,5 miliardi di dollari nel 2021), mostrano l'importanza di queste tecnologie. In questo caso, le applicazioni principali riguardano il processamento dati e il controllo qualità tramite computer vision, con un'adozione di chatbot relativamente bassa ma in crescita.

È bene sottolineare che, vista la complessità dei processi interni, le aziende tendono a introdurre gradualmente delle novità che possono impattare in maniera considerevole sugli stessi. È quindi coerente il fatto che, in tutti i report fin qui menzionati, emerga la tendenza delle aziende a creare dei casi d'uso più o meno piccoli prima di integrare l'IA in maniera rilevante all'interno dei propri processi. In particolare, tramite progetti pilota, l'idea è quella di fare una prima verifica delle potenzialità e, soprattutto, delle limitazioni (es. competenze, costi) che potrebbero impedire un impatto ottimale dell'IA, sia che si tratti di IA tradizionale che di IA generativa.

In generale, si desume che la previsione è quella per cui l'introduzione di tecnologie legate all'IA all'interno dei processi aziendali porterà a considerevoli benefici non solo in termini di

produttività ed efficienza ma aprirà le porte anche alla definizione di nuovi modelli di business. Nonostante ciò, come sottolineato dal report della Corte dei Conti Europea sulle ambizioni UE per l'IA, è necessario che vengano definite delle politiche mirate e delle azioni di controllo specifiche, in modo da monitorare al meglio l'utilizzo dei fondi stanziati e l'implementazione dell'IA *(European Court of Auditors*, 2024).

1. IA e nuovi modelli di business

L'adozione dell'IA all'interno delle aziende ha il potenziale di abilitare nuovi modelli di business, ovvero l'insieme di logiche che definisco come un'azienda crea, distribuisce e raccoglie il valore. Vi sono molti aspetti che devono essere considerati nel momento in cui si definisce un nuovo modello di business, spesso visualizzati sulla base del Business Model Canvas proposto da Osterwalder (Osterwalder & Pigneur, 2010). Tra gli aspetti principali da considerare nella fase di definizione di un business model vi sono la "creazione del valore", la "distribuzione del valore" e la "cattura del valore" (Bocken et al., 2014). In particolare, la "creazione del valore" fa riferimento all'insieme di prodotti e servizi che un'azienda sviluppa per rispondere alle esigenze di un cliente. La "distribuzione del valore" comprende tutte le attività e risorse che un'azienda utilizza per fornire il valore al cliente. Con "cattura del valore" si intende invece la struttura di costi e ricavi che permettono all'azienda fornitrice di monetizzare l'offerta di prodotti e/o servizi offerti al cliente.

Negli ultimi decenni, nel panorama industriale, soprattutto in quello manifatturiero, i servizi hanno assunto sempre maggior importanza, permettendo alle aziende di creare nuove opportunità di business sfruttando la combinazione di prodotti e servizi offerti ai clienti in un fenomeno chiamato servitizzazione (Vandermerwe & Rada, 1988). In particolare, la servitizzazione viene descritta come il processo per cui le aziende, tradizionalmente prodotto-centriche, aggiungono servizi alla loro offerta per diventare più competitive, incrementare i ricavi, differenziarsi dalla concorrenza e occupare maggiori fette di

mercato. La realizzazione, in termini di offerta di mercato, della servitizzazione è associata ai sistemi prodotto-servizio, cioè delle combinazioni di prodotti e servizi, offerti in vari mix, che permettono di creare legami di lungo termine tra aziende fornitrici e clienti, generando maggior valore per entrambi gli attori coinvolti (Mont, 2002). Tradizionalmente i sistemi prodotto-servizio sono classificati in product-oriented, use-oriented, e result-oriented (Tukker, 2004). Queste tre categorie, che a loro volta contengono delle sottocategorie, sono contraddistinte da diverse politiche sulla proprietà del prodotto, sul modello di ricavi, e sull'assunzione di rischio all'interno dell'offerta, permettendo quindi di ampliare l'offerta ai clienti adattandola alle loro necessità. Più recentemente, con la diffusione dell'Industria 4.0, sempre più aziende hanno approcciato il tema della digitalizzazione dei processi e dell'analisi dati, aprendo quindi a un'ulteriore trasformazione, che ha portato a parlare non più solo di servitizzazione ma di servitizzazione digitale, basata quindi su offerte di business abilitate da servizi digitali (Pirola et al., 2020).

In particolare, si è iniziato anche a parlare di come l'IA può essere utilizzata all'interno delle offerte di prodotti e servizi per creare nuovi modelli di business, molto spesso orientati alla circolarità e sostenibilità grazie all'ottimizzazione dell'utilizzo delle risorse e alla riduzione gli sprechi (Romero et al., 2023). L'integrazione dell'IA offre opportunità di "creazione di valore" attraverso una maggiore efficienza, riduzione dei costi, crescita dei ricavi e miglioramento del processo decisionale (Åström et al., 2022; Toorajipour et al., 2024). Abilita anche nuove tipologie di "distribuzione del valore", che possono sfruttare connessioni remote e catene di fornitura ottimizzate grazie all'analisi dati. Oltre a ciò, è anche possibile definire nuovi metodi di "cattura del valore", basati ad esempio sull'utilizzo delle funzionalità IA o sul risultato dell'utilizzo delle funzionalità IA (sempre in ottica di servitizzazione ad esempio) (Åström et al., 2022). Questo ovviamente senza tralasciare il continuo perfezionamento delle soluzioni di IA, necessario per adattarle ai progressi tecnologici e alle esigenze di sostenibilità. Nella servitizzazione digitale, l'IA può

creare e acquisire valore supportando i risultati dei clienti e consentendo nuovi flussi di entrate (Black et al., 2024). I servizi basati sull'IA ottimizzano l'uso e la manutenzione dei prodotti nelle operazioni dei clienti, migliorando la fornitura di valore. I meccanismi di acquisizione del valore garantiscono un'equa distribuzione dei profitti in tutta la catena di creazione del valore, affrontando le strutture dei costi e i potenziali flussi di entrate.

Fondamentalmente, quindi, l'IA offre due funzionalità critiche in grado di abilitare nuovi modelli di business: *augmentation* (miglioramento delle capacità e dei processi) e *automation* (automazione di attività e processi) (Sjödin et al., 2023).

Tramite l'*augmentation* è possibile ottimizzare l'utilizzo delle risorse, aumentare la produttività e migliorare l'efficienza operativa. Ad esempio, l'uso di veicoli a guida autonoma (*Automated Guided Vehicles - AGV*) abilitati dall'IA, la manutenzione predittiva, la gestione della flotta e i contratti di ottimizzazione degli impianti. Queste offerte mirano ad aumentare l'efficienza e le prestazioni di prodotti, flotte e impianti industriali sfruttando l'IA e l'analisi dati. Ciò comporta l'intensificazione e l'ottimizzazione dell'utilizzo delle risorse, l'estensione dei cicli di vita dei prodotti e l'ottimizzazione del sistema (Madanaguli et al., 2024).

Tramite l'*automation*, d'altra parte, è possibile utilizzare l'IA per automatizzare le attività di routine, aumentando l'efficienza e riducendo i costi. L'automazione può ridurre il consumo di risorse, intensificare l'utilizzo degli asset aziendali e prolungare la vita utile dei prodotti, riducendo l'impatto ambientale. I modelli di business di automazione basati sull'IA offrono vantaggi sostanziali, contribuendo a ridurre le spese operative e sostenendo principi di circolarità (Sjödin et al., 2023).

Questo abilita inoltre la possibilità di usare l'IA per ottimizzare l'efficienza e ridurre gli sprechi offrendo questi benefici come servizi per i clienti, permettendo così di confezionare offerte sostenibili e concorrenziali. Queste offerte sono abilitate dalle capacità "percettive", "predittive" e "prescrittive" dell'IA (Sjödin et al., 2023). Le capacità percettive dell'IA consentono ai produttori

di ottenere informazioni dettagliate sull'utilizzo dei prodotti nelle operazioni industriali. Le capacità predittive consentono alle aziende di analizzare dati multidimensionali, anticipare i cambiamenti e gestire in modo proattivo le risorse. Le capacità prescrittive aiutano a identificare le azioni ideali per massimizzare il valore sostenibile attraverso le simulazioni.

Nell'ottica di modelli di business sostenibili, è possibile considerare tre meccanismi chiave: "rallentare", "restringere" e "chiudere" i cicli delle risorse (Madanaguli et al., 2024). Il "rallentamento" comporta l'aumento della durata di vita dei prodotti e dei loro componenti attraverso funzionalità di IA predittiva e prescrittiva, come l'apprendimento automatico sui dati di utilizzo e manutenzione. Il "restringimento" si concentra sulla riduzione del consumo di materiali ed energia ottimizzando aspetti specifici della catena del valore attraverso l'IA. La "chiusura" permette di convertire parti di prodotti a fine vita in materiali che possono essere usati per nuovi prodotti, con nuovi cicli di vita, anche questa viene facilitata da progetti basati sull'IA che sfruttano design per il riutilizzo.

Nonostante la promessa dell'IA nel plasmare nuovi modelli di business, sono necessarie ulteriori ricerche per comprenderne appieno il potenziale (Romero et al., 2023). L'integrazione dell'IA con i principi dell'economia circolare nella servitizzazione digitale rimane un tema di discussione aperto e di interesse. Ciononostante, la capacità dell'IA di supportare pratiche sostenibili, un processo decisionale intelligente e la manutenzione predittiva la posiziona come un fattore chiave per la creazione di modelli di business innovativi, efficienti e sostenibili (Romero et al., 2023).

2. Applicazioni

Come spiegato in precedenza, la diffusione di applicazioni di IA nel campo industriale è in forte crescita grazie alla possibilità di supportare l'ottimizzazione processi, ridurre gli sprechi, e generare nuovo valore. Se, come naturale, in un primo momento le applicazioni si sono basate sull'IA tradizionale, oggi l'interesse

verso l'IA generativa è sempre maggiore, viste anche la crescente popolarità di strumenti legati a questa nuova tipologia di IA quali ChatGPT o Gemini (LUISS & Minsait, 2024). Nonostante ciò, è bene ricordare che, dal punto di vista dell'impiego, la maggior parte delle imprese è ancora oggi in fase di adozione di applicazioni IA tradizionali nei processi produttivi e gestionali, anche grazie alle novità portate dall'Industria 4.0 (prima e ora) e dall'Industria 5.0 (ora e nei prossimi anni) (AFIL et al., 2024; LUISS & Minsait, 2024).

Applicazioni di IA all'interno del mondo industriale possono trovarsi lungo tutta la catena del valore (*value chain*), ovvero l'insieme di attività che permettono alle aziende di creare e gestire il valore. Facendo riferimento alla catena del valore identificata da Porter (Porter, 2001), si possono distinguere processi primari (contribuiscono in maniera diretta alla produzione dei beni e dei servizi dell'azienda) e processi di supporto (creano le condizioni necessarie affinché le attività primarie possano svolgere la propria funzione).

Di seguito viene fornito un elenco di processi primari e di supporto, con relativa spiegazione, e, per ognuno, vengono dettagliati degli esempi di implementazioni di IA, permettendo di evidenziarne i benefici applicativi. È importante sottolineare che l'elenco di seguito riportato è da intendersi come esemplificativo e non esaustivo in quanto alcune delle applicazioni di seguito classificate possono trovare impiego anche a supporto di altri processi primari o di supporto. Allo stesso modo, la descrizione fornita per ciascuna applicazione, ha l'obiettivo di chiarire, concettualmente, alcuni dei possibili vantaggi senza andare troppo nel dettaglio.

Processi primari

Logistica in entrata

La logistica in entrata comprende tutte le attività che servono per la ricezione di materie prime, semilavorati e input per la produzione, la loro conservazione e diffusione all'interno dei processi produttivi. Sono legate quindi anche alla

movimentazione dei materiali, al loro immagazzinamento, al controllo dell'inventario, e alla gestione dei resi ai fornitori. In questo contesto, l'IA vede applicazioni quali:

- Ottimizzazione della catena di fornitura (S*upply Chain Optimization*). Tramite l'analisi dei dati di domanda (es. storico ordini, prezzi, inventario), fornitura (es. tempi di consegna, caratteristiche e scorte dei fornitori), e altre tipologie utili, permette di migliorare l'organizzazione della catena di fornitura, ottimizzando la disponibilità e il flusso in entrata dei materiali, permettendo di garantire la produzione così come richiesto dai clienti (Kamal et al., 2024; Walter, 2023).

- Gestione del magazzino (*Inventory Management*). Il processamento di dati storici e in tempo reale riguardanti lo stato delle scorte a magazzino, gli ordini di produzione e le previsioni di domanda, le tempistiche di produzione ed altri dati utili, permette di migliorare la gestione del magazzino con l'obiettivo di prevenire situazioni di stock out (mancanza di scorte di materiali per la produzione o di prodotti finiti), e ridurre i costi associati alle scorte (minimizzando la quantità necessaria) (Ayhan & Kır, 2024). Tramite previsioni basate sull'IA, è anche possibile automatizzare parte dei processi di gestione del magazzino, permettendo ad esempio il riordino automatico di alcuni prodotti (Dave & Sarkar, 2023).

- Veicoli a guida autonoma (*Automated Guided Vehicles – AGV*). L'introduzione di veicoli a guida autonoma in ambito produttivo e di gestione della logistica permette di avere maggiore efficienza nei processi, spostando gli operatori umani su altre attività, e permettendo di ridurre i rischi in termini di infortuni legati al trasporto di materiale (Kuo & Wu, 2023; Schweitzer et al., 2023).

Attività operative

Sono tutte quelle attività che permettono di trasformare gli input in prodotto finito. Tra le attività operative troviamo quindi: lavorazioni, assemblaggio, packaging, manutenzione degli impianti e dei macchinari, controllo qualità. In questo contesto, l'IA vede applicazioni quali:

- Ottimizzazione dei processi e delle performance (*Process and Performance Optimization*). L'analisi dei dati storici e di quelli raccolti in tempo reale permette di identificare colli di bottiglia nel processo produttivo andando a ridurre le inefficienze, anticipando i problemi e migliorando il controllo sul processo produttivo, definendo quindi strategie per l'ottimizzazione della pianificazione e programmazione delle attività produttive. Ad esempio, è possibile ottimizzare l'assegnazione delle attività ai macchinari e l'assegnazione delle risorse (tecnologiche e umane) alle varie attività del processo. Come nei casi precedenti, questo porterà a maggiore produttività, riduzione di costi e scarti, maggiore qualità generale del processo (Kalir et al., 2023).

- Manutenzione predittiva (*Predictive Maintenance*). L'obiettivo, tramite la manutenzione predittiva, è quello di prevedere i guasti dei macchinari prima che questi si verifichino. Questo ha poi il beneficio di ridurre i tempi di fermo non programmati (permettendo quindi continuità di produzione), garantire la qualità della produzione (macchinari con dei problemi possono portare a produzioni non conformi e ad un aumento degli scarti), ridurre dei costi di manutenzione e ottimizzazione delle scorte a magazzino (la prevenzione dei guasti permette di evitare rotture di componenti che possono essere molto costose, oltre a fermare la produzione per un tempo maggiore se le parti di ricambio non sono disponibili a

magazzino), e migliorare la sicurezza (macchinari che non funzionano correttamente possono costituire un pericolo per il personale). Tali benefici si possono ottenere studiando e analizzando dati storici (es. storico di manutenzione, dati raccolti dai sensori per definire le soglie ottimali di funzionamento) e dati raccolti tempo reale (es. gli stessi dati raccolti dai sensori, ma in tempo reale) dai macchinari. Al tempo stesso, anche la manutenzione su condizione (*Condition-based Maintenance*), può beneficiare di applicazioni IA grazie a una maggior sensibilità nell'identificazione delle soglie di buon funzionamento e nell'identificazione di pattern di funzionamento anomali (Ferraz Júnior et al., 2023; Liu et al., 2021; Marti-Puig et al., 2024).

- Diagnosi e identificazione dei guasti (*Fault Detection and Diagnosis*). L'adozione di questi algoritmi è orientata all'identificazione di pattern di funzionamento non corretti con l'obiettivo di diagnosticare malfunzionamenti e identificare e prevenire i guasti. Allenare gli algoritmi con dei dati storici e in tempo reale rilevati da sensori (es. vibrazioni, pressioni, correnti) e dati di processo (buon funzionamento e cattivo funzionamento) permette di distinguere condizioni di funzionamento normale, anormale (ma corretto), e di guasto. Come nel caso delle applicazioni di manutenzione predittiva, i benefici risiedono nella riduzione dei tempi di fermo, maggiore qualità dei prodotti, minori costi di manutenzione, migliori prestazioni del sistema e riduzione dei problemi di sicurezza (Cohen et al., 2022; Dubaish & Jaber, 2024).

- Machine vision per lo smistamento e la classificazione (*Machine vision for sorting and grading*). Implementa sistemi di visione artificiale basati sull'IA per ordinare e classificare i prodotti in base a criteri predefiniti.

In questo caso dati quali immagini, video e dati provenienti da sensori (es. LiDAR) vengono utilizzati per addestrare gli algoritmi a riconoscere i pezzi prodotti e smistarli lungo il processo produttivo a seconda delle necessità (es. smistamento pezzi in base a dimensione, colore, forma, tipo). L'introduzione di tali algoritmi permette di controllare in tempo reale i prodotti, ed è necessaria una grande quantità di dati per insegnare all'algoritmo a distinguere i prodotti (Taatali et al., 2024; Uhlemann et al., 2017; Zhang et al., 2019).

- Machine vision per il controllo qualità (*Machine vision for Quality Control*). L'obiettivo in questo caso è utilizzare gli algoritmi per riconoscere non-conformità nei prodotti in modo da rimuovere quelli che non possono essere venduti e intercettare immediatamente problematiche relative alla qualità del processo produttivo. Come nel caso precedente, l'introduzione di tali algoritmi necessita di una grande quantità di dati per insegnare all'algoritmo quali prodotti sono buoni e quali sono difettosi. In termini di benefici, si segnalano una riduzione dei prodotti di scarto dal processo produttivo e un aumento della produttività per quanto riguarda l'attività di controllo qualità che, essendo automatizzata riesce ad essere più veloce di un controllo manuale umano (Chouhad et al., 2021; Park & Jeong, 2022; Urgo et al., 2024).

- Robot autonomi (*Autonomous Robots*). In questo caso il concetto alla base dell'applicazione dell'IA è quello di migliorare l'interazione tra le risorse umane ed i robot, migliorando l'esecuzione delle attività assegnate grazie al processamento dei dati. Esegue compiti complessi come la saldatura, la verniciatura e la movimentazione dei materiali con elevata precisione (Umbrico et al., 2022).

- Supporto alla progettazione tramite design generativo (*Generative Design*). Utilizza algoritmi di IA per creare prodotti ottimizzati esplorando una vasta gamma di possibili configurazioni a seconda delle esigenze (ad esempio, ottimizzazione del processo produttivo, di servizio, sostenibilità). Le applicazioni possono quindi riguardare lo sviluppo di nuovi prodotti, l'ottimizzazione del loro design e anche la relativa personalizzazione sulla base delle esigenze degli attori coinvolti. Di fatto, la possibilità di sfruttare questo tipo di IA permette di ridurre sensibilmente i tempi di progettazione e sviluppo dei prodotti senza inficiare la qualità, permettendo però di ridurre i costi. In questo caso l'addestramento dell'algoritmo si basa sulla presenza di dati storici riguardanti il design di prodotti precedenti di vario tipo e con varie caratteristiche, in modo che l'IA possa esplorare varie configurazioni a seconda delle specificità del caso (Bartlett & Camba, 2024; Lee et al., 2024).

- Generazione dati sintetici (*Synthetic Data Generation*). Lo scopo di queste applicazioni è creare dei dati sintetici per addestrare i modelli di machine learning in situazioni in cui i dati reali sono scarsi o costosi da ottenere. È ovviamente necessario avere a disposizione dei dati reali su cui allenare l'algoritmo che poi sarà responsabile della generazione dei dati. Oltre a ciò, è necessario dare indicazioni precise al modello perché sia in grado di creare dati usabili per situazioni specifiche. In termini di benefici, si può identificare una riduzione dei costi per quanto riguarda la campagna di raccolta dati che, in alcuni casi può essere onerosa sia in termini di costi che di tempo necessario per raccogliere la quantità necessaria. Se ben addestrati, i modelli che svolgono questa funzione sono in grado di generare dati coerenti con la realtà, garantendo quindi la qualità richiesta oltre che la quantità di dati richiesta. Di fatto, la possibilità di

generare dati sintetici può essere utile per testare prodotti, creare campagne di marketing personalizzate, o generare informazioni per i clienti tramite chatbot (De & Mitra, 2024; Urgo et al., 2024).

Logistica in uscita

Comprende attività che riguardano la raccolta e distribuzione fisica dei prodotti. Includono anche aspetti di movimentazione dei materiali, dei veicoli per la consegna, e l'elaborazione e pianificazione degli ordini. Oltre ad applicazioni simili a quelle della logistica in entrata, in questo contesto, ci sono applicazioni quali:

- Previsione della domanda (*Demand Forecasting*). L'obiettivo di questi algoritmi è quello di sfruttare l'analisi di dati storici riguardanti la domanda e i fattori che possono influire su di essa, per prevedere la domanda futura dei prodotti. I benefici sono simili a quelli menzionati con la logistica in entrata, quindi l'ottimizzazione della gestione del magazzino, la migliore pianificazione della produzione, una riduzione dei costi di trasporto e logistica (Groene & Zakharov, 2024; Walter, 2023).

Marketing e vendite

Queste attività sono associate alla promozione dei prodotti tramite pubblicità, identificazione di canali ottimali per raggiungere il cliente, definizione di politiche di prezzo ottimali e ottimizzazione della forza vendita. In questo contesto, l'IA vede applicazioni quali:

- Personalizzazione dei prodotti basata sull'IA (*AI-driven Product Customization*). Vengono sfruttati algoritmi in grado di sfruttare dati relativi a prodotti (ad esempio caratteristiche prodotto, tendenze di mercato) e clienti (es. storico acquisti, dati di profilazione) per personalizzare l'offerta e l'esperienza di acquisto. I principali benefici risiedono in una maggiore soddisfazione

dei clienti, cui vengono proposti solo prodotti di interesse, un aumento delle vendite e una riduzione dei costi associati alla produzione (prioritizzazione del mix produttivo sulla base delle richieste dei clienti). L'analisi dei dati di vendita e di interazione con i clienti può anche essere utile per identificare nuovi trend e anticipare la concorrenza sul mercato (Powell et al., 2024; Tejasvi et al., 2024).

Servizi alla clientela

Riguardano tutte le attività associate alla percezione che i clienti hanno del valore dei prodotti, con l'obiettivo di mantenerlo stabile o innalzarlo. Tra queste attività si possono menzionare l'installazione e riparazione dei prodotti, la formazione del personale e la fornitura di ricambi. Oltre alle applicazioni di manutenzione predittiva e identificazione dei guasti discusse in precedenza, in questo contesto, l'IA vede applicazioni quali:

- Automazione dell'elaborazione dei documenti (*Document Processing Automation*). In questo caso, vengono sfruttati algoritmi di elaborazione del linguaggio naturale (*Natural Language Processing - NLP*) per automatizzare l'estrazione, l'elaborazione e la gestione dei dati testuali da documenti come fatture, ordini e rapporti sulla qualità. L'adozione di questo tipo di algoritmi e applicazioni è fondamentale per la creazione di un sistema di gestione della conoscenza che permette di ottimizzare i processi a vario livello e di condividere la conoscenza in maniera ottimale (Muludi et al., 2024; Sala et al., 2023).

- Chatbot per il supporto dei clienti (*Chatbots for Customer Support*). Come suggerisce il nome, l'idea è quella di utilizzare chatbot per automatizzare la gestione di alcune richieste dei clienti, permettendo così di fare intervenire tecnici e risorse umane solo nel momento in cui le necessità dei clienti lo richiedono.

L'implementazione di questo tipo di chatbot permette di migliorare l'efficienza del processo di gestione clienti e la relativa soddisfazione mettendo a disposizione uno strumento di assistenza disponibile 24 ore su 24 e 7 giorni su 7. Di conseguenza, questo permette anche di razionalizzare l'impiego di risorse umane su determinate attività ottenendo anche una riduzione dei costi associati. È però necessario avere a disposizione un'adeguata base di conoscenza per istruire questi chatbot, con informazioni che spaziano dallo storico delle interazioni con i clienti, alle informazioni (anche tecniche) su prodotti e servizi. L'organizzazione e il processamento delle informazioni per trasformarle in conoscenza utilizzabile richiede l'adozione di tecniche di NLP (El-Ansari & Beni-Hssane, 2023; Gupta et al., 2024).

Processi di supporto

Approvvigionamenti

Sono attività associate alla funzione di acquisto degli input che vengono utilizzati dall'impresa e trasformati in output attraverso i processi produttivi. Possono includere materie prime e altri articoli di consumo. Ricadono in questa categoria anche i macchinari, e le attrezzature per laboratorio e ufficio. In questo contesto, l'IA vede applicazioni quali:

- Previsione e gestione del consumo energetico (*Energy Consumption Forecasting and Management*). L'applicazione di questi algoritmi si basa sull'analisi di dati storici e in tempo reale riguardo i consumi degli impianti e delle attività produttive. L'uso di questi algoritmi permette l'identificazione di eventuali sprechi energetici, portando a risparmi anche in termini di efficienza operativa e sostenibilità (Danish, 2023).

- Relazioni con i fornitori (*Supplier Relationship Management*). Sfruttando un mix di dati strutturati e non

strutturati è possibile identificare dei pattern di comportamento e delle caratteristiche che in grado di supportare attività come la selezione dei fornitori e le successive relazioni, permettendo di ridurre i costi tramite l'ottimizzazione dei processi di approvvigionamento e una migliore gestione dei rischi associati alla variabilità del mercato (Kamal et al., 2024; Walter, 2023).

Sviluppo delle tecnologie

Riguardano tutti quegli aspetti legati al miglioramento dei processi e dei prodotti o servizi offerti dall'azienda. Tramite lo sviluppo delle tecnologie è possibile creare un know-how aziendale da sfruttare per il miglioramento dell'offerta ai clienti. In questo contesto, l'IA vede applicazioni quali:

- Gemelli digitali (*Digital Twin*). Come dice il nome stesso, un *digital twin* è una replica, digitale, di un oggetto fisico che può essere sfruttato per effettuare simulazioni e modellazioni con l'obiettivo di ottimizzare i processi produttivi, ridurre costi di sperimentazione e prevedere problematiche relative a specifiche configurazioni prima che queste vengano adottate. L'utilizzo dell'IA nel contesto dei *digital twin* permette di migliorare le simulazioni e ottenere vantaggi in termini di costi e qualità. Lo sviluppo di gemelli digitali è infatti molto costoso e richiede tempo per essere fatto con metodi tradizionali, sfruttando l'IA, è possibile creare *digital twin* in tempi e a costi ridotti a patto di avere a disposizione tutti i dati necessari (es. dati di progettazione, funzionamento, manutenzione). Applicazioni specifiche in termini di ottimizzazione della produzione, manutenzione predittiva e controllo qualità, oltre che gestione dei flussi in entrata e in uscita sono poi ottenibili (Chen et al., 2024; Sadeghi et al., 2024; Uhlemann et al., 2017; Urgo et al., 2024).

- Personalizzazione strumentazione per attività specifiche (*Custom Tooling and Fixtures Design*). I vantaggi di questa tipologia di applicazione sono riconducibili a quelli espressi quando si è parlato di supporto alla progettazione tramite design generativo. Se però nel caso precedente l'obiettivo della progettazione era rivolto all'esterno, in questo caso l'obiettivo è migliorare la progettazione degli strumenti e macchinari che servono per la produzione. I benefici si ritrovano poi principalmente in termini di ottimizzazione del processo produttivo, riduzione dei tempi di fermo, maggiore qualità e sicurezza. Come nel caso del design, è necessario fornire all'algoritmo, tra gli altri, dati riguardo al design dei pezzi da produrre, ai materiali, al processo produttivo, alle normative (Bartlett & Camba, 2024; Lee et al., 2024).

Gestione delle risorse umane

Sono attività e applicazioni di AI associate alla sfera della forza lavoro e possono riguardare la ricerca di nuovo personale e la gestione del personale già in azienda, ad esempio attraverso attività di formazione. Riguardano tutta la catena del valore, in quanto l'introduzione di risorse umane tocca sia i processi primari che quelli secondari. In questo contesto, l'IA vede applicazioni quali:

- Sistemi di gestione della conoscenza (*Knowledge Management Systems*). Questo tipo di attività si concentra sull'organizzazione, recupero e fornitura informazioni da grandi quantità di documentazione tecnica. Le informazioni, sotto forma di testo, video, immagini e/o audio, vengono elaborate e organizzate per un più semplice recupero che ha l'obiettivo di contribuire a una maggiore efficienza dei processi, maggiore produttività, miglioramento del processo decisionale, miglior servizio clienti. Come in altri casi, una delle conseguenze identificate riguarda la riduzione dei costi

associati all'esecuzione di attività e processi (Meyers et al., 2024; Walter, 2023).

- Automazione dei processi di business (Business Process Automation). Ha l'obiettivo di automatizzare tramite software le attività ripetitive che possono trovarsi in ambiente aziendale (es. inserimento o estrazione dati). L'utilizzo di IA permette di eseguire quindi azioni ripetitive in maniera affidabile, riducendo il numero di attività operative e a basso valore aggiunto che un operatore umano potrebbe trovarsi a fare, riducendo sia la probabilità di errore che i tempi di esecuzione richiesti. Di fatto, questo porta a una maggiore efficienza e a minori costi di esecuzione, con la possibilità di scalare sulla capacità di analisi in tempi brevi. I dati richiesti per l'allenamento degli algoritmi e per il loro utilizzo variano a seconda della casistica d'uso. Spesso, quando vengono utilizzati dei software per imitare azioni umane (es. digitazione, clic, navigazione), si parla di Robotic Process Automation (RPA) (Moorthy et al., 2023; Zebec & Indihar Štemberger, 2024).

- Sistemi di tutoring intelligenti (*Intelligent Tutoring Systems*). Si tratta di applicazioni che permettono di fornire istruzioni e materiale di apprendimento personalizzato in base alle esigenze degli utilizzatori (Sewunetie & Kovács, 2022; Shamsuddinova et al., 2024). L'allenamento di tali algoritmi richiede informazioni relative ai contenuti da fornire, al livello di difficoltà, alle caratteristiche degli utilizzatori. È ovviamente anche necessario raccogliere feedback da parte degli utilizzatori per continuare a migliorare la proposta dell'algoritmo. In questo caso, i maggiori benefici risiedono sicuramente nella riduzione degli errori nel contesto dell'attività oggetto di tutoring con un conseguente incremento della

produttività e, di riflesso, una riduzione dei costi. Questi sistemi possono anche essere rivolti alla clientela.

Attività infrastrutturali

Questa categoria di attività vuole supportare aspetti quali la contabilità, la gestione generale, la gestione delle attività finanziarie e altre. In questo contesto, l'IA vede applicazioni quali:

- Controllo compliance (*Compliance Monitoring*). L'obiettivo di questa applicazione è assicurare l'aderenza delle attività e dei processi aziendali alle normative vigenti. È quindi necessario fornire all'algoritmo di IA i requisiti legislativi per lo svolgimento delle attività aziendali e le informazioni su come vengono effettivamente svolti in azienda per assicurare la coerenza tra i due aspetti. Si tratta poi, nello specifico, di implementare sistemi di monitoraggio e segnalazione per identificare eventuali correzioni da apportare. Oltre a un miglioramento dell'efficienza riguardo al monitoraggio degli aspetti legali, tra i benefici si possono individuare un minor rischio di incorrere in sanzioni legali, un migliorato processo decisionale riguardo alcuni aspetti critici (Koshiyama et al., 2024; Robaldo et al., 2023; Ryan et al., 2024).

- Generazione automatica di report (*Automated Report Generation*). Questa applicazione sfrutta l'IA con lo scopo di processare grandi quantità di dati e restituire report sintetici con i punti e i messaggi fondamentali richiesti per prendere decisioni o comunicare con altri attori (anche clienti). L'automatizzazione di questa attività permette una maggior efficienza e un miglioramento dei processi decisionali a fronte di una riduzione dei costi e una maggiore standardizzazione delle informazioni restituite. I dati richiesti per questa attività possono essere sia di tipo strutturato che non strutturato, a seconda del caso d'uso (Jafari et al., 2021; McMaster et al., 2023).

3. Ostacoli

Nonostante il numero consistente di benefici ottenibili tramite l'adozione dell'IA, è necessario discutere anche le possibili limitazioni e fattori che ne possono impedire il raggiungimento. Secondo il campione sondato da LUISS & Minsait (2024) la percentuale di aziende, non facendo distinzioni a livelli di dimensione, che hanno preso in considerazione l'utilizzo di tecnologie di IA ma non le hanno ancora utilizzate è il 4,4% (15,3% tra le grandi). La ragione di questo risultato risiede, secondo le aziende intervistate, principalmente in tre fattori:

- la mancanza di competenze,
- i costi elevati,
- la mancanza di dati di qualità necessari all'uso di applicativi AI.

Per quanto riguarda le competenze, il tema risulta essere tra i più discussi. L'introduzione dell'IA porterà infatti all'introduzione di nuove figure, in possesso di competenze non solo legate ai processi tradizionali, ma anche a quelli capaci di sfruttare l'IA per funzionare in maniera più efficiente. Il tutto, in molti casi, si lega anche alla quarta e quinta rivoluzione industriale che, oltre a necessitare di competenze tecniche, richiedono anche sulle competenze di resilienza, di sostenibilità e sociali. Particolarmente importante è la riflessione da fare sull'introduzione di IA porta a rendere obsolete alcune competenze e richiede un continuo processo di apprendimento di nuove skill per stare al passo con l'evoluzione tecnologica (Tamayo et al., 2023). Da questo punto di vista, può essere visto come un fattore di stimolo per aziende e lavoratori per continuare a investire sulla formazione e sull'aggiornamento.

Per quanto riguarda i costi vanno fatte considerazioni relative alle attività infrastrutturali (es. potenza computazionale richiesta, scalabilità delle applicazioni), alle competenze richieste per lo sviluppo, validazione, e l'utilizzo di applicativi di IA in ambito

aziendale compresa l'integrazione con i sistemi già in uso in azienda. Infatti, il passaggio dall'utilizzo di sistemi non IA a sistemi di IA (tradizionale o generativa) richiede un ripensamento dei processi di condivisione, gestione e utilizzo dati. È necessario, infatti, assicurarsi che l'introduzione di tali sistemi non causi problemi con i sistemi già in uso in azienda al fine di evitare interruzioni di servizio che possono portare a perdite ingenti in termini economici per l'azienda in termini di produzione o di servizio ai clienti. Da un punto di vista puramente economico, è necessario sottolineare che l'investimento inizialmente richiesto per sviluppare un'IA in grado di supportare i processi e le attività aziendali può essere molto alto in termini di raccolta dati, di infrastruttura e di addestramento. Non solo, durante l'utilizzo è necessario aggiornare continuamente l'IA ed eseguire attività manutentive orientate al mantenerla correttamente funzionante. Questo aspetto riguarda soprattutto il garantire che l'aggiunta di nuovi dati al modello non causi delle variazioni negative nel processo decisionale, richiedendo quindi di rendere questo processo il più robusto possibile alle variazioni che possono essere introdotte nei processi aziendali. Questo richiede una fase di re-training periodica. Inoltre, non è garantito il ritorno sull'investimento effettuato per rendere funzionante l'IA.

La mancanza di dati affidabili è un fattore particolarmente critico, in quanto l'affidabilità e l'accuratezza degli algoritmi si basa sul modo in cui questi vengono addestrati, è necessario che i dati necessari all'addestramento e utilizzo quotidiano siano disponibili e di buona qualità, evitando di utilizzare fonti dati poco affidabili (per mancanza di integrità o informazioni, contenenti errori, di formati diversi), anche in termini di numerosità e disponibilità dei dati. Inoltre, è importante, ove necessario, considerare anche la condivisione di dati tra diverse fonti.

Un altro tema da considerare è quello dell'etica e della conformità ai regolamenti e leggi, motivo per cui è necessario gestire i dati in maniera specifica a seconda della loro tipologia. Sempre a proposito di questo tema, è necessario considerare la limitazione riguardo alla trasparenza e comprensione delle

decisioni dell'IA, che porta poi a possibili discussioni circa possibili pregiudizi che l'IA potrebbe avere riguardo certi argomenti a causa del modo in cui è stata addestrata.

Infine, un aspetto fondamentale è quello della sicurezza e della vulnerabilità dell'IA agli attacchi informatici dovuti a input volutamente errati o modifiche nel processo di addestramento dell'IA. Per questo è necessario creare delle pipeline di addestramento ed uso sicure, in grado di garantire gli standard richiesti per la raccolta, trasmissione e processamento dei dati all'interno dell'IA.

Sempre secondo il rapporto LUISS & Minsait (2024), il 14,3% delle aziende intervistate non trova utilità nell'applicazione delle tecnologie di IA. Nello specifico, nel caso dell'IA generativa, è la mancanza di un chiaro caso d'uso a livello di business che frena l'adozione di questa tecnologia. A questo si aggiungono possibili rischi normativi e, di nuovo, la mancanza di competenze adeguate.

4. Spunti di riflessione

È quindi necessario evidenziare come il cammino delle aziende verso l'adozione in maniera massiccia di applicazioni IA sia ancora agli albori, nonostante i grossi benefici che si possono intravedere, soprattutto in termini di ottimizzazione processi, supporto alle decisioni e gestione dei rapporti con i clienti. Questo ha come effetto anche la possibilità di abilitare nuove offerte di business basate sui servizi grazie a una maggior conoscenza di prodotti e bisogni dei clienti. In particolare, è l'utilizzo combinato di IA tradizionale e IA generativa che può spingere la produttività ai massimi livelli. In particolare, l'uso combinato di IA tradizionale e IA generativa promette di ridurre sensibilmente il tempo impiegato per eseguire molte operazioni, soprattutto routinarie.

Nonostante i benefici previsti, l'ostacolo più grande individuato dalle aziende risiede nella disponibilità di personale con le competenze adeguate a sfruttare al meglio le possibilità offerte dall'IA. La questione si conferma prioritaria anche guardando gli investimenti che, secondo il rapporto (*European Commission.*

Joint Research Centre, 2022), per la maggior parte riguardano questo aspetto. La questione è discussa nel dettaglio anche nel Programma Strategico Intelligenza Artificiale 2022-2024 (Governo Italiano, 2021), che propone una serie di linee guida e obiettivi da raggiungere per lo più incentrati sulla creazione di competenze legate all'IA per supportare l'economia italiana. La rilevanza dell'aspetto finanziario, così come la disponibilità dei dati necessari alle applicazioni, varia a seconda della dimensione delle aziende. Quelle più grandi sono tendenzialmente più pronte rispetto alle aziende medio-piccole, che non hanno, in generale, infrastrutture e risorse adeguate e che quindi necessiterebbero di risorse specifiche (*European Court of Auditors*, 2024). A ciò si aggiunge il fatto che molte aziende identificano nella mancanza di una strategia di utilizzo o di casi d'uso (50% delle aziende intervistate dal campione LUISS) utili una delle ragioni per la non adozione. Infine, la mancanza di un quadro normativo chiaro pone un freno anche ai tentativi di *proof of concept* delle aziende (13% del campione LUISS). Proprio per questo motivo, si sottolinea la necessità di supportare le aziende, soprattutto quelle medio-piccole, nell'adozione di tecnologie IA e nell'individuazione di casi d'uso appropriati, oltre che a investire nella formazione del personale. Per quanto riguarda l'adozione di IA generativa all'interno dei processi aziendali, emerge la preoccupazione rispetto ai dati generati e/o al problema di allucinazioni degli algoritmi che potrebbero portare a decisioni errate.

A livello europeo, il rapporto della Corte dei Conti (*European Court of Auditors*, 2024) evidenzia come le misure della Commissione Europea e quelle nazionali non siano ad oggi abbastanza coordinate, principalmente per la limitata disponibilità di strumenti di controllo e/o la loro parziale implementazione, oltre alla definizione di target non competitivi, soprattutto rispetto ai leader a livello globale.

5. Prospettive future

Sebbene l'integrazione dell'IA nei processi aziendali sia già una realtà, la rapidità dell'evoluzione dei modelli impone di mantenere

lo sguardo in avanti. In particolare, lo sviluppo degli LLM e dei modelli multimodali rende verosimili scenari in cui l'IA potrebbe raggiungere livelli di autonomia impensabili con le tecnologie attuali.

La capacità di tali modelli di apprendere compiti nuovi, interagire organicamente con le persone e operare in modo proattivo differenziano l'IA da altri strumenti, e permettono a queste tecnologie di sostituirsi sempre più agli umani su alcuni compiti. L'interazione uomo-macchina in questo contesto ha tratti inediti, e porta a chiedersi quale diventerà il ruolo dell'IA generativa in azienda: l'IA sarà uno strumento al supporto dell'azione del lavoratore, o diventerà sempre di più un nuovo "collega" (o "stagista") con cui interagire e collaborare? come si struttura l'interazione tra esseri umani ed IA rispetto alle interazioni lavorative tra esseri umani? Quest'ultima prospettiva rende sempre più rilevante l'adozione di un approccio multidisciplinare allo studio del fenomeno, che integri gli aspetti tecnologici ed economici, con punti di vista sociologici, psicologici, e legati alle risorse umane. Si pone inoltre il tema della responsabilità: l'intelligenza artificiale ha la possibilità di prendere decisioni autonome? E se sì, su chi ricade la responsabilità di tali scelte?

L'argomento dell'interazione e della collaborazione uomo-intelligenza artificiale non è nuovo nella letteratura, ma è solo alla luce di più recenti sviluppi tecnologici che ha iniziato ad assumere più risvolti pratici. Lo studio di O'Neill et al (2022) presenta un'analisi della letteratura, parlando più precisamente di *human-autonomy teaming* (HAT) per descrivere "l'interazione tra umani e agenti autonomi ed intelligenti che lavorano in modo interdipendente alla realizzazione di obiettivi comuni". L'agency degli agenti autonomi si manifesta attraverso l'indipendenza, l'auto-governo e la proattività, e può variare su una scala di autonomia (*Levels of Abstraction*, LOA) che spazia da agenti senza autonomia che richiedono il totale controllo umano, fino ad agenti altamente autonomi. In questa scala i concetti di automazione e autonomia sono su un continuo, in cui

l'automazione caratterizza i gradi più bassi della scala, mentre i gradi più alti sono ascrivibili agli agenti autonomi. È interessante notare come questa scala faccia riferimento più al grado di autonomia che al livello di "intelligenza" di un sistema: ad esempio un sistema di IA che fornisce un set di decisioni/azioni alternative che l'umano può implementare (ad esempio un LLM che fornisce una serie di testi di presentazione di un prodotto) è, secondo questa scala un soggetto privo di autonomia. Un soggetto altamente autonomo è invece capace di decidere e agire autonomamente ignorando l'umano, o informandolo delle azioni prese.

La ricerca empirica sta iniziando ad analizzare le dinamiche delle interazioni tra umani ed intelligenze artificiali in contesti di lavoro di team: Dennis et al. (2023) studiano le percezioni degli utenti rispetto a queste collaborazioni in un contesto sperimentale controllato per valutarne i comportamenti e i bias, anche in relazione alle performance. Gli autori nell'esperimento riscontrano come le IA nel contesto sperimentale siano giudicate in modo simile ad altri umani, e che non ci siano particolari differenze nella fiducia e nell'attitudine alla collaborazione con questi agenti. Si mostra inoltre che, nei casi in cui le performance delle IA siano buone, gli umani percepiscano meno conflitti nel team rispetto ad un team di umani con simile livello di competenze. Gli esperimenti condotti mostrano però che la presenza di agenti IA porti ad una riduzione della soddisfazione dei membri del team. Considerazioni analoghe a queste relative all'interazione con IA, possono essere fatte nell'ambito della robotica, dove gli avanzamenti nell'IA rendono più concrete le prospettive di modelli organizzativi in cui esseri umani e robot collaborino in modo meno strutturato di quanto facciano ora (si veda ad esempio il lavoro di Li et al, 2022, che analizza il campo della collaborazione umano-robot proattiva).

Gli studi accademici e l'esperienza pratica ci mostrano in ogni caso che, come per qualsiasi cosa, esiste una curva d'apprendimento, ed è fondamentale sviluppare competenze e

soft-skill specifiche per un contesto dove l'uso dell'IA è pervasivo. Per sfruttare a pieno il potenziale dell'IA generativa è necessario capire chiaramente cosa un utente o un membro di un team possa fare con essa e sviluppare best practices condivise. Una mancata alfabetizzazione rischia di portare ad un uso poco efficace di queste tecnologie o di riporre in esse aspettative troppo alte, con il risultato di risultati sub-ottimali, o di portare a perdere più tempo di quello risparmiato.

Bibliografia

AFIL, Università degli Studi di Bergamo, & Politecnico di Milano. (2024). R&I Priorities for enhancing Artificial Intelligence applications in Manufacturing in Lombardy. https://afil.it/afil-journal/cluster-meet-regions-milano-il-ruolo-fondamentale-dellintelligenza-artificiale-per-lindustria-manifatturiera-lombarda

Åström, J., Reim, W., & Parida, V. (2022). Value creation and value capture for AI business model innovation: A three-phase process framework. *Review of Managerial Science*, 16(7), 2111–2133.

Ayhan, H. M., & Kır, S. (2024). MI-driven approaches to enhance inventory planning: Inoculant weight application in casting processes. *Computers & Industrial Engineering*, 110280.

Bartlett, K. A., & Camba, J. D. (2024). *Generative Artificial Intelligence in Product Design Education: Navigating Concerns of Originality and Ethics*.

Black, S., Samson, D., & Ellis, A. (2024). Moving beyond 'proof points': Factors underpinning AI-enabled business model transformation. *International Journal of Information Management*, 77, 102796.

Bocken, N. M. P., Short, S. W., Rana, P., & Evans, S. (2014). A literature and practice review to develop sustainable business model archetypes. *Journal of Cleaner Production*, 65, 42–56.

Chen, Z., Surendraarcharyagie, K., Granland, K., Chen, C., Xu, X., Xiong, Y., Davies, C., & Tang, Y. (2024). Service oriented digital twin for additive manufacturing process. *Journal of Manufacturing Systems*, 74, 762–776.

Chouhad, H., El Mansori, M., Knoblauch, R., & Corleto, C. (2021). Smart data driven defect detection method for surface quality control in manufacturing. *Measurement Science and Technology*, 32(10), 105403.

Cohen, J., Jiang, B., & Ni, J. (2022). EveSyncIAI: Event synchronization industrial augmented intelligence for fault diagnosis. *IEEE Transactions on Semiconductor Manufacturing*, 35(3), 446–456.

Danish, M. S. S. (2023). A framework for modeling and optimization of data-driven energy systems using machine learning. *IEEE Transactions on Artificial Intelligence*, 5 (5), 2434–2443.

Dave, R., & Sarkar, B. (2023). AI-Powered Inventory Optimization in Industrial Manufacturing. *International Journal of Engineering Trends and Technology*, 71, 13–25.

De, S., & Mitra, P. (2024). A Novel Technique of Synthetic Data Generation for Asset Administration Shells in Industry 4.0 Scenarios. *IEEE Transactions on Artificial Intelligence*.

Dennis, A. R., Lakhiwal, A., & Sachdeva, A. (2023). AI Agents as Team Members: Effects on Satisfaction, Conflict,

Trustworthiness, and Willingness to Work With. *Journal of Management Information Systems*, 40(2), 307-337.

Dubaish, A. A., & Jaber, A. A. (2024). Comparative Analysis of SVM and ANN for Machine Condition Monitoring and Fault Diagnosis in Gearboxes. *Mathematical Modelling of Engineering Problems*, 11(4).

El-Ansari, A., & Beni-Hssane, A. (2023). Sentiment analysis for personalized chatbots in e-commerce applications. *Wireless Personal Communications*, 129(3), 1623–1644.

European Commission. (2024). Digital Europe Programme (DIGITAL) | EU Funding & Tenders Portal. https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/programmes/digital

European Commission. Joint Research Centre. (2022). *AI Watch: Estimating AI investments in the European Union*. Publications Office. https://data.europa.eu/doi/10.2760/702029

European Court of Auditors. (2024). *Special report 08/2024: EU Artificial intelligence ambition* (8). European Court of Auditors.

Eurostat. (2023). Artificial intelligence by NACE Rev.2 activity (isoc_eb_ain2) [Dataset]. https://doi.org/10.2908/ISOC_EB_AIN2

Eurostat. (2023). Artificial intelligence by size class of enterprise (isoc_eb_ai) [Dataset]. https://doi.org/10.2908/ISOC_EB_AI

Eurostat. (2024). Use of artificial intelligence in enterprises. Eurostat. https://ec.europa.eu/eurostat/statistics-explained/index.php?oldid=568530

Ferraz Júnior, F., Romero, R. A. F., & Hsieh, S.-J. (2023). Machine Learning for the Detection and Diagnosis of Anomalies in Applications Driven by Electric Motors. *Sensors*, 23(24), 9725.

Governo Italiano. (2021). Programma Strategico Intelligenza Artificiale 2022-2024.

Groene, N., & Zakharov, S. (2024). Introduction of AI-based sales forecasting: How to drive digital transformation in food and beverage outlets. *Discover Artificial Intelligence*, 4(1), 1–17.

Gupta, M., Dheekonda, V., & Masum, M. (2024). Genie: Enhancing information management in the restaurant industry through AI-powered chatbot. *International Journal of Information Management Data Insights*, 4(2), 100255.

Jafari, P., Al Hattab, M., Mohamed, E., & AbouRizk, S. (2021). Automated extraction and time-cost prediction of contractual reporting requirements in construction using natural language processing and simulation. *Applied Sciences*, 11(13), 6188.

Kalir, A. A., Lo, S. K., Goldberg, G., Zingerman-Koladko, I., Ohana, A., Revah, Y., Chimol, T. B., & Honig, G. (2023). Leveraging Machine Learning for Capacity and Cost on a Complex Toolset: A Case Study. *IEEE Transactions on Semiconductor Manufacturing*. 36 (4), 611—618.

Kamal, C. R., Agnes, G., Jemima, L., & Chandrakala, M. (2024). Automation of Business Processes Using Robots in the Fields of Supply Chain Management, Intelligent Transportation, and Logistics. In AI in Business: *Opportunities and Limitations: Volume 1* (pp. 477–489). Springer.

Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, L., Pavey, G., Ahamat, G., Leutner, F., Goebel, R., Knight, A., & others. (2024). Towards algorithm auditing: Managing legal, ethical and

technological risks of AI, ML and associated algorithms. *Royal Society Open Science*, 11(5), 230859.

Kuo, Y.-H., & Wu, E. H.-K. (2023). Advanced, Innovative AIoT and Edge Computing for Unmanned Vehicle Systems in Factories. Electronics, 12(8), Articolo 8.

Lee, S., Law, M., & Hoffman, G. (2024). When and How to Use AI in the Design Process? Implications for Human-AI Design Collaboration. *International Journal of Human–Computer Interaction*, 1–16.

Li, S., Zheng, P., Liu, S., Wang, Z., Wang, X. V., Zheng, L., & Wang, L. (2023). Proactive human–robot collaboration: Mutual-cognitive, predictable, and self-organising perspectives. *Robotics and Computer-Integrated Manufacturing*, 81, 102510.

Liu, Y., Yu, W., Dillon, T., Rahayu, W., & Li, M. (2021). Empowering IoT predictive maintenance solutions with AI: A distributed system for manufacturing plant-wide monitoring. *IEEE Transactions on Industrial Informatics*, 18(2), 1345–1354.

LUISS & Minsait. (2024). Intelligenza Artificiale in Italia: La rivoluzione che sta cambiando il business. https://landing.luiss.it/book/report/minsait-luiss.html

Madanaguli, A., Sjödin, D., Parida, V., & Mikalef, P. (2024). Artificial intelligence capabilities for circular business models: Research synthesis and future agenda. *Technological Forecasting and Social Change*, 200, 123189.

Marti-Puig, P., Touhami, I. A., Perarnau, R. C., & Serra-Serra, M. (2024). Industrial AI in condition-based maintenance: A case study in wooden piece manufacturing. *Computers & Industrial Engineering*, 188, 109907.

McMaster, C., Chan, J., Liew, D. F., Su, E., Frauman, A. G., Chapman, W. W., & Pires, D. E. (2023). Developing a deep learning natural language processing algorithm for automated reporting of adverse drug reactions. *Journal of Biomedical Informatics*, 137, 104265.

Meyers, B., Vangheluwe, H., Lietaert, P., Vanderhulst, G., Van Noten, J., Schaffers, M., Maes, D., & Gadeyne, K. (2024). Towards a knowledge graph framework for ad hoc analysis in manufacturing. *Journal of Intelligent Manufacturing*, 1–22.

Mont, O. (2002). Clarifying the concept of product-service system. *Journal of Cleaner Production*, 10(3), 237–245.

Moorthy, C., Srivastava, A., Vasundhara, D., Reddy, V., & Prasad, A. (2023). Robotic Process Automation-The Ineluctable Virtual Workforce in Various Business Sectors: Post Covid-19 Scenario. International *Journal of Engineering Trends and Technology*, 143–154.

Muludi, K., Fitria, K. M., Triloka, J., & others. (2024). Retrieval-Augmented Generation Approach: Document Question Answering using Large Language Model. *International Journal of Advanced Computer Science & Applications*, 15(3).

OECD.AI. (2024). Live data from OECD.AI. https://oecd.ai/en/data
Osterwalder, A., & Pigneur, Y. (2010). *Business model generation: A handbook for visionaries, game changers, and challengers* (Vol. 1). John Wiley & Sons.

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–autonomy teaming: A review and analysis of the empirical literature. *Human factors*, 64(5), 904-938.

Park, M., & Jeong, J. (2022). Design and implementation of machine vision-based quality inspection system in mask manufacturing process. *Sustainability*, 14(10), 6009.

Pirola, F., Boucher, X., Wiesner, S., & Pezzotta, G. (2020). Digital technologies in product-service systems: A literature review and a research agenda. *Computers in Industry*, 123, 103301.

Porter, M. E. (2001). The value chain and competitive advantage. *Understanding business processes*, 2, 50–66.

Powell, C., Zhu, E., Xiong, Y., & Yang, S. (2024). A data-driven approach to predicting consumer preferences for product customization. *Advanced Engineering Informatics*, 59, 102321.

Robaldo, L., Batsakis, S., Calegari, R., Calimeri, F., Fujita, M., Governatori, G., Morelli, M. C., Pacenza, F., Pisano, G., Satoh, K., & others. (2023). Compliance checking on first-order knowledge with conflicting and compensatory norms: A comparison among currently available technologies. *Artificial Intelligence and Law*, 1–51.

Romero, D., Taisch, M., Acerbi, F., Khan, M. A., Andersen, A.-L., Arioli, V., Bressanelli, G., Chari, A., Frank, A. G., Ebel, M., & others. (2023). 2023 *World Manufacturing Report: New Business Models for the Manufacturing of the Future*. World Manufacturing Foundation.

Ryan, D., Harris, E., & O'Connor, G. M. (2024). Explainable machine learning for the regulatory environment: A case study in micro-droplet printing. *Additive Manufacturing*, 104237.

Sadeghi, A., Bellavista, P., Song, W., & Yazdani-Asrami, M. (2024). Digital Twins for Condition and Fleet Monitoring of Aircraft: Towards More-Intelligent Electrified Aviation Systems. *IEEE Access*, 12, 99806-99832.

Sala, R., Pirola, F., Pezzotta, G., & Cavalieri, S. (2023). Improvement of maintenance-based Product-Service System offering through field data: A case study. *Production & Manufacturing Research*, 11(1), 2278313.

Schweitzer, F., Bitsch, G., & Louw, L. (2023). Choosing Solution Strategies for Scheduling Automated Guided Vehicles in Production Using Machine Learning. *Applied Sciences*, 13(2), Articolo 2.

Sewunetie, W. T., & Kovács, L. (2022). Comparison of template-based and multilayer perceptron-based approach for automatic question generation system. *Indonesian Journal of Electrical Engineering and Computer Science*, 28(3), 1738–1748.

Shamsuddinova, S., Heryani, P., & Naval, M. A. (2024). Evolution to revolution: Critical exploration of educators' perceptions of the impact of Artificial Intelligence (AI) on the teaching and learning process in the GCC region. *International Journal of Educational Research*, 125, 102326.

Sjödin, D., Parida, V., & Kohtamäki, M. (2023). Artificial intelligence enabling circular business model innovation in digital servitization: Conceptualizing dynamic capabilities, AI capacities, business models and effects. *Technological Forecasting and Social Change*, 197, 122903.

Stanford University. (2024). Artificial Intelligence Index Report 2024. https://aiindex.stanford.edu/report/

Taatali, A., Sadaoui, S. E., Louar, M. A., & Mahiddini, B. (2024). On-machine dimensional inspection: Machine vision-based approach. *The International Journal of Advanced Manufacturing Technology*, 131(1), 393–407.

Tamayo, J., Doumi, L., Goel, S., Kovács-Ondrejkovic, O., & Sadun, R. (2023, settembre 1). Reskilling in the Age of AI. *Harvard Business Review*. https://hbr.org/2023/09/reskilling-in-the-age-of-ai

Tejasvi, K., Vinya, V. L., Padmaja, J. N., Begum, R., & Jabbar, M. A. (2024). Explainable Artificial Intelligence (XAI) for Managing Customer Needs in E-Commerce: A Systematic Review. In L. Gaur & A. Abraham (A c. Di), *Role of Explainable Artificial Intelligence in E-Commerce* (pp. 17–31). Springer Nature Switzerland.

Toorajipour, R., Oghazi, P., & Palmié, M. (2024). Data ecosystem business models: Value propositions and value capture with Artificial Intelligence of Things. *International Journal of Information Management*, 78, 102804.

Tukker, A. (2004). Eight Types of Product-Service System: Eight Ways to Sustainability? Experiences from Suspronet. *Business Strategy and Environment*, 13, 246–260.

Uhlemann, T. H.-J., Lehmann, C., & Steinhilper, R. (2017). The digital twin: Realizing the cyber-physical production system for industry 4.0. *Procedia Cirp*, 61, 335–340.

Umbrico, A., Orlandini, A., Cesta, A., Faroni, M., Beschi, M., Pedrocchi, N., Scala, A., Tavormina, P., Koukas, S., Zalonis, A., & others. (2022). Design of advanced human–robot collaborative cells for personalized human–robot collaborations. *Applied Sciences*, 12(14), 6839.

Unioncamere. (2024, marzo 4). Intelligenza artificiale: Meno del 10% delle imprese la utilizza già.https://www.unioncamere.gov.it/comunicazione/comunicati-stampa/intelligenza-artificiale-meno-del-10-delle-imprese-la-utilizza-gia

Urgo, M., Terkaj, W., & Simonetti, G. (2024). Monitoring manufacturing systems using AI: A method based on a digital factory twin to train CNNs on synthetic data. *CIRP Journal of Manufacturing Science and Technology*, 50, 249–268.

Vandermerwe, S., & Rada, J. (1988). Servitization of business: Adding value by adding services. *European Management Journal*, 6(4), 314–324.

Walter, S. (2023). AI impacts on supply chain performance: A manufacturing use case study. *Discover Artificial Intelligence*, 3(1), 18.

Zebec, A., & Indihar Štemberger, M. (2024). Creating AI business value through BPM capabilities. *Business Process Management Journal*, 30(8), 1–26.

Zhang, B., Liu, S., & Shin, Y. C. (2019). In-Process monitoring of porosity during laser additive manufacturing process. *Additive Manufacturing*, 28, 497–505.

Autrici e autori

Il Tavolo Interdipartimentale sull'Intelligenza Artificiale dell'Università degli Studi di Bergamo è stato istituito nel novembre 2023 e riunisce ricercatori e ricercatrici degli otto Dipartimenti dell'Università degli studi di Bergamo.

Francesca Cerea è Ricercatrice di diritto privato presso il Dipartimento di Giurisprudenza. È *research fellow* presso l'Institut für Italienisches Recht di Innsbruck e l'Université Lumière Lyon 2. Tra i suoi interessi di ricerca rientrano le questioni di responsabilità civile derivanti dall'impiego di sistemi di IA nei settori della sanità e dell'*automotive*.

Stefano Coniglio è Ricercatore in Informatica presso il Dipartimento di Scienze Economiche. Ha una lunga esperienza di ricerca e insegnamento all'estero maturata in oltre dieci anni. I suoi interessi scientifici comprendono modelli e metodi di ottimizzazione matematica e di *machine learning* e aspetti algoritmici e architetturali del *deep learning*, con applicazioni in diversi settori, tra cui la medicina, l'ingegneria e l'economia.

Hagen Lehmann è Ricercatore presso il Dipartimento di Scienze Umane e Sociali, dove insegna Technologies for Caring and Learning. I suoi principali interessi di ricerca sono la didattica assistita dai robot, la tecnologia per l'insegnamento e l'apprendimento, l'interazione uomo-robot, la robotica sociale, la terapia assistita dai robot, le interfacce uomo-macchina, i principi dell'evoluzione sociale e l'*enaction*.

Maria Francesca Murru è Professoressa associata di Sociologia dei Processi Culturali e Comunicativi presso il Dipartimento di Lettere, Filosofia e Comunicazione. Ha svolto ricerca teorica ed empirica sui pubblici, sulla sfera pubblica digitale, sulle culture civiche e mediali. Ultimamente i suoi studi si concentrano sulle dinamiche sociali ed epistemiche del pensiero

cospirazionista e sull'adozione di sistemi di intelligenza artificiale nei processi di *newsmaking*.

Giuseppe Previtali è Ricercatore in Cinema, Fotografia, Televisione e Media Audiovisivi presso il Dipartimento di Lingue, Letterature e Culture Straniere. Insegna cinema e cultura visuale e si occupa prevalentemente di forme estreme della visualità contemporanea, media e *digital literacy* ed epistemologia critica delle *digital humanities*.

Federico Leo Redi è Ricercatore presso il Dipartimento di Ingegneria e Scienze Applicate. Attivo dal 2012 in vari esperimenti al CERN di Ginevra, la sua attività si concentra sui problemi sperimentali della fisica fondamentale, con particolare attenzione alle ricerche dirette di nuova fisica. I suoi studi sono estesi anche ai rivelatori e alle loro simulazioni parametriche. Si occupa infine di problemi fenomenologici della materia oscura.

Roberto Sala è Ricercatore presso il Dipartimento di Ingegneria Gestionale, dell'Informazione e della Produzione dove è anche membro del gruppo di ricerca CELS. Svolge attività di ricerca sui temi della manutenzione, Industria 4.0/5.0, sistemi prodotto-servizio e *natural language processing*. Ha partecipato a numerosi progetti a livello regionale, nazionale ed europeo in questi ambiti di ricerca.

Gabriele Torri è Ricercatore presso il Dipartimento di Scienze Aziendali. Si occupa di modelli quantitativi per la finanza, con un focus specifico sulla gestione di portafoglio, gli investimenti sostenibili e la misurazione del rischio sistemico. Insegna Matematica Finanziaria, Asset Pricing and Risk Analysis e Quantitative Methods for Business and Data Analysis.