Alfio Quarteroni

L'intelligenza creata

L'AI e il nostro futuro

HOEPLI

L’INTELLIGENZA CREATA

ALFIO QUARTERONI

L’INTELLIGENZA CREATA

L’AI e il nostro futuro

EDITORE ULRICO HOEPLI MILANO

Copyright © Ulrico Hoepli Editore S.p.A. 2025
via Hoepli 5, 20121 Milano (Italy)
e-mail hoepli@hoepli.it

www.hoeplieditore.it

Seguici su Twitter: @Hoepli_1870

Tutti i diritti sono riservati a norma di legge
e a norma delle convenzioni internazionali

ISBN EBOOK 978-88-360-2009-6

Realizzazione editoriale: Ibidem, ibidemstudio.it
Copertina: Emanuele Ragnisco
Immagini realizzate da Paola Gervasio

Realizzazione digitale: Promedia, Torino

SOMMARIO

Ringraziamenti
Introduzione
1. Le prime pagine dell'AI: tra sogno e realtà
2. AI o non AI?
3. Machine learning, ovvero quando i computer imparano (da soli!)
4. Matematica... il minimo sindacale
5. Quali dati per l'apprendimento, e quale apprendimento?
6. AI generativa: il boom all'improvviso
7. Non di sola generativa vive l'AI
8. L'AI nella scienza
9. Uno sguardo al futuro: a che punto siamo?
10. Intelligenza artificiale e intelligenza naturale
11. Scatole nere, bianche e grigie: oltre l'AI, dai Big Data alla Big Science
12. Dialogando con ChatGPT
13. Conclusione, con alcuni spunti di riflessione

Addendum
Glossario
Riferimenti bibliografici
Circa l'autore
Informazioni sul Libro

A Lara, Luca, Bianca Sofia e Leonardo Alexander,
perché non abbiano mai a temere il loro futuro.

Ringraziamenti

Ringrazio Francesca Bonadei, Paola Gervasio, David Moscato, Silvia Quarteroni, Francesco Regazzoni e Luigi Simeone per le loro osservazioni e i loro consigli su una versione preliminare di questo testo.

Introduzione

La questione dell'intelligenza artificiale (AI) e del suo impatto su vari aspetti della vita umana è un tema che richiede una riflessione profonda.

Quando parliamo di intelligenza artificiale, non possiamo fare a meno di porci una domanda che mi viene rivolta spesso quando intervengo in pubblico: "Ma è davvero intelligenza, come quella umana?" Un quesito naturale, che ne porta con sé molti altri: "Se l'AI fosse davvero intelligente, potrebbe mai superare le capacità umane? E potrebbe arrivare un punto di non ritorno, in cui l'umanità diventerà succube di queste macchine? Che ne sarà dei posti di lavoro in un mondo dominato dall'AI? E come possiamo proteggere la nostra privacy e i nostri dati da un loro uso indiscriminato?"

Cercherò di rispondere a queste domande, ma anche di spiegare perché, per alcune di esse, non esista una risposta definitiva o esaustiva. Il mio obiettivo è quello di fare chiarezza su cosa sia realmente l'intelligenza artificiale, sfatando miti e falsi concetti, e fornendo un quadro rigoroso e non ambiguo di questa rivoluzione tecnologica. Delineando il suo enorme potenziale, i successi già raggiunti, le speranze che alimenta, ma anche i rischi che presenta, non solo per noi come individui, ma per la società nel suo complesso.

L'intenzione non è quella di rincorrere l'ultima novità dell'AI generativa o l'ultima architettura di rete neurale profonda. Piuttosto, mi concentrerò sui concetti fondamentali che reggono l'AI, concetti destinati a durare nonostante il ritmo incalzante

dell’innovazione. Mostrerò anche il legame tra l’AI generativa e le altre forme di AI, chiarendo che non tutte le applicazioni che fanno uso di algoritmi possono essere etichettate come “intelligenza artificiale”.

Sin dagli anni ’50, pionieri come Alan Turing e John McCarthy hanno gettato le basi di quella che oggi definiamo AI. Da allora, la definizione stessa di AI è cambiata continuamente, riflettendo la molteplicità degli obiettivi che ha assunto: dalla cosiddetta AI debole, o ristretta, che ha obiettivi limitati, all’AI generale, simile all’intelligenza umana, fino all’idea di una superintelligenza artificiale che, in teoria, potrebbe persino superare quella umana. Al momento, siamo ancora dentro i confini dell’AI ristretta, che aiuta l’uomo in compiti specifici senza entrare in competizione con l’intelligenza umana in senso più ampio.

Ciononostante, l’AI possiede abilità straordinarie, che noi umani fatichiamo a replicare: la capacità di elaborare enormi quantità di dati, di sintetizzare rapidamente informazioni complesse, anticipare comportamenti di consumo con una precisione sorprendente e, da poco, generare testi, poesie e racconti, immagini e film, in frazioni di secondo. Nella sanità, poi, le reti neurali possono scovare diagnosi che potrebbero sfuggire agli specialisti. In molti casi, l’AI sfida la nostra concezione dell’agire umano e della presa di decisioni, sollevando interrogativi sul controllo e sulla responsabilità.

Il cuore del progresso dell’AI risiede nella capacità dei computer di apprendere autonomamente, principalmente attraverso l’uso delle reti neurali artificiali. Questo processo si basa sulla capacità di addestramento di queste ultime grazie alla disponibilità di dati (spesso, ma non sempre, di *Big Data*) più che sulla teoria scientifica tradizionale. Naturalmente i dati di addestramento devono essere significativi relativamente al compito che si intende affrontare. In un certo senso, l’esperienza codificata nei dati ha preso il sopravvento sulla teoria, sovvertendo il paradigma della scienza moderna.

Esplorerò le aree in cui l’AI ha ottenuto i risultati più sorprendenti, offrendo previsioni sul suo impatto futuro e sulle

problematiche aperte, come l'enorme consumo di risorse necessarie per addestrare reti neurali profonde. È un tema che non possiamo ignorare, poiché il *carbon footprint* dell'AI cresce vertiginosamente e pone seri problemi sulla sostenibilità ambientale di questo sviluppo.

Un altro tema di grande rilevanza sociale è l'impatto sul mondo del lavoro: se da un lato l'AI eliminerà molte professioni, soprattutto quelle ripetitive e basate su complesse elaborazioni di dati, dall'altro ne creerà nuove. Ma la transizione non sarà senza difficoltà, e molte persone potrebbero trovarsi escluse da questa trasformazione. Per questo motivo, sarà necessario adottare politiche proattive in ambito educativo e formativo, per garantire che i benefici dell'AI siano il più possibile equamente distribuiti.

L'AI rappresenta un cambiamento epocale nel modo in cui interagiamo con il mondo e tra di noi. Nell'espletamento di tanti compiti supera di gran lunga le nostre capacità, tuttavia i suoi processi sono spesso opachi, creando il cosiddetto effetto *black-box*: nemmeno i suoi creatori possono spiegare sempre come un sistema di AI prenda certe decisioni. Questo fenomeno ci porta a ripensare concetti fondamentali come responsabilità e fiducia, soprattutto quando l'AI viene applicata in ambiti critici come la salute, la giustizia o la guerra.

Un altro argomento riguarda l'impatto dell'intelligenza artificiale sulla governance e la geopolitica. Oggi, l'AI sta assumendo un ruolo strategico nelle relazioni di potere globale, simile a quello che la tecnologia nucleare rivestì nel XX secolo. Le nazioni all'avanguardia nello sviluppo dell'AI potrebbero ottenere un'influenza sproporzionata, attraverso il dominio economico, la superiorità militare, il controllo dei dati. Questo scenario solleva questioni cruciali su nuove forme di disuguaglianza globale e sull'eventuale innesco di una corsa agli armamenti alimentata dalla tecnologia. L'AI ha il potenziale per ridisegnare gli equilibri di potere tradizionali e diventa quindi fondamentale che i governi elaborino politiche lungimiranti per la regolamentazione e l'integrazione di questi sistemi nella società. L'intelligenza

artificiale è già impiegata per ottimizzare i servizi pubblici, prevenire reati e persino influenzare i comportamenti elettorali attraverso algoritmi mirati. Tuttavia, il suo utilizzo nella governance porta con sé preoccupazioni legate alla privacy e al rispetto delle libertà individuali. Senza una gestione attenta, l'AI potrebbe trasformarsi in uno strumento di sorveglianza o manipolazione dell'opinione pubblica, mettendo a rischio i fondamenti stessi della democrazia.

Gli esseri umani tendono a essere attratti da tutto ciò che è innovativo, ma allo stesso tempo provano timore verso ciò che non comprendono completamente o non riescono a controllare. È naturale temere ciò che è sconosciuto. La stessa denominazione "intelligenza artificiale" evoca una sfida. Indipendentemente dalle sensazioni che ognuno di noi può avvertire, l'AI sta diventando onnipresente ed è probabilmente destinata a innescare una trasformazione epocale, nella società, nell'economia, nella politica. In questo libro non voglio né celebrarla né demonizzarla. Cercherò invece di considerare le sue implicazioni finché queste rimangono nel regno della comprensione umana. L'obiettivo è di fornire strumenti per comprendere meglio questo fenomeno, promuovendo un approccio critico e consapevole nei confronti della trasformazione che stiamo vivendo. E anche di provare a immaginare un futuro che vada *oltre* l'AI, o meglio che *integri* in modo sempre più funzionale l'AI con la conoscenza scientifica che ha consentito alla nostra umanità di raggiungere, nello scorso secolo, i progressi sensazionali che tutti conosciamo.

1
LE PRIME PAGINE DELL'AI: TRA SOGNO E REALTÀ

Nella prima metà del XX secolo, la fantascienza ha reso familiare il concetto di robot (artificialmente) intelligenti. Tutto è iniziato con il robot umanoide che impersonava Maria in *Metropolis* di Fritz Lang nel 1927 ed è continuato con il "cuore senza cuore" dell'Uomo di Latta nel film *Il Mago di Oz* diretto da Victor Fleming nel 1939.

Negli anni '50, una generazione di scienziati, matematici e filosofi aveva ormai assimilato culturalmente il concetto di intelligenza artificiale. Tra questi c'era Alan Turing, un giovane britannico geniale che, oltre al suo sensazionale contributo allo sviluppo di *Enigma*, la macchina che ha permesso di decodificare i messaggi tedeschi durante la Seconda guerra mondiale (Copeland 2024), fu un autentico pioniere nell'esplorazione dei fondamenti matematici dell'intelligenza artificiale. Turing suggerì che se gli esseri umani usano le informazioni disponibili e la ragione per risolvere problemi e prendere decisioni, perché non potrebbero fare lo stesso le macchine? Questo era il quadro logico del suo articolo del 1950, *Computing Machinery and Intelligence*, pubblicato sulla rivista "Mind" (Turing 1950). In questo articolo, Turing formulò la celeberrima domanda: le macchine possono pensare? Domanda, precisava lo stesso Turing, per rispondere alla quale sarebbe innanzitutto necessario fornire una precisa definizione di *macchina* e di

pensiero. Turing discusse come costruire macchine intelligenti e come testare la loro intelligenza, grazie a quello che in seguito sarebbe diventato il celebre *test di Turing*: mettiamo un computer in una stanza, una persona in un'altra, si neutralizzi qualunque elemento indiziario (voce, calligrafia ecc.) e si cerchi, attraverso dei quesiti, di far indovinare a un terzo soggetto, all'oscuro di tutto, con chi stia parlando, se con la persona o il computer. La macchina, naturalmente, ha la facoltà di "imbrogliare". L'idea è che, se gli algoritmi con cui è stata "allenata" sono abbastanza potenti, la macchina può diventare intelligente a tal punto da "travestirsi" da essere umano, non facendosi individuare dal soggetto terzo. Usando la stessa articolazione di Turing,

"Intelligence is demonstrated by a machine's ability to engage in natural language conversation and deceive a human into believing it is human."[1]

In questo modo, Turing eluse secoli di discussioni filosofiche sulla natura dell'intelligenza. Il "gioco dell'imitazione", che introdusse, suggeriva che se una macchina fosse in grado di operare così bene da rendere impossibile per gli osservatori distinguere il suo comportamento da quello di un essere umano, allora si dovrebbe etichettare la macchina come intelligente.

Turing fu però troppo precursore e non poté dare un seguito operativo alle sue intuizioni. A quel tempo, ai computer mancava un prerequisito fondamentale per l'intelligenza: non potevano memorizzare comandi, ma solo eseguirli. Potevano essere istruiti su *cosa fare*, ma non potevano ricordare *cosa avevano fatto*. Gli elevatissimi tempi di calcolo e gli elevati costi rappresentavano un altro grande ostacolo. Questi e altri elementi fecero dire a Hans Moravec, all'epoca uno studente di dottorato di John McCarthy (l'autore della prima definizione di AI, lo scopriremo fra poco), che "i computer erano ancora milioni di volte troppo deboli per mostrare intelligenza". Nei primi anni '50 del secolo scorso, il costo per noleggiare un computer poteva arrivare fino a 200.000 dollari al mese. Essi erano pertanto accessibili solo alle

università più ricche e prestigiose e alle grandi aziende tecnologiche americane. Era necessario convincere le agenzie di finanziamento che la ricerca sull'intelligenza delle macchine valesse la pena di essere sostenuta.

Pochi anni dopo, Allen Newell, Herbert A. Simon e Cliff Shaw scrissero *Logic Theorist*, un software deliberatamente progettato per eseguire ragionamenti automatizzati. Logic Theorist, oggi considerato il primo programma di intelligenza artificiale, fu capace di dimostrare 38 dei primi 52 teoremi nel secondo capitolo dei *Principia Mathematica* di Alfred N. Whitehead e Bertrand Russell, trovando addirittura nuove e più brevi dimostrazioni per alcuni di essi. Logic Theorist aveva come ambizione quella di imitare le capacità di risoluzione dei problemi di un essere umano e per questo venne finanziato dall'allora Research and Development (RAND) Corporation.

Una definizione mutevole

Nel 1955, quando Newell e Simon iniziarono a lavorare al Logic Theorist, il campo dell'intelligenza artificiale non esisteva ancora. Anche il termine stesso di intelligenza artificiale non sarebbe stato coniato fino all'estate successiva. In effetti, il battesimo si ebbe nel corso di una storica conferenza, *The Dartmouth Summer Research Project on Artificial Intelligence*, al Dartmouth College a Hanover nel New Hampshire nell'agosto 1956. In questa conferenza, John McCarthy, Marvin Minsky, Nathaniel Rochester e Claude Shannon riunirono i migliori ricercatori di vari campi per una discussione aperta sull'intelligenza artificiale, termine che McCarthy coniò proprio per quell'occasione. Nella sua proposta, McCarthy, un giovane professore ventottenne di matematica del Dartmouth College, affermò che la conferenza doveva "procedere sulla base della congettura che ogni aspetto dell'apprendimento o qualsiasi altra caratteristica dell'intelligenza potesse essere descritto in modo così preciso da

poter essere simulato da una macchina". E propose la seguente definizione di AI:

"*The science and engineering of making intelligent machines, especially intelligent computer programs*."[2]

È importante osservare come, analogamente a quanto fatto da Turing, si accosti il concetto di intelligenza a quello di macchina (o computer). Fino a poco tempo prima, si riteneva che le macchine, in particolare i computer, non potessero essere intelligenti e che solo gli esseri umani fossero capaci di perseguire obiettivi in totale autonomia e prendere decisioni.

Nonostante il sostanziale insuccesso nello stabilire metodi standard per il nuovo campo, tutti i partecipanti alla conferenza concordarono con entusiasmo che l'AI fosse realizzabile. In effetti, l'evento del Dartmouth College rappresentò un volano per i successivi vent'anni di ricerca sull'AI.

Da questa premessa si comprende come i confini dell'AI non siano tracciati in modo indelebile: essi in effetti sono soggettivi (la percezione di ciò che è prerogativa umana varia da persona a persona) e si spostano anche nel tempo man mano che nuovi traguardi vengono raggiunti. Questo fenomeno, noto come *AI effect*, è stato ben sintetizzato nel 1970 dalla frase dell'informatico statunitense Larry Tesler:

"*AI is whatever hasn't been done yet*."[3]

Molte altre definizioni sarebbero state proposte in seguito, anche da parte di istituzioni pubbliche e organizzazioni culturali. Le intuizioni di Turing e McCarthy sull'AI sono comunque diventate dei punti di riferimento da allora nel riposizionare l'attenzione nella definizione di intelligenza verso le prestazioni (comportamenti che *sembrano* intelligenti) piuttosto che sulle dimensioni filosofiche, cognitive o neuroscientifiche più profonde del termine intelligenza.

Ad esempio Marvin Minsky, uno dei pionieri dell'AI, nel 1985 formulò questa definizione:

"*AI is the science of making machines do things that would require intelligence if done by men*."[4]

Secondo il MIT (Massachusetts Institute of Technology),

"*AI is the quest to build machines that can think, act, and learn like humans*."[5]

Questa definizione cattura un'idea più generale di cognizione e apprendimento, piuttosto che riferirsi a obiettivi specifici.

Nel 2019 la Commissione Europea ha proposto la seguente definizione:

"*Artificial intelligence (AI) refers to systems that display intelligent behavior by analyzing their environment and taking actions – with some degree of autonomy – to achieve specific goals*."[6]

Si sottolinea in questo caso l'autonomia nel perseguire specifici obiettivi. Secondo l'OECD (Organization for Economic Co-operation and Development) AI è

"*A machine-based system that is capable of influencing the environment by producing an output (predictions, recommendations or decisions) for a given set of objectives. It uses machine and/or human-based data and inputs to: (i) perceive real and/or virtual environments; (ii) abstract these perceptions into models through analysis in an automated manner (e.g. with machine learning), or manually; and (iii) use model inference to formulate options for outcomes. AI systems are designed to operate with varying levels of autonomy*."[7]

Secondo l'Oxford English Dictionary, l'AI è

"The capacity of computers or other machines to exhibit or simulate intelligent behaviour",[8]

una definizione che si focalizza sull'abilità delle macchine di simulare comportamenti che abitualmente associamo all'intelligenza umana.

Come ci si poteva aspettare da quel lontano 1956, la definizione di AI è cambiata nel corso degli anni ma sostanzialmente restando all'interno di un perimetro in cui si enfatizza la simulazione delle capacità umane, il comportamento razionale, l'apprendimento autonomo e la capacità di decisione. Ancora oggi, non esiste una definizione universalmente riconosciuta. Secondo la scuola di pensiero più diffusa,

"L'AI è l'insieme di quelle abilità dimostrate dai sistemi hardware e software che, agli occhi di un osservatore umano, potrebbero sembrare esclusive dell'intelligenza umana".

Inverni e primavere

Riportiamo per un momento l'orologio indietro a quell'anno di grazia, il 1956. Dal 1957 al 1974, l'AI prosperò. I computer potevano memorizzare più informazioni e divennero più veloci, economici e accessibili. Gli algoritmi di apprendimento automatico, quelli che sono alla base del *Machine Learning* (ML) ovvero della capacità dei computer di apprendere autonomamente (li incontreremo nel capitolo 5 e ci faranno compagnia attraverso l'intero libro), diventarono sempre più efficaci e le persone più abili nel sapere quale algoritmo applicare al proprio problema. Fra le prime dimostrazioni possiamo citare il General Problem Solver (GPS), un programma per computer creato nel 1957 da Herbert A. Simon, J. Cliff Shaw e Allen Newell, con lo scopo di risolvere problemi *generali* (purché fossero adeguatamente formalizzati). Venne creato

principalmente per risolvere problemi matematici di geometria e anche per giocare a scacchi. Un altro è *ELIZA*, un *chatbot*, ovvero un programma in grado di interagire vocalmente con una persona, scritto nel 1966 da Joseph Weizenbaum, un analizzatore lessicale e un insieme di regole (o script) capaci di simulare una conversazione in diverse lingue (inglese, gallese o tedesco). Sia GPS sia ELIZA fornivano le prime evidenze in merito alla possibilità di costruire programmi capaci di interpretare il linguaggio scritto e parlato. (Interessante osservare che il nome ELIZA traeva spunto da Eliza Doolittle, protagonista del *Pigmalione* di George Bernard Shaw.)

Questi successi, insieme al sostegno dei principali ricercatori, convinsero le agenzie governative come la Defense Advanced Research Projects Agency (DARPA) a finanziare la ricerca sull'AI in diverse istituzioni. Il governo era particolarmente interessato a una macchina che potesse trascrivere e tradurre il linguaggio parlato, oltre che elaborare grandi quantità di dati. L'ottimismo era alto e le aspettative erano ancora più alte, al punto da indurre il già citato Marvin Minsky (uno dei quattro co-organizzatori della Conferenza del Dartmouth College) a scrivere nel 1970 su "Life Magazine":

"*Tra tre e otto anni avremo una macchina con l'intelligenza generale di un essere umano medio*."

Superfluo osservare che la profezia non si è avverata, come peraltro molte altre, divenute ormai celebri, nel campo dei computer.[9]

Per il grande pubblico è utile ricordare come Minsky sia stato uno dei principali consiglieri di Stanley Kubrick per la realizzazione del celeberrimo film *2001 Odissea nello Spazio*. Minsky nel 1969 vinse anche il prestigioso Turing Award per la sua influenza nella nascita dell'AI. McCarthy se lo aggiudicò nel 1971.

Mentre la *proof of concept* era lì, c'era ancora molta strada da fare prima di poter raggiungere gli obiettivi finali del

processamento del linguaggio naturale, del pensiero astratto e dell'auto-riconoscimento. Negli anni '80, l'AI venne rilanciata grazie a due fattori: un'evoluzione delle tecniche algoritmiche e un aumento dei finanziamenti. John Hopfield e David Rumelhart resero popolari le tecniche di *deep learning* che permettevano ai computer di apprendere dall'esperienza. Hopfield avrebbe vinto il Nobel per la Fisica nel 2024, grazie a questi studi pionieristici. Incontreremo le reti neurali artificiali profonde, la base del deep learning, più avanti. D'altra parte, Edward Feigenbaum introdusse i *sistemi esperti*, che imitavano il processo decisionale di un esperto umano. Questi programmi chiedevano a un esperto di un determinato campo come rispondere in specifiche situazioni e, una volta apprese le risposte per quasi ogni situazione, i non esperti potevano ricevere consigli dal programma. Questi sistemi furono ampiamente utilizzati nelle industrie. Il governo giapponese finanziò massicciamente i sistemi esperti e altri progetti legati all'AI come parte del loro Progetto di Computer di Quinta Generazione (FGCP). Dal 1982 al 1990 vennero investiti 400 milioni di dollari con l'obiettivo di rivoluzionare l'elaborazione dei computer, implementare la programmazione logica e migliorare l'intelligenza artificiale. Purtroppo, la maggior parte degli obiettivi più ambiziosi non fu raggiunta. Tuttavia, si potrebbe sostenere che gli effetti indiretti del FGCP ispirarono una nuova generazione di ingegneri e scienziati di talento. Nonostante ciò, i finanziamenti del FGCP cessarono e l'AI sembrò uscire di scena.

Paradossalmente, in assenza di finanziamenti governativi e senza clamore pubblico, l'AI tornò a prosperare. Negli anni '90 e 2000 molti degli obiettivi fondamentali dell'intelligenza artificiale furono raggiunti. Nel 1997, il campione mondiale di scacchi e gran maestro Garry Kasparov fu sconfitto da *Deep Blue* di IBM, un programma di scacchi per computer. Questa partita, molto pubblicizzata, rappresentò un grande passo avanti verso un programma decisionale dotato di intelligenza artificiale. Nello stesso anno, il software di riconoscimento vocale sviluppato da Dragon Systems fu implementato su Windows, segnando un

ulteriore progresso, questa volta nell'interpretazione del linguaggio parlato. Sembrava che non ci fosse un problema che le macchine non potessero affrontare. Anche le emozioni umane erano alla loro portata, come dimostrato da Kismet,[10] una testa robotica creata negli anni '90 presso il MIT da Cynthia Breazeal come esperimento di *affective computing*, una macchina che può riconoscere e simulare emozioni.

Un cambio di paradigma

Cos'era cambiato per consentire questi spettacolari progressi? Una prima decisiva spiegazione è da ricercare nel miglioramento della tecnologia. Il limite fondamentale della memoria del computer che ci bloccava 30 anni prima non era più un problema. La *legge di Moore*, una congettura formulata nel 1965 dal cofondatore di Intel Gordon Moore secondo la quale la memoria e la velocità dei computer raddoppiano ogni 18 mesi, aveva finalmente raggiunto e in molti casi superato le nostre esigenze. È proprio così che *Deep Blue* è stato in grado di sconfiggere Garry Kasparov nel 1997, così come *Alpha Go* è stato in grado di sconfiggere il campione cinese di Go, Ke Jie, nel 2017.

Questo creò in alcuni l'illusione che si potesse spiegare l'effetto montagne russe che aveva caratterizzato l'evoluzione della ricerca in AI sin dagli inizi: esauriamo le capacità dell'AI al livello della nostra attuale potenza di calcolo (memoria e velocità di elaborazione del computer) e poi aspettiamo che la *legge di Moore* venga di nuovo in nostro soccorso. Ma, soprattutto, questo progresso fu essenzialmente dovuto a un cambio di paradigma nell'approccio all'AI. Mentre in settori che richiedono caratterizzazioni precise, come il gioco degli scacchi, la manipolazione algebrica e l'automazione dei processi aziendali, l'intelligenza artificiale aveva fatto notevoli progressi, in altri ambiti, come la traduzione linguistica e il riconoscimento visivo degli oggetti, l'ambiguità intrinseca ne aveva ostacolato lo

sviluppo. Questi sistemi formali e rigidi hanno avuto successo in quei contesti in cui le attività potevano essere eseguite mediante la codifica di regole chiare.

Dalla fine degli anni ’80 e per tutti gli anni ’90, il campo dell’AI era entrato in una fase conosciuta come “inverno dell’AI”. Quando applicata a compiti più dinamici, l’intelligenza artificiale si era dimostrata fragile, producendo risultati che non superavano il test di Turing, ovvero che non raggiungevano o imitavano le prestazioni umane. Negli anni ’90 si verificò la svolta. I ricercatori compresero che era necessario adottare un nuovo approccio, che permettesse alle macchine di apprendere *in autonomia*, producendo un cambiamento di paradigma: passare dall’idea di codificare intuizioni umane nelle macchine a quella di *delegare* il processo di apprendimento stesso a queste ultime. Sebbene l’apprendimento automatico esistesse sin dagli anni ’50, sono state le nuove scoperte a consentire applicazioni pratiche. I metodi più efficaci in questo campo estraggono schemi da ampi set di dati attraverso reti neurali. Ad esempio, per riconoscere l’immagine di un gatto, i ricercatori hanno capito che una macchina deve imparare diverse rappresentazioni visive dei gatti osservando gli animali in vari contesti. In questo modo, ciò che conta per l’apprendimento automatico è la pluralità delle diverse rappresentazioni di un oggetto, piuttosto che il suo ideale (o, peggio ancora, la sua definizione). I moderni algoritmi di AI hanno la capacità di apprendere tipi e schemi, ad esempio raggruppamenti di parole che si trovano spesso insieme quando si voglia analizzare il linguaggio naturale, o caratteristiche più frequentemente presenti in un’immagine quando si tratta di un gatto, e poi di dare un senso alla realtà identificando reti di somiglianze e analogie con ciò che l’AI già conosceva. Anche se l’AI non potrà probabilmente mai *conoscere* nello stesso modo in cui lo sa fare una mente umana, un accumulo di corrispondenze con gli schemi della realtà potrebbe approssimare e talvolta superare le prestazioni della percezione e della ragione umane.

È così che ha preso forma il moderno campo dell’apprendimento automatico, o machine learning, costituito

da programmi che apprendono attraverso l'esperienza.

Parallelamente allo sviluppo della strategia algoritmica che ha generato il moderno approccio all'apprendimento automatico, gli scienziati non hanno smesso di riflettere su quanto ambiziosi si potesse essere nel perseguire il sogno dell'intelligenza artificiale. Lungo questa linea di pensiero l'AI è stata declinata in tanti modi, in ragione degli obiettivi che ci si era proposti di raggiungere. Le definizioni che si sono imposte nel corso del tempo possono essere riassunte nei termini seguenti.

- L'*AI ristretta*, detta anche *AI debole*, si riferisce all'automazione di compiti specifici, ma non ha la capacità di sviluppare una comprensione al di fuori dell'ambito per il quale è stata programmata. Esempi sono il riconoscimento di un animale in una fotografia o la traduzione di un testo in un'altra lingua. Altri esempi includono assistenti vocali come Siri e Alexa, sistemi di raccomandazione, e le auto a guida autonoma.
- L'*AI forte*, detta anche *AI generale*, invece, implica la capacità di apprendere e pensare *come gli esseri umani*, ovvero si riferisce a sistemi con la capacità di comprendere, apprendere e applicare conoscenze su una vasta gamma di compiti, con performance simili alle capacità cognitive di un essere umano. Attualmente, l'AI debole può in diversi ambiti addirittura superare gli esseri umani nei compiti specifici per cui è stata progettata, ma opera con molti più vincoli rispetto all'intelligenza umana anche più elementare. Tutte le AI disponibili oggi possono essere considerate AI deboli. Possiamo quindi affermare che l'AI forte, oggi, ancora non esista (Togelius 2024).
- Infine, la cosiddetta *Superintelligenza Artificiale* è un concetto del tutto teorico in cui l'AI supererebbe l'intelligenza umana in tutti gli aspetti, inclusi la creatività, la risoluzione di ogni classe di problemi e sarebbe dotata di un'intelligenza sociale. Rappresenta un punto in cui le

> macchine non solo eguagliano, ma superano di gran lunga le capacità intellettuali umane.

L'intelligenza artificiale generale è un concetto ambiguo e fuorviante, e lo è ancor di più quello di superintelligenza, soprattutto perché non abbiamo una chiara definizione né di "intelligenza" né di "generalità". Sebbene continuiamo a fare progressi impressionanti nella tecnologia dell'AI, siamo ancora lontani dall'avere un sistema unico in grado di compiere tutte le attività che associamo all'idea di intelligenza.

Altri esempi di ambienti di sviluppo proposti nell'ambito dell'AI includono *macchine reattive* e la cosiddetta *teoria della mente*. Le macchine reattive sono sistemi AI basilari che reagiscono a input specifici con risposte pre-programmate. Non hanno funzionalità basate sulla memoria né capacità di apprendimento. Un esempio è *Deep Blue* di IBM, programmato per giocare a scacchi e sconfiggere il campione Garry Kasparov. La teoria della mente è un tipo di AI ancora in fase di ricerca con l'ambizione di comprendere e replicare le emozioni e i pensieri umani. Questo tipo di AI sarebbe in grado di percepire, interpretare e rispondere alle emozioni e alle interazioni sociali umane, permettendo interazioni più sfumate e adattive con gli esseri umani.

Queste classificazioni delineano lo sviluppo e gli obiettivi della ricerca sull'AI, tracciando un percorso da sistemi specifici e semplici a entità complesse e autoconsapevoli. Il passaggio dall'AI debole alla forte implicherebbe progressi significativi nella comprensione della cognizione umana, nell'apprendimento automatico e nelle considerazioni etiche.

Quando si legge e si scrive di AI, è rilevante il contesto in cui viene utilizzato il termine ed è opportuno ricordare che il concetto stesso di AI è tuttora oggetto di ampio dibattito tra ingegneri, scienziati e filosofi. Ne daremo maggiore concretezza nei due capitoli che seguono. Il futuro dell'AI offrirà spazio a molteplici prospettive e approcci tecnici diversi.

[1] L'intelligenza è dimostrata dalla capacità di una macchina di impegnarsi in una conversazione in linguaggio naturale e di ingannare un umano facendogli credere di essere umana.

[2] La scienza e l'ingegneria della costruzione di macchine intelligenti, in particolare di programmi informatici intelligenti.

[3] AI è tutto ciò che non è ancora stato fatto.

[4] AI è la scienza che permette alle macchine di fare cose che richiederebbero intelligenza se fatte da uomini.

[5] AI è la ricerca di macchine in grado di pensare, agire e imparare come gli esseri umani.

[6] L'intelligenza artificiale (AI) si riferisce a sistemi che mostrano un comportamento intelligente analizzando l'ambiente circostante e intraprendendo azioni – con un certo grado di autonomia – per raggiungere obiettivi specifici.

[7] Un sistema basato su macchine in grado di influenzare l'ambiente producendo un output (previsioni, raccomandazioni o decisioni) per un determinato insieme di obiettivi. Utilizza dati e input basati sulla macchina e/o sull'uomo per: (*i*) percepire ambienti reali e/o virtuali; (*ii*) astrarre queste percezioni in modelli attraverso l'analisi in modo automatizzato (ad esempio con l'apprendimento automatico) o manuale; e (*iii*) utilizzare l'inferenza del modello per formulare opzioni per i risultati. I sistemi di AI sono progettati per operare con vari livelli di autonomia (https://oecd.ai/en/ai-principles).

[8] La capacità dei computer o di altre macchine di mostrare o simulare un comportamento intelligente.

[9] Si veda ad esempio https://www.pcworld.com/article/532605/worst_tech_predictions.html.

[10] Il nome Kismet deriva da una parola turca che significa destino o a volte fortuna, https://robotsguide.com/robots/kismet.

2
AI O NON AI?

È interessante osservare come, nella percezione comune, regni una diffusa confusione su *cosa sia* (o *cosa non sia*) l'AI. Spesso, si tende ad attribuire all'AI molti processi che sfuggono alla nostra comprensione, talvolta addirittura a confondere con AI qualunque azione che presupponga la presenza di un algoritmo (un altro grande incompreso dei nostri tempi). Ancora peggio, molto frequentemente, AI è considerato sinonimo di ChatGPT (un sistema di intelligenza artificiale generativa), ma quest'ultima esprime solo un sottoinsieme dell'intero mondo AI. Naturalmente questa confusione si può spiegare, almeno parzialmente: abbiamo notato in precedenza come la definizione stessa di AI sia cambiata molte volte nel corso degli anni. Tuttavia, qualche elemento di chiarezza è opportuno stabilirlo, vista la delicatezza dell'argomento e la sua immanenza nella nostra società.

Un fattore che a me sembra distintivo e peculiare dell'AI, oggi, è la presenza della componente di machine learning, ovvero la capacità di apprendimento autonomo degli algoritmi.

Algoritmi che apprendono autonomamente

Gli algoritmi tradizionali sono sequenze di istruzioni, tipicamente formulate in termini di enti e operazioni matematiche, che seguono regole rigide e predefinite. Funzionano in modo deterministico, cioè il loro comportamento e il loro output sono prevedibili e ripetibili, determinati dai dati di input e dalle istruzioni che li governano. Non cambiano né si adattano in base alle esperienze passate o ai nuovi dati.

La caratteristica distintiva degli algoritmi di AI basati sul machine learning è che, attraverso l'apprendimento automatico, essi modificano i propri comportamenti e risultati in base ai dati che ricevono. Questo implica una capacità di "evoluzione" o "adattamento" che non esiste negli algoritmi tradizionali. L'AI utilizza tecniche come l'apprendimento supervisionato, non supervisionato o rinforzato per migliorarsi nel tempo (lo vedremo nel capitolo 3): può gestire dati complessi, fare previsioni o identificare schemi senza che ogni singolo comportamento sia specificato da un programmatore umano.

Dunque, mentre un algoritmo tradizionale è statico (non si evolve con i nuovi dati), un sistema di AI apprende e si adatta. È capace di generalizzare partendo da dati e modelli preesistenti per applicarli a nuovi casi o scenari che non ha mai incontrato prima. Questo *apprendimento adattivo* è cruciale per compiti come il riconoscimento delle immagini, la comprensione del linguaggio naturale o le raccomandazioni personalizzate.

Negli ultimi cinquant'anni, le macchine non sono state in grado di dimostrarsi intelligenti secondo il test di Turing, ma questo blocco sembra stia per essere superato. Fino a un paio di decenni fa, i computer hanno lavorato seguendo codici rigorosamente definiti, il che ha portato a risultati altrettanto limitati e statici. I programmi tradizionali erano capaci di gestire grandi volumi di dati e di effettuare calcoli complessi, ma faticavano a riconoscere immagini di oggetti semplici o ad adattarsi a input imprecisi. La natura incerta e concettuale del pensiero umano ha rappresentato un notevole ostacolo nello sviluppo dell'intelligenza artificiale. Tuttavia, negli ultimi due decenni, le innovazioni tecnologiche hanno dato vita a forme di

intelligenza artificiale che hanno cominciato a uguagliare o addirittura superare le performance umane in questi ambiti. A titolo di esempio, gli algoritmi dell'intelligenza artificiale generativa traducono i testi non semplicemente sostituendo parole, ma identificando e utilizzando frasi idiomatiche e schemi linguistici. Inoltre, gli algoritmi sono dinamici, poiché evolvono in risposta a situazioni in continuo cambiamento, e si presentano come emergenti, in quanto riescono a scoprire nuove soluzioni per gli esseri umani. A differenza della maggior parte di quelli tradizionali, gli attuali algoritmi di AI operano in modo *non lineare* e, in alcuni casi, *stocastico* (ovvero grazie a processi probabilistici), il che li rende capaci di trovare relazioni complesse e non evidenti nei dati. Analogamente a un algoritmo classico, un algoritmo di apprendimento automatico (ML) è composto da una serie di passaggi precisi. Tuttavia, a differenza degli algoritmi classici, questi passaggi non portano direttamente a un risultato specifico. Essi si distaccano dalla precisione e dalla prevedibilità degli algoritmi tradizionali. Questa flessibilità è ancora più evidente nel deep learning, dove le reti neurali profonde (le incontreremo nel capitolo 4) "imparano" modelli altamente complessi che sarebbero impossibili da definire manualmente attraverso regole rigide. Sono tutte caratteristiche che riflettono il carattere adattivo e l'autonomia nell'apprendere dai dati dei moderni algoritmi di AI.

Se accettiamo questo principio discriminativo, ovvero che il fattore distintivo e peculiare dell'AI, oggi, sia la presenza della componente ML, ovvero la capacità di apprendimento autonomo degli algoritmi, allora possiamo condividere le considerazioni che seguono.

I sistemi di navigazione, ovvero i dispositivi che utilizzano il Global Positioning System (GPS) per determinare la posizione della nostra auto e fornire indicazioni stradali, mappe e altre informazioni utili al conducente, non sono necessariamente AI. Se dobbiamo raggiungere una destinazione e chiediamo al nostro navigatore di indicarci la strada, lo farà risolvendo "dietro le quinte" (ovvero nel cloud) un complesso problema matematico

di ottimizzazione combinatoria. Ovvero, fra tutte le combinazioni possibili di strade, sceglierà quella che “ottimizza” il percorso. Il criterio di ottimizzazione è quello che avremo impostato noi: minimizzazione del tempo di percorrenza, oppure del chilometraggio, giusto per fare due esempi. Ma senza una fase di addestramento che consenta all’algoritmo di apprendere (dalle nostre scelte precedenti, o dalle scelte fatte dagli altri viaggiatori sullo stesso tipo di percorso). In definitiva, anche se usano grandi quantità di dati, questi algoritmi non imparano autonomamente, in quanto soddisfano criteri predefiniti (es. calcolo del percorso più breve o più veloce) basate su algoritmi classici come ad esempio Dijkstra o A* (Hart *et al*. 1968; Dijkstra 1959).[1]

Analogamente, sebbene avanzati, molti sistemi di assistenza alla guida, come ad esempio il *cruise control* adattivo o i sensori di parcheggio, non sono veri sistemi di AI. Si basano su sensori e regole predefinite, senza un reale processo di apprendimento dai dati. Usano algoritmi che calcolano traiettorie, tempi di frenata, angoli minimi compatibili con gli spazi disponibili, basati ad esempio su leggi di moto di tipo deterministico, ma non apprendono in modo autonomo, diversamente dagli algoritmi che sottendono alla guida autonoma. Non è (probabilmente) AI nemmeno il pilota automatico che assiste i piloti di aerei in tutte le fasi del volo, ad eccezione di quelle di decollo e atterraggio.

E ancora: i processi che consentono ai nostri telefoni cellulari di scambiare immagini o film in tempo reale sono basati su algoritmi di compressione e decompressione di immagini e non sono AI. Algoritmi come JPEG o MPEG sono progettati per ridurre la dimensione dei file in base a regole precise, senza apprendimento.

Sono algoritmi di intelligenza artificiale, invece, gli *spam filter* della nostra posta elettronica, perché si addestrano in continuazione sulla base delle nostre scelte in merito ai messaggi che riceviamo e che giudichiamo sgraditi, o inutili (cestinandoli, come faremmo noi). Sono algoritmi di AI quelli che

fanno previsione sulla propensione all'acquisto di un consumatore usando modelli di machine learning per analizzare comportamenti passati. Imparano dalle preferenze e dai comportamenti dei clienti per fare previsioni personalizzate. Anche gli algoritmi di *churn analysis* rientrano nell'AI. Si tratta di modelli predittivi che utilizzano machine learning per identificare i clienti di compagnie telefoniche o distributrici di energia che sono a rischio di abbandono, basandosi su dati storici e aggiornandosi man mano che vengono raccolte nuove informazioni. Altri algoritmi di AI sono quelli di *riconoscimento vocale*, ovvero sistemi come quelli utilizzati da assistenti vocali (es. Siri, Alexa), che apprendono a riconoscere meglio la voce dell'utente. O quelli di *riconoscimento facciale*, dove il sistema migliora la sua accuratezza nel tempo apprendendo da nuove immagini. E infine, i *motori di raccomandazione*, quali ad esempio quelli di Netflix o Amazon, che imparano dai comportamenti degli utenti per offrire suggerimenti all'acquisto sempre più personalizzati.

Che poi AI non sia sinonimo di processo "governato da algoritmi" è alquanto evidente. In particolare, tutti gli algoritmi che derivano da modelli matematici ispirati a leggi fisiche (ne parleremo diffusamente nel capitolo 8) non hanno generalmente alcuna componente di apprendimento automatico, dunque non sono derubricabili come AI. Non sono addestrati dai dati, ma implementano equazioni matematiche che codificano leggi fondamentali le quali a loro volta descrivono processi naturali, biologici, economici e sociali. Esempi notevoli sono i modelli per le previsioni meteorologiche, quelli che si usano per analisi aerodinamiche o strutturali (come la *crash analysis*) nell'industria automobilistica, o ancora, quelli che simulano l'impatto di terremoti in area ad alto rischio sismico, solo per citarne alcuni.

Storie di successo

Visione artificiale, riconoscimento e utilizzo del linguaggio scritto e parlato, intelligenza generativa (di testi, immagini e film), robotica, guida autonoma di veicoli e mezzi di trasporto pubblici, sistemi esperti (quelli che pongono in essere procedure adeguate alla risoluzione di problemi come solo un esperto del settore saprebbe fare): ecco alcuni dei domini in cui l'AI ha riportato notevoli successi (figura 2.1).

In breve facciamo qualche distinguo.

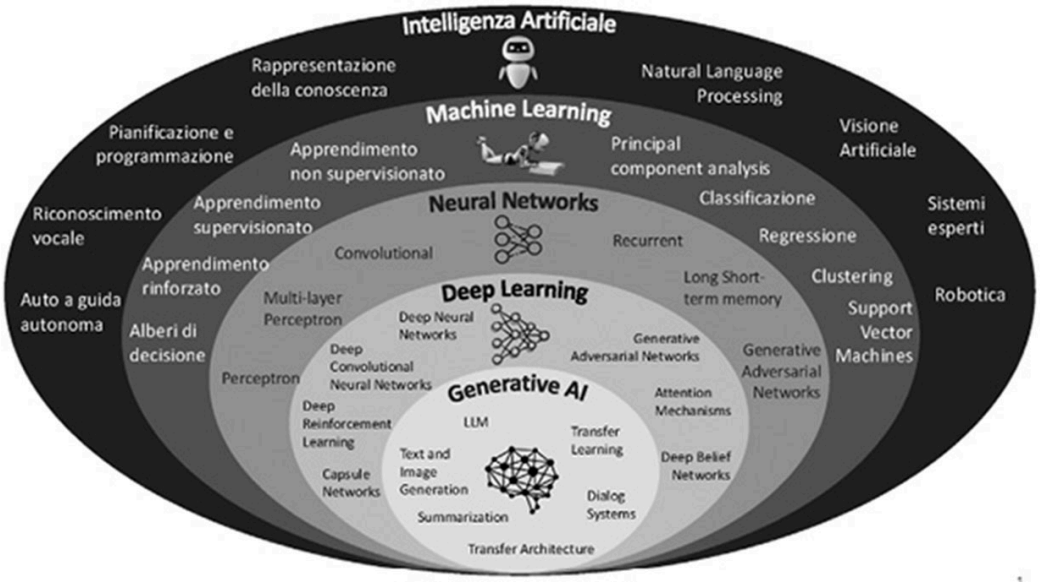

FIGURA 2.1. Intelligenza artificiale, machine learning, reti neurali, deep learning, intelligenza generativa.

- La *visione artificiale* è un ramo della tecnologia che consente alle macchine di vedere, interpretare e reagire al mondo visivo che le circonda. Telecamere e sensori catturano immagini che vengono poi elaborate con algoritmi. Combina aspetti dell'ottica e dell'informatica per creare sistemi capaci di eseguire compiti visivi complessi che una volta richiedevano l'intervento umano. Il processo si articola in diverse fasi. Utilizzando telecamere e sensori, i sistemi di visione artificiale catturano immagini del loro ambiente, semplici fotografie o complesse

rappresentazioni tridimensionali. In seguito, le immagini vengono elaborate e analizzate con algoritmi in grado di evidenziare motivi (*pattern*), riconoscere oggetti, misurare e classificare. La fase finale è l'interpretazione dei dati analizzati e la presa di decisioni basate su queste informazioni. In un'azienda di processo, questo può includere azioni come la classificazione dei prodotti o l'identificazione di difetti. La qualità delle immagini catturate dai sistemi dipende in larga misura dall'ottica e dai sensori utilizzati, che determinano la precisione e l'efficacia del sistema di visione artificiale. Infine, il software utilizzato per elaborare e analizzare le immagini è essenziale per l'efficacia dei sistemi di visione artificiale, garantendo la possibilità di gestire grandi volumi di dati in tempo reale. Si usa talvolta il termine *neural vision* per caratterizzare l'uso dell'AI nella visione artificiale.

- La *robotica* è un campo multidisciplinare che integra ingegneria, informatica e tecnologia per progettare, costruire e operare robot. Questi sono generalmente progettati per replicare o migliorare le attività umane, svolgendo compiti ripetitivi, o pericolosi, o che richiedono alta precisione. Le componenti fondamentali che contribuiscono alla realizzazione dei robot includono la progettazione meccanica, che riguarda la struttura fisica e le capacità di movimento dei robot, spesso ispirate alle forme umane o animali; i sistemi elettrici, che comprendono le fonti di alimentazione e i sistemi di controllo che permettono ai robot di funzionare; infine, il software e gli algoritmi che consentono ai robot di elaborare informazioni, prendere decisioni ed eseguire compiti.

  È difficile descrivere in modo esauriente le diverse tipologie di robot, ma possiamo semplificare dicendo che esistono robot industriali, utilizzati nelle linee di produzione per svolgere compiti ripetitivi con alta precisione, come ad esempio saldatura, verniciatura e

assemblaggio; robot di servizio, progettati per assistere gli esseri umani in vari settori, compresi i robot sanitari per chirurgia o riabilitazione e i robot domestici per la pulizia e la manutenzione; robot autonomi, in grado di eseguire compiti senza intervento umano, come droni e auto a guida autonoma; e robot umanoidi, che imitano l'aspetto e il comportamento umano, spesso utilizzati per ricerca, intrattenimento e compagnia. La robotica è una disciplina in continua evoluzione, guidata dai progressi tecnologici e dalle crescenti applicazioni in vari settori, che promette di rivoluzionare il modo in cui vengono eseguiti i compiti e di espandere le capacità sia degli esseri umani che delle macchine.

- I *sistemi esperti* sono programmi informatici progettati per risolvere problemi complessi e fornire capacità decisionali simili a quelle di un esperto umano. Essi utilizzano sia fatti che considerazioni euristiche, proprio come farebbe un esperto umano. Vengono così chiamati perché contengono la conoscenza esperta di un dominio specifico e possono in teoria risolvere qualsiasi problema complesso di quel particolare dominio. Un sistema esperto è costituito principalmente da un'interfaccia utente, un motore di inferenza e una base di conoscenza. Il primo sistema esperto, realizzato nel 1970, rappresentò uno dei primi approcci di successo dell'intelligenza artificiale. Questi sistemi sono progettati per un dominio specifico, come ad esempio una branca della medicina. La loro *performance* dipende dalla conoscenza dell'esperto immagazzinata nella sua base di conoscenza. Più grande è la conoscenza, più il sistema sarà in grado di migliorare le proprie prestazioni. Uno degli esempi comuni di un sistema esperto è il suggerimento di errori di ortografia mentre si digita nella casella di ricerca di Google. È opportuno ricordare che un sistema esperto non è utilizzato per sostituire gli esperti umani; piuttosto, viene impiegato per assistere l'essere umano nel prendere

decisioni complesse. Questi sistemi non possiedono le capacità di pensiero umano, ma funzionano limitatamente alla conoscenza contenuta nella base di conoscenza del dominio specifico in cui operano.

- Il *riconoscimento vocale* (*speech recognition*), chiamato anche riconoscimento automatico del parlato, riconoscimento vocale del computer o conversione della voce in testo, è una tecnologia che consente a un programma di convertire il parlato umano in testo scritto. Spesso si confonde il riconoscimento vocale con il riconoscimento della voce, ma esiste una differenza sostanziale: il riconoscimento vocale si concentra sulla traduzione del discorso da formato verbale a testuale, mentre il riconoscimento della voce mira "semplicemente" a identificare la voce di un singolo utente. Numerose applicazioni e dispositivi per il riconoscimento vocale sono disponibili sul mercato. Questi sistemi integrano elementi come grammatica, sintassi, struttura e composizione dei segnali audio per comprendere ed elaborare il linguaggio umano. L'ideale è che questi sistemi apprendano continuamente, migliorando le loro risposte a ogni interazione. I sistemi migliori permettono anche alle aziende di personalizzare e adattare la tecnologia in base a specifiche esigenze, che possono riguardare la lingua, le sfumature del parlato e il riconoscimento del marchio. Si possono così perseguire obiettivi significativi, quali ad esempio la ponderazione linguistica, attribuendo un peso maggiore a parole frequentemente usate, oltre ai termini già presenti nel vocabolario di base, l'etichettatura dei parlanti, fornendo trascrizioni che identificano i contributi di ciascun partecipante in conversazioni con più interlocutori; addestramento acustico, adattando il sistema all'ambiente sonoro (come il rumore di fondo in un call center) e agli stili di parlato (come tonalità, volume e velocità della voce).

- L'*elaborazione del linguaggio naturale* (*Natural Language Processing*, *NLP*) è un'area interdisciplinare che si occupa principalmente di fornire ai computer la capacità di elaborare dati codificati in linguaggio naturale ed è quindi strettamente legata al recupero delle informazioni, alla rappresentazione della conoscenza e alla linguistica computazionale, un ramo della linguistica. Tipicamente, i dati vengono raccolti in corpora di testi, utilizzando approcci basati su regole, di tipo statistico, oppure basati su reti neurali artificiali per l'apprendimento automatico. L'NLP ha fornito l'humus naturale da cui è germinata l'AI generativa; alimenta motori di ricerca, supportando chatbot per il servizio clienti con comandi vocali, sistemi GPS operati con la voce e assistenti digitali sugli smartphone. L'NLP ha anche un ruolo crescente nelle soluzioni che aiutano ad automatizzare le operazioni aziendali, aumentare la produttività dei dipendenti e semplificare i processi aziendali più critici. Fino agli anni '80, la maggior parte dei sistemi di elaborazione del linguaggio naturale si basava su complessi insiemi di regole scritte manualmente. Tuttavia, dalla fine degli anni '80, si è verificata una rivoluzione in questo campo grazie all'introduzione di algoritmi di apprendimento automatico per l'elaborazione del linguaggio. Un cambiamento reso possibile sia dall'aumento costante della potenza di calcolo (vedi la già citata legge di Moore) sia dal graduale declino delle teorie linguistiche chomskiane (Chomsky 1986), come la grammatica trasformazionale, i cui principi teorici scoraggiavano l'uso della linguistica dei corpora, essenziale per l'approccio dell'apprendimento automatico nell'elaborazione del linguaggio.
  Un ambito di particolare successo dell'NLP è quello della *traduzione automatica* da una lingua a un'altra. Questo processo richiede una comprensione approfondita delle strutture linguistiche, dei significati contestuali e delle sfumature culturali, rendendola una sfida complessa e

affascinante nell'ambito dell'NLP. Invece di limitarsi all'addestramento su testi tradotti con attenzione, vengono proposti articoli e altri testi in varie lingue che affrontano un argomento comune, senza preoccuparsi delle traduzioni dettagliate. Questo metodo, che prevede l'addestramento su corpora di testo approssimativamente corrispondenti ma non tradotti, è conosciuto come tecnica dei corpora paralleli ed è paragonabile al passare da un corso di lingua di base a un programma di immersione totale. Sebbene l'addestramento sia meno preciso, la quantità di dati disponibili è notevolmente maggiore: gli sviluppatori possono includere nuovi articoli, recensioni di libri e film e praticamente qualsiasi altra pubblicazione, formale o informale, su argomenti trattati da autori in diverse lingue. Il successo di questo approccio ha portato a un uso più diffuso dell'apprendimento parzialmente supervisionato, in cui si utilizzano informazioni altamente approssimative o parziali per l'addestramento. Il miglioramento che si è ottenuto grazie all'uso dei corpora paralleli rispetto al tradizionale approccio di traduzione automatica ha consentito a Google Translate di migliorare di oltre il 50% la qualità delle sue performance.

- Ovviamente, la capacità di tradurre testi e classificare immagini è una cosa, mentre la capacità di *generare* e creare nuovi testi, immagini e suoni è un'altra. Le *reti neurali generative* sono in grado di produrre contenuti. Mentre una rete neurale standard può riconoscere un'immagine di un volto umano, una rete generativa può realizzare un'immagine di un volto umano che appare reale. Una tecnica di addestramento comune per la creazione di intelligenza artificiale generativa prevede l'uso di due reti, chiamate reti generative avversarie (*Generative Adversarial Networks*, *GAN*), che hanno obiettivi di apprendimento complementari. L'obiettivo della rete generatrice è quello di creare potenziali output, mentre la rete discriminatrice ha il compito di impedire la

generazione di output di bassa qualità. Durante la fase di addestramento, il generatore e il discriminatore vengono addestrati in alternanza, mantenendo uno dei due fisso per addestrare l'altro e viceversa. Addestrate su enormi quantità di dati provenienti principalmente da Internet, le reti neurali generative possono anche convertire il testo in immagini e viceversa, espandere e condensare descrizioni, e svolgere compiti simili. Spesso, i risultati ottenuti sembrano molto intelligenti, ma altre volte possono apparire sciocchi o incomprensibili. Poiché hanno la capacità di influenzare diversi settori, inclusi quelli creativi, suscitano un grande interesse via via che ricercatori e sviluppatori ne indagano i punti di forza, le limitazioni e le possibili applicazioni.

[1] Va tuttavia osservato che, dopo aver proposto la mappa da seguire, alcuni navigatori stradali (come Google Maps e Waze) ricevono dinamicamente informazioni dai cellulari e grazie a queste possono aggiornare la mappa grazie ad algoritmi di *machine learning*.

3
MACHINE LEARNING, OVVERO QUANDO I COMPUTER IMPARANO (DA SOLI!)

Le radici dell'AI si fondano su un intreccio di idee e sforzi visionari, ispirati dalle prime riflessioni sulla possibilità di dotare le macchine di intelligenza. Da queste origini, l'AI ha attraversato un percorso di sviluppo accelerato, ridefinendo costantemente i confini della nostra interazione con la tecnologia intelligente. Tuttavia, nonostante i suoi settant'anni di storia, gli spettacolari progressi che abbiamo sperimentato nell'ultimo quarto di secolo sono dovuti anche a due altri imprescindibili fattori: la straordinaria potenza dei computer, in particolare quelli basati su GPU (*Graphics Processing Units*) che sono particolarmente efficienti nell'implementazione degli algoritmi dell'apprendimento automatico dei computer, il machine learning, e la straordinaria disponibilità di dati, i Big Data, che in molti ambiti sono necessari per addestrare gli algoritmi del ML (figura 3.1).

Quando si parla di machine learning, ovvero di apprendimento automatico da parte dei computer, la percezione più comune può così riassumersi: si forniscono ai computer tanti dati di addestramento affinché *imparino* a fare cose senza che noi umani dobbiamo specificare esattamente come. Ovvero,

senza doverli programmare allo scopo. Come se fossero cagnolini. Ma succede veramente così?

L'apprendimento automatico è un ramo dell'intelligenza artificiale che esplora come simulare computazionalmente (o addirittura superare) l'intelligenza umana. Esso guida la maggior parte dei progressi attuali del campo concentrandosi su un unico obiettivo: utilizzare algoritmi per migliorare automaticamente le prestazioni di altri algoritmi.

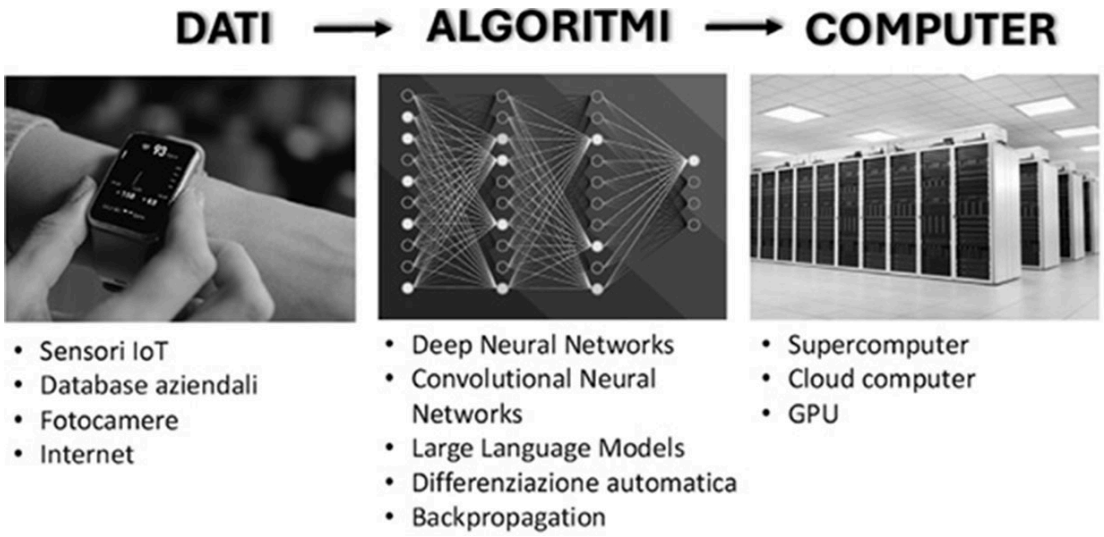

FIGURA 3.1. I tre fattori determinanti dietro la recente esplosione dell'AI.

Ecco come funziona in pratica, per un tipo comune di apprendimento automatico chiamato apprendimento supervisionato. Il processo inizia con un compito, ad esempio "riconoscere quali foto, all'interno di un certo insieme di foto, rappresentino un gatto". L'obiettivo è trovare una funzione matematica, che chiameremo *modello*, che possa svolgere questo compito. Essa prende numeri in input – in questo caso, fotografie digitalizzate – e li trasforma in altri numeri in output, che potrebbero rappresentare etichette come "gatto" o "non gatto", diciamo 1 o 0, per semplificare. Il modello ha una struttura matematica di base caratterizzata da un certo insieme

di numeri, detti parametri, ma inizialmente è improbabile che produca risultati accurati.

Come si addestra un modello di AI

Ora è il momento di *addestrare* il modello, ed è qui che entra in gioco un altro tipo di algoritmo. Prima, una diversa funzione matematica, chiamata *funzione obiettivo*, o anche *funzione costo*, o ancora *loss function*, indicata con il simbolo *J*, calcola un numero che rappresenta la "distanza" attuale tra l'output del modello e il risultato desiderato, in corrispondenza a tutte le foto dell'insieme di partenza che chiameremo insieme di addestramento, o *training set*. Nell'esempio specifico, questo numero indica la percentuale di volte in cui il modello fallisce nell'identificare correttamente le foto che rappresentano gatti. Poi, l'algoritmo di addestramento utilizza la misura di distanza della funzione obiettivo per aggiustare la forma del modello originale. Non deve "sapere" nulla su cosa rappresenta il modello, ovvero i parametri che lo definiscono. Semplicemente, aggiorna il valore dei parametri in modo tale che la funzione obiettivo (e dunque l'errore percentuale del modello) risulti la più piccola possibile.

Una volta effettuati questi aggiustamenti, il processo ricomincia, dando luogo a iterazioni successive. Il modello aggiornato trasforma gli input dell'insieme di addestramento in output (leggermente migliori), poi la funzione obiettivo indica un altro (leggermente migliore) aggiustamento al modello. Dopo un numero sufficiente di iterazioni, il modello addestrato dovrebbe essere in grado di produrre output accurati per la maggior parte dei suoi esempi di addestramento. Ecco il vero obiettivo: mantenere quelle prestazioni su nuovi esempi del compito, purché non siano troppo diversi da quelli inclusi nell'insieme di addestramento.

La messa in moto di questo processo automatizzato fa emergere una rappresentazione matematica del compito (il modello), senza che gli esseri umani debbano specificare quali dettagli sono importanti. Con un algoritmo efficiente, una funzione obiettivo ben scelta e un insieme di addestramento sufficientemente ricco, l'apprendimento automatico può creare potenti modelli computazionali che svolgono compiti in modo molto soddisfacente. Una descrizione matematica dell'intero processo verrà abbozzata (senza esagerare con i tecnicismi!) nel prossimo capitolo.

I compiti di classificazione e previsione – come identificare i gatti nelle foto o lo spam nelle email – di solito si basano sull'apprendimento *supervisionato*, il che significa che i dati di addestramento sono già etichettati in anticipo e l'etichetta è funzione dell'output desiderato: le foto contenenti gatti, per esempio, sono etichettate "gatto" o, equivalentemente, con il numero 1; tutte le altre con lo zero. Il processo di addestramento troverà quella funzione matematica (il modello) che può mappare quanto più input possibile sul suo output (conosciuto) corrispondente. In seguito, il modello addestrato sarà in grado di etichettare esempi sconosciuti, senza commettere troppi errori. Ad esempio, all'inizio del 2020, ricercatori del MIT, il Massachusetts Institute of Technology, hanno utilizzato l'apprendimento supervisionato per scoprire un nuovo antibiotico, la halicina. A tal fine, è stato impiegato un database contenente 2000 molecole per addestrare un modello, nel quale la struttura molecolare veniva fornita come input e l'efficacia dell'antibiotico come output (Marchant 2020).

L'apprendimento *non supervisionato*, invece, trova strutture all'interno di esempi non etichettati, raggruppandoli in gruppi non specificati in anticipo. In questo caso, i dati di input non hanno alcuna etichetta. Viene spesso utilizzato per insiemi di dati grezzi e ha il compito di convertirli in dati strutturati. In molti contesti, sono generate grandi quantità di dati, inclusi i file di registro generati dai computer. Per questo motivo, l'apprendimento non supervisionato rappresenta una delle sfide

più grandi del machine learning. I sistemi di raccomandazione di contenuti che apprendono dal comportamento passato di un consumatore, così come alcuni compiti di riconoscimento di oggetti nella visione artificiale, possono basarsi sull'apprendimento non supervisionato. Ad esempio, servizi di streaming video come Netflix impiegano algoritmi per individuare gruppi di utenti con abitudini di visione simili, in modo da suggerire loro contenuti aggiuntivi da guardare. Alcuni compiti, come la modellazione del linguaggio eseguita da sistemi LLM (*Large Language Models*) come GPT-4, utilizzano combinazioni intelligenti di tecniche supervisionate e non supervisionate, note come apprendimento auto-supervisionato e semi-supervisionato.

L'apprendimento *semi-supervisionato* è quello in cui l'insieme di addestramento contiene sia elementi etichettati che elementi non etichettati. Si pensi ad esempio a un reparto di neurochirurgia in cui si acquisiscono tantissime immagini tomografiche, ma (per mancanza di tempo o incapacità) solo per un numero ristretto il neurochirurgo è riuscito a fornire una descrizione della patologia osservata (ovvero, etichettarle), precisando se si tratti o meno di un tumore e, in caso affermativo, di quale tipo sia.

Infine, l'apprendimento *per rinforzo* genera una funzione-modello utilizzando un segnale di ricompensa anziché esempi di risultati desiderati. Esso mira a insegnare agli agenti, entità in grado di prendere decisioni e intraprendere azioni, ad adottare una sequenza di decisioni in un ambiente per ottimizzare le ricompense cumulative totali. Massimizzando questa ricompensa attraverso tentativi (migliaia, decine di migliaia o oltre, realizzati in tempi rapidissimi grazie a computer potenti) ed errori, un modello può migliorare le sue prestazioni su compiti dinamici e sequenziali come giocare a dama o a scacchi o a Go, o controllare il comportamento di agenti reali e virtuali (come le auto a guida autonoma o i chatbot). L'obiettivo principale dell'apprendimento per rinforzo è permettere a un agente di interagire con l'ambiente, osservare le conseguenze delle proprie

azioni e adattare il proprio comportamento in base a tali osservazioni. L'apprendimento per rinforzo richiede il coinvolgimento umano nella creazione dell'ambiente di addestramento dell'AI: gli esseri umani definiscono un simulatore e una funzione di ricompensa, e l'AI si allena sulla base di questi elementi. Per ottenere risultati significativi è fondamentale scegliere accuratamente il simulatore e la funzione di ricompensa. Nel 2017 AlphaZero, il programma sviluppato da Google DeepMind, ha sconfitto Stockfish, il più potente programma di scacchi esistente fino ad allora. Questo evento ha segnato un momento significativo nella storia degli scacchi e dell'intelligenza artificiale, poiché AlphaZero ha dimostrato di poter apprendere e sviluppare strategie di gioco senza alcuna supervisione umana. Per farlo, si è realizzato un programma di apprendimento con rinforzo che ha giocato contro se stesso. Per valutare le sue prestazioni, ha utilizzato una funzione di ricompensa che valutava le sue mosse in base alle opportunità che queste creavano.

I sistemi matematici basati sul machine learning sono spesso estremamente potenti e accurati, tuttavia non sono scevri da debolezze di carattere strutturale. La più ricorrente è il cosiddetto *overfitting*: un modello "overfitted" costruisce una funzione-modello che si adatta così bene ai valori dei dati usati per l'addestramento da non poter generalizzare in modo affidabile per rappresentare nuovi dati non compresi nell'insieme di addestramento. Per fare un esempio, si pensi a un algoritmo di machine learning che deve riconoscere se una certa foto (data in input) rappresenti o meno un gatto, che fallisce quando la foto rappresenta effettivamente un gatto ma è capovolta rispetto a quelle con cui era stato addestrato. Oppure quando nei dati di addestramento c'è troppa inaccuratezza, o "rumore", tali da indurre la funzione modello ad assumere una struttura troppo ricca per poterle consentire di cogliere questa eccessiva variabilità nei dati. In questo caso, la stessa funzione modello non sarà poi capace di "generalizzare", ovvero comprendere il vero significato di dati che non siano stati

ricompresi nell'insieme di addestramento. Se volessimo ad esempio addestrare un modello a prevedere la temperatura giornaliera media nell'arco dei tre mesi primaverili in una località specifica e avessimo usato come insieme di addestramento dati oscillanti perché affetti da rumore nel processo di acquisizione, la funzione modello potrebbe ad esempio non essere in grado di assumere un comportamento monotòno crescente, come in realtà ci si aspetterebbe da marzo a giugno.

I bias nei dati sono in genere da attribuire a campionamenti non significativi. Essi possono derivare da distorsioni o pregiudizi presenti in un dataset, dovuti a errori sistematici nel modo in cui sono raccolti, selezionati, rappresentati o interpretati. Se ad esempio tutti si riferissero a controindicazioni di un certo farmaco testato solo su popolazione maschile, ci potremmo aspettare una cattiva capacità del modello a generalizzare alla popolazione femminile. In altri termini, i bias presenti nei dati di addestramento potrebbero essere amplificati dal processo di addestramento, portando a risultati distorti – o addirittura errati. E anche quando un modello funziona, non è sempre chiaro il perché. Gli algoritmi di deep learning sono particolarmente afflitti da questo problema di *interpretabilità*.

Ripartiamo dalle definizioni

Proviamo a rimettere le definizioni e i concetti nel loro giusto ordine. Forse la definizione più antica di machine learning è stata fornita in una citazione apocrifa di Arthur Samuel, spesso attribuita al suo articolo del 1959 sul machine learning per il gioco della dama:

"Il machine learning è il campo di studio che conferisce ai computer la capacità di apprendere senza essere esplicitamente programmati per farlo."

Una definizione più concisa è stata proposta in un più recente rapporto della Royal Society del 2017:

"*Il machine learning è un insieme di regole che permette ai sistemi di apprendere direttamente dagli esempi, dai dati e dall'esperienza*."

In un articolo di revisione di Michael Jordan e Tom Mitchell (2015), si definisce "apprendimento" il *processo di trasformazione delle informazioni in competenza o conoscenza*; "apprendimento automatico" (machine learning, appunto) l'*apprendimento automatizzato*.

Il processo di machine learning assume dunque la forma di un algoritmo che prende informazioni come input e produce conoscenza come output. Come precedentemente osservato, gli algoritmi di machine learning sono generalmente suddivisi in quattro principali categorie: apprendimento supervisionato, semi-supervisionato, non supervisionato e per rinforzo.

Gli algoritmi di apprendimento automatico (ML) sono in generale efficaci quando sono disponibili grandi volumi di dati e quando gli sviluppatori conoscono sufficientemente bene la verità di base (ovvero l'output corretto) per verificare i risultati, spesso anche durante l'addestramento dei modelli. Non sono invece particolarmente efficaci quando i dataset sono scarsi, parziali e non etichettati (di solito quando l'acquisizione dei dati è costosa), quando la verità di base è sconosciuta e non sono disponibili dataset di riferimento, e per decisioni ad alto impatto che richiedono modelli o output interpretabili dagli esseri umani.

Nel caso degli algoritmi supervisionati di machine learning, i loro principali compiti possono essere individuati fra i seguenti:

- *Regressione*: consiste nell'estrarre da una distribuzione discreta di dati (input e output) una funzione che ne descriva il comportamento. È la cosiddetta modellazione matematica basata sui dati. Ad esempio, ricostruire una funzione che descriva al variare del tempo il valore di una certa azione allo Stock Exchange. In questo caso il set di

addestramento è composto da *N* coppie formate da un dato di input e un corrispondente valore di verità di base, ovvero *N* date diverse e le corrispondenti quotazioni della specifica azione. Un altro esempio potrebbe essere la rappresentazione attraverso una curva del comportamento della temperatura in una certa località partendo da una distribuzione giornaliera della temperatura stessa.

- *Classificazione*: consiste nel prevedere un valore da un insieme finito, ad esempio l'etichetta da associare a un'immagine. Il set di addestramento è composto da coppie di un dato di input e un valore di verità di base, in questo caso un'etichetta pescata da un insieme finito. L'approccio standard prevede di assegnare un punteggio a ciascuna classe potenziale, con la classe corretta che ha il punteggio massimo. La si applica ad esempio al problema della classificazione delle immagini (gatto o non gatto), al filtro antispam (mail da cestinare o tenere), alla cosiddetta *sentiment analysis* (finalizzata a determinare se il contenuto di un testo esprima un sentimento positivo, negativo o neutro) ecc.

Nel caso dell'addestramento non supervisionato possiamo invece evidenziare i seguenti compiti.

- *Clustering*: si tratta di raggruppare i dati in modo che quelli simili tra loro si trovino nello stesso gruppo, anche senza etichette predefinite. Ad esempio, per una certa distribuzione di punti nel piano cartesiano si vogliono individuare sottoinsiemi che condividano specifiche somiglianze. Se fossero individui, raggruppandoli sulla base di preassegnate fasce di reddito. Serve per il rilevamento di comunità in biologia, nelle scienze sociali, in economia, nella segmentazione di clienti per un determinato tipo di acquisto ecc.

- *Associazioni*: si riferisce all'identificazione di relazioni tra variabili nei dati, come nelle analisi di mercato, dove si cercano correlazioni. Questo approccio è utilizzato per scoprire la probabilità della co-occorrenza di elementi in una raccolta di dati. Un esempio classico è rappresentato dai sistemi di raccomandazione che analizzano quali prodotti sono spesso acquistati insieme, come nel caso delle analisi di mercato.
- *Stima della densità* (ad esempio, rilevamento di anomalie, generazione di scenari realistici): questo compito si concentra sul trovare i pattern comuni all'interno di una popolazione di dati senza presupporre una struttura rigida. È utile per rilevare anomalie (dati che non si conformano ai pattern attesi) o per generare scenari realistici, simili a quelli descritti dai dati di addestramento.
- *Riduzione della dimensionalità*: si tratta di trovare una rappresentazione compatta di dati complessi, riducendone la dimensionalità pur mantenendo il più possibile le informazioni più rilevanti. È molto utile per affrontare problemi di dati ad alta dimensionalità, rendendone più semplice la visualizzazione e la gestione.

4
MATEMATICA... IL MINIMO SINDACALE

Prima di avventurarci in quello che per molti lettori risulterà il passaggio più ostico di questo libro, penso vada messo in chiaro un punto cruciale. Ormai dovrebbe essere evidente, dopo tante premesse, che gli algoritmi del machine learning sono una gioiosa macchina da guerra (*sic*) costruita su quanti più dati possibili (preferibilmente, Big Data). I dati, come il termine stesso indica, sono "dati" (participio passato del verbo dare), ovvero quantità *note*! Ed è vietato confonderli con le *soluzioni*! Queste ultime sono incognite, dunque vanno trovate. Ogni volta che si affronta un problema matematico, la sfida è quella di passare dai dati alle soluzioni, ovvero dall'input all'output. *Come* passarvi, è quello di cui stiamo trattando in questo libro. Possiamo intendere il machine learning come un procedimento che ci consente di associare a un certo insieme di dati la soluzione corrispondente. Si tratti di attribuire un'etichetta che consente di stabilire se una certa foto rappresenti un cane o altro, oppure di riconoscere se una certa immagine medica (una TAC, o una risonanza magnetica, ad esempio) nasconda o meno la presenza di un tumore, e in caso affermativo di quale tumore si tratti, o di partizionare un certo insieme di individui in sottoinsiemi con caratteristiche simili (ad esempio la loro capacità di spesa, o l'attitudine a sviluppare determinate patologie), o ancora di formulare previsioni meteo per il prossimo weekend a partire dai dati che caratterizzano venti, pressione atmosferica e umidità al momento attuale. Come si può intuire da questi esempi, il *tipo* di

dati può essere diversissimo, e diversissime saranno le soluzioni associabili. Quello che rimane invariato è l'assetto matematico: il trinomio

dati → procedimento matematico → soluzioni.

Il procedimento matematico, a sua volta, può assumere forme molto differenti. Nel caso della partizione in sottoinsiemi di cui sopra si possono usare metodi statistici di tipo *clustering*, mentre in quello della meteorologia bisognerà ricorrere a un complesso sistema di equazioni che traducono matematicamente i processi fisici della nostra atmosfera.

I mattoncini della costruzione ML

Anche il machine learning può essere considerato un esempio (particolarmente notevole) di procedimento matematico. Per comprendere come si estrinseca il processo di apprendimento automatico, possiamo ricorrere alla formulazione generale proposta da Tom M. Mitchell nel lontano 1997:

"*Un programma apprende da un'esperienza E rispetto a uno specifico obiettivo T (task) e una assegnata misura di performance P se la sua performance rispetto a T migliora con l'esperienza E*."

La formalizzazione del processo inizia a prendere corpo, anche se in forma piuttosto astratta. Per dare concretezza a questo enunciato riconsideriamo l'ormai abusato problema del riconoscimento di immagini: si vuole scrivere un programma il quale, per un'assegnata immagine chiamata *x*, dovrà restituire una cifra binaria *y* che varrà 1 se l'immagine di partenza è un gatto, 0 in tutti gli altri casi. Possiamo immaginare che un algoritmo di ML funzioni all'incirca in questo modo:

- (E = esperienza): si parte da un insieme di addestramento (*training set*) che consiste in un insieme di N immagini x_i (rappresentanti gatti o altri animali) e un corrispondente valore y_i, per i compreso fra 1 e N.
- (M = modello): si introduce un possibile *modello*, ovvero una funzione che associa a una generica immagine x un valore binario y (0 oppure 1). In simboli, $y = f(x, p)$ dove x e p sono valori da determinare (genericamente denominati *parametri*) e f è la funzione che associa alla coppia (x, p) il risultato y. Il modo in cui f dipende da x e p identifica il modello scelto.
- (P = performance): addestriamo il modello (*training*) scegliendo fra tutti i possibili parametri quelli in corrispondenza dei quali il modello, applicato alle immagini del training set, commette il più piccolo errore possibile. In altri termini, i valori $t_i = f(x_i, p)$ si discostano il meno possibile da quelli veri (del training set) y_i. Una possibilità per calcolare questo errore è sommare il quadrato delle discrepanze di t_i da y_i, ovvero calcolando

  $$p_{opt} \text{ tale che } J(p_{opt}) = min_{p \in \mathbb{R}^n} J(p)$$

  dove

  $$J(p) = \sum_i [y_i - f(x_i; p)]^2$$

  è la funzione obiettivo (o funzione costo). Naturalmente vi sono altre possibili definizioni di errore, consistenti in una scelta diversa della funzione obiettivo J. La scelta di tale funzione caratterizza la misura di performance P del nostro modello.

Per chi detesta il formalismo matematico, al di là del tecnicismo potremmo dire che nel caso in esame con il machine learning si propone un modello di relazione fra una qualunque immagine x e il suo valore semantico (y, che indica se l'immagine sia o meno quella di un gatto) dipendente da parametri che sono scelti in

modo tale da ottimizzare la performance del modello quando applicato alle immagini usate per l'addestramento.

A questo punto, determinati i valori numerici dei parametri, e quindi determinata completamente la funzione *f*, il nostro modello sarà pronto per essere testato su una qualsiasi *nuova immagine* che non sia già presente fra quelle del training set.

Naturalmente, cambiando il contesto applicativo, potremmo partire ora da un qualunque insieme di input *x* (ad esempio soggetti affetti da Covid-19) per generare un output *y* (per esempio il numero stimato di giorni necessari per la loro guarigione).

Vi sono diversi criteri per realizzare la fase di esperienza (E), che come si è già descritto vanno di solito sotto il nome di apprendimento non supervisionato, supervisionato, semi-supevisionato o rinforzato. In quello non supervisionato solo i valori x_i – non i valori corrispondenti y_i – sono inclusi nel training set (l'addestramento è condotto "alla cieca"). In quello supervisionato lo sono entrambi. Nell'apprendimento rinforzato il training set non è fissato a priori, ma cambia e si aggiorna dinamicamente, in funzione dei risultati e del contesto.

Con un linguaggio meno matematico ma più immaginifico, potremmo dire che, in generale, nell'apprendimento supervisionato si forniscono nella fase di addestramento nozioni specifiche che permettono di costruire un vero e proprio database di informazioni e di esperienze. In questo modo, quando il computer affronta un problema nuovo, non dovrà fare altro che attingere alle esperienze inserite nel proprio sistema, analizzarle e decidere quale risposta dare sulla base di esperienze già codificate.

L'apprendimento non supervisionato prevede invece che le informazioni inserite non siano codificate, ovvero che il computer abbia la possibilità di attingere a determinate informazioni senza avere alcun esempio del loro utilizzo e, quindi, senza avere conoscenza dei risultati attesi a seconda della scelta effettuata. Sarà il computer stesso che dovrà

catalogare tutte le informazioni in suo possesso, organizzarle e apprendere il risultato a cui esse portano.

L'apprendimento per rinforzo è quello di maggiore complessità. Si suppone che il computer sia dotato di sistemi e strumenti in grado di migliorare il proprio apprendimento e, soprattutto, di comprendere le caratteristiche dell'ambiente circostante. Questo apprendimento è tipico, ad esempio, delle auto a guida autonoma in cui si beneficia di un complesso sistema di telecamere e sensori di supporto.

Come si può intuire, la scelta del buon modello *f* è cruciale per la realizzazione di un efficace algoritmo di apprendimento automatico. Naturalmente, numerose opzioni sono disponibili al riguardo. I modelli più semplici prevedono una dipendenza lineare (ovvero, proporzionale) di *f* sia da *x* che da *p*. Altri una dipendenza non lineare (polinomiale) da *x*, con coefficienti polinomiali rappresentati dai parametri da ottimizzare.

Vi sono poi quelli di tipo *random forest* (letteralmente, foresta casuale) in cui la scelta della risposta *y* è basata su una successione di scelte dicotomiche. Questo modello è composto da più alberi decisionali, ognuno dei quali deve fornire la risposta a una specifica domanda. Per fare un esempio: "Dovrei acquistare un appartamento" Da lì, si possono porre una serie di domande per determinare una risposta, come ad esempio: "Quante stanze dovrebbe avere?" oppure "Qual è il budget disponibile?" E ancora: "In quale zona della città dovrebbe essere?" Queste domande costituiscono i nodi decisionali dell'albero, che servono a partizionare i dati. Ogni domanda aiuta l'individuo a prendere una decisione finale, che viene indicata dal nodo foglia. Le osservazioni che soddisfano i criteri seguiranno il percorso "sì", mentre quelle che non lo soddisfano seguiranno il percorso alternativo. Gli alberi decisionali cercano di trovare la migliore suddivisione dei dati e sono in genere addestrati attraverso uno specifico algoritmo, detto *Classification and Regression Tree* (CART). Ogni albero decisionale è un esempio di problema di classificazione, in cui le etichette delle classi sono "acquista" e "non acquista". In generale, quanti più alberi

decisionali formano un insieme nell'algoritmo foresta casuale, tanto più precisi saranno i risultati, in particolare se i singoli alberi non sono correlati tra loro.

Proseguiamo ricordando le *support vector machines*, ovvero macchine a supporto vettoriale, che rientrano nel caso dell'apprendimento supervisionato. Partendo da un training set i cui elementi sono etichettati con la classe di appartenenza (gatto/non gatto, nell'esempio iniziale), viene costruito un modello che assegna i nuovi esempi a una delle due classi, ottenendo un classificatore binario deterministico. Un modello a supporto vettoriale rappresenta gli esempi come punti nello spazio, mappati in modo tale che gli esempi appartenenti alle due diverse categorie siano chiaramente separati da uno spazio il più ampio possibile. I nuovi esempi sono quindi mappati nello stesso spazio e la predizione della categoria alla quale appartengono viene fatta sulla base del semispazio nel quale ricadono.

Tuttavia, i modelli che di gran lunga hanno assunto un ruolo primario sono quelli basati su reti neurali artificiali.

Le reti neurali artificiali

Le *Artificial Neural Network*, *ANN*, ovvero reti neurali artificiali, sono algoritmi matematici che cercano di emulare il comportamento delle reti neurali biologiche, composte da neuroni, assoni e dendriti. Un neurone biologico è una cellula specializzata del sistema nervoso, responsabile della trasmissione di informazioni attraverso segnali elettrici e chimici. Possiamo immaginarla come composta da diverse parti (figura 4.1).

Il soma (o corpo cellulare) contiene il nucleo e altre strutture cellulari essenziali. È responsabile del mantenimento della cellula e dell'integrazione dei segnali ricevuti dai dendriti, ramificazioni che ricevono segnali da altri neuroni. L'assone è una lunga

estensione che trasmette segnali dal soma ad altri neuroni, muscoli o ghiandole. Finisce con le terminazioni assoniche. La guaina mielinica è il rivestimento che avvolge l'assone (in molti neuroni) e aumenta la velocità di trasmissione del segnale. È costituita da cellule gliali come gli oligodendrociti nel sistema nervoso centrale e le cellule di Schwann nel sistema nervoso periferico. Le terminazioni assoniche (o sinapsi) indicano la fine dell'assone, dove vengono rilasciati i neurotrasmettitori per comunicare con altri neuroni.

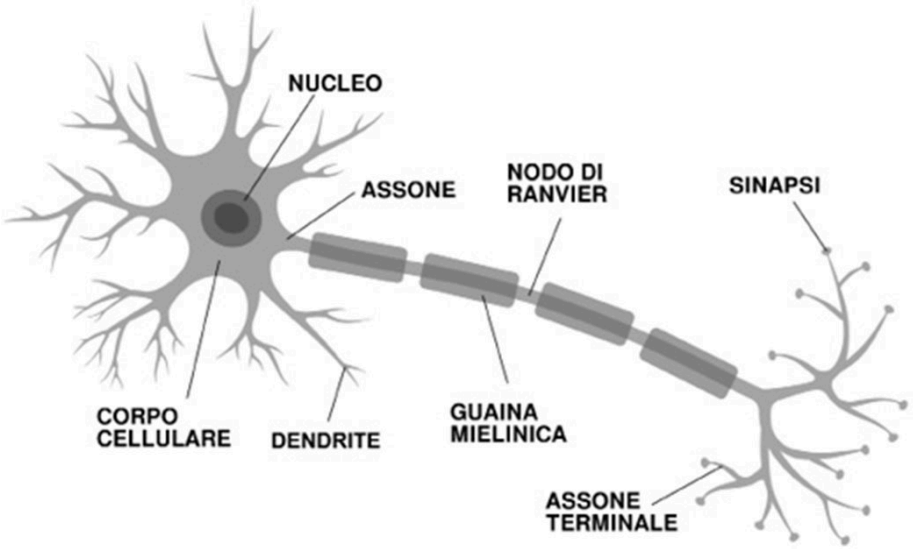

FIGURA 4.1. Un neurone biologico e le sue componenti.

A riposo, un neurone mantiene una carica elettrica diversa tra l'interno e l'esterno della sua membrana, con l'interno più negativo (circa –70 mV). Questo equilibrio è regolato da speciali "pompe" che controllano il passaggio di ioni sodio (Na^+) e potassio (K^+).

Quando il neurone riceve un segnale abbastanza forte, si verifica un rapido cambiamento: i canali del sodio si aprono e gli ioni Na^+ entrano nella cellula, rendendo l'interno più positivo. Subito dopo, i canali del sodio si chiudono e quelli del potassio si

aprono, permettendo agli ioni K^+ di uscire, riportando la carica al suo stato iniziale. Questo processo genera un impulso elettrico, chiamato potenziale d'azione, che si propaga lungo il neurone come un'onda.

Nei neuroni rivestiti di mielina, il segnale si muove saltando tra piccoli spazi scoperti chiamati nodi di Ranvier, rendendo la trasmissione molto più veloce. Alla fine dell'assone, questo impulso elettrico stimola il rilascio di neurotrasmettitori, molecole che attraversano lo spazio tra due neuroni e si legano ai recettori della cellula successiva. Questo può attivare un nuovo segnale o innescare altre risposte, come la contrazione di un muscolo o il rilascio di ormoni.

Il neurone è una parte essenziale della comunicazione all'interno del sistema nervoso. Insieme, i neuroni formano reti complesse che permettono il funzionamento di tutte le attività cerebrali e corporee, dalle percezioni sensoriali alle risposte motorie, ai processi cognitivi complessi come il pensiero e la memoria.

Le reti neurali artificiali sono modelli non lineari (composizionali, ovvero ottenute applicando in sequenza funzioni non lineari) il cui nucleo elementare è il percettrone (*perceptron*), il neurone artificiale inventato nel 1943 da Warren McCulloch e Walter Pitts, che simula il comportamento logico di un neurone biologico (McCulloch e Pitts 1943). Tuttavia, il termine percettrone fu introdotto solo nel 1958 da Frank Rosenblatt, ricercatore presso il Cornell National Laboratory (Rosenblatt 1958). Il modello di McCulloch e Pitts può essere visto come un caso particolare di percettrone. L'obiettivo era quello di sviluppare un metodo per codificare l'informazione simile a quello utilizzato dal cervello umano, che connette circa cento miliardi di neuroni con quadrilioni (10^{21}!) di sinapsi. Una rete neurale artificiale codifica le relazioni tra nodi (simili ai neuroni) e pesi numerici che rappresentano l'intensità delle connessioni (tramite le sinapsi) tra gli stessi nodi. Per decenni, la mancanza di potenza di calcolo e di algoritmi avanzati ha

rallentato lo sviluppo delle reti neurali, che sono rimaste perlopiù rudimentali. Tuttavia, recenti progressi hanno finalmente liberato gli sviluppatori di intelligenza artificiale da queste restrizioni.

I neuroni artificiali, che rappresentano le componenti essenziali di una rete neurale artificiale, sono costituiti da diverse componenti chiave (figura 4.2). Analogamente a quanto avviene per un neurone biologico grazie ai dendriti, un neurone artificiale riceve segnali in ingresso da altri neuroni artificiali. Ogni ingresso è associato a un peso (*w*) che rappresenta l'importanza relativa di quell'ingresso. I pesi sono valori numerici che modulano la forza di ciascun ingresso e sono generalmente modificati durante il processo di apprendimento. A questo punto, una funzione calcola la somma pesata degli ingressi, grazie a un'espressione della forma

$$z = \sum_i w_i x_i + b$$

dove x_i sono gli ingressi, w_i i pesi associati e b è un termine di bias che permette di spostare la funzione di attivazione. Quest'ultima applica una trasformazione non lineare alla somma pesata per determinare l'output del neurone. Si veda per una rappresentazione la figura 4.2. Le funzioni di attivazione comuni includono la sigmoide, la *Rectified Linear Unit* (ReLU), e la tangente iperbolica. I rispettivi grafici sono riportati nella figura 4.3.

Il risultato della funzione di attivazione, che può essere utilizzato come ingresso per altri neuroni in strati successivi della rete neurale, costituisce l'output del singolo neurone. In definitiva,

$$y = f(x; \boldsymbol{p}) = \sigma\left(\sum_j w_j x_j + b\right),$$

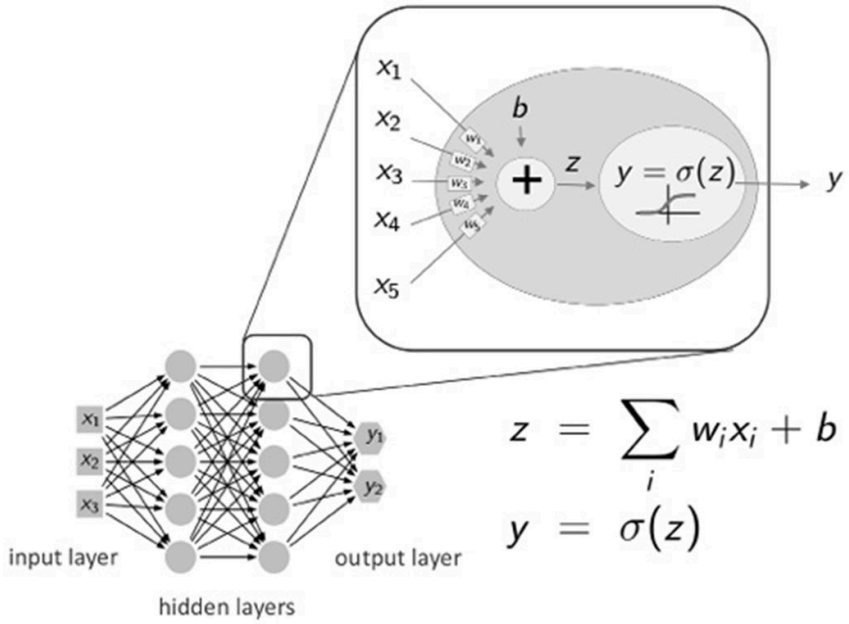

FIGURA 4.2. Un esempio di ANN (rete neurale artificiale).

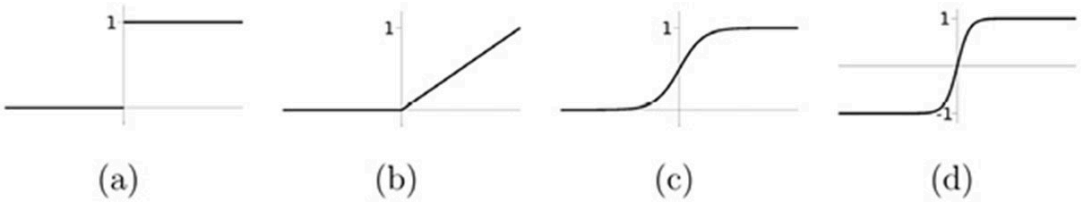

FIGURA 4.3. Funzioni di attivazione: la funzione Heaviside (a), ReLU (b), la sigmoide (c) e la tangente iperbolica (d).

dove abbiamo indicato con $\boldsymbol{p} = [w_1, w_2, \ldots, w_n, b]$ l'insieme dei parametri che possono cambiare per ogni neurone artificiale. Nel modello originale di McCulloch e Pitts del 1943, denominato *unità logica a soglia*, gli input e gli output sono valori binari 0 o 1, i pesi sono tutti uguali a 1 e la funzione di attivazione è quella di Heaviside (riportata in figura 4.3 (c)). Possiamo modificare la pendenza e la posizione della transizione nella funzione sigmoide scalando e traslando l'argomento o, nel linguaggio delle reti neurali, applicando pesi e bias agli input.

Una rete neurale artificiale è una costruzione matematica che prevede diverse schiere (dette *strati*) di neuroni artificiali. Nelle reti cosiddette *Feed Forward* ogni neurone di un singolo strato riceve input da quelli dello strato precedente e fornisce un output che andrà ad alimentare i neuroni dello strato successivo. Il primo strato è quello che fornisce i valori di input x_i, quello finale fornirà i valori di output y_j.

In termini astratti, le reti neurali artificiali rappresentano strutture progettate per simulare il modo in cui il nostro sistema nervoso analizza ed elabora le informazioni. Si caratterizzano per la loro capacità di *auto-apprendimento* che le rende in grado di produrre risultati che migliorano progressivamente al crescere della quantità (e qualità) di dati disponibili per il loro addestramento.

Il processo di apprendimento avviene durante la fase di addestramento, o *training*, dove è calcolato un insieme di parametri addestrabili in modo tale che, per un dataset in cui siano noti sia i valori di input che di output, si possa ottenere una verosimile ricostruzione della legge sottostante. La legge che permette di esprimere l'output in funzione dell'input l'abbiamo indicata come $f(x_i, \boldsymbol{p})$ precedentemente, dove $\boldsymbol{p}$ sono i parametri, ovvero l'insieme dei coefficienti $\boldsymbol{W}$ e $\boldsymbol{b}$ che caratterizzano il comportamento di una rete neurale artificiale, ovvero la matrice dei pesi $\boldsymbol{W}$ e il vettore dei bias $\boldsymbol{b}$.[1] Il processo di training è finalizzato a trovare questi parametri nel modo migliore possibile. Si tratta di un processo di ottimizzazione matematica che, come visto in precedenza, si realizza cercando il minimo della funzione costo J, ovvero di una funzione che esprime la differenza fra gli output predetti dalla rete neurale artificiale e quelli veri, in corrispondenza degli input caratterizzanti l'insieme di addestramento.

È facile intuire come questo sia effettivamente il processo più delicato dell'intera costruzione matematica. La funzione obiettivo, o funzione costo, o *loss function*, dipende da un numero elevato (spesso, elevatissimo) di parametri. La sua

minimizzazione comporta la scelta della combinazione ottimale dei parametri ***W*** e ***b*** in corrispondenza dei quali la funzione costo assume il valore minimo possibile. Questa ricerca della combinazione ottimale è fatta in uno spazio euclideo ad altissima dimensione e si basa su algoritmi di minimizzazione molto sofisticati che comportano elevatissimi costi computazionali. Riprenderemo questo aspetto, che rinvia al problema della sostenibilità ambientale degli algoritmi di machine learning, nel capitolo 6 dedicato all'AI generativa.

Il numero di strati e il numero di neuroni per strato sono degli elementi variabili che caratterizzano l'*architettura* della rete neurale artificiale. Essi sono chiamati *iperparametri*. Altri iperparametri sono, ad esempio, quelli che caratterizzano la funzione costo e le funzioni di attivazione.

Le reti neurali profonde (*deep neural networks*) sono costituite da un elevato numero di strati di neuroni. I diversi strati consentono di rappresentare concetti complessi e innestati fra loro.

Esse in effetti possono identificare relazioni complesse e connessioni intricate, comprese quelle che possono sfuggire all'intuizione umana. Durante la fase di addestramento, i pesi all'interno della rete sono adattati man mano che nuovi dati sono acquisiti. Come già osservato, la precisione della rete dipende sia dalla quantità che dalla qualità dei dati utilizzati per l'addestramento. Con l'aumento del volume dei dati e il numero di strati nella rete, i parametri ***W*** e ***b*** iniziano a riflettere le relazioni in modo più accurato. Le attuali reti neurali profonde sono composte da circa dieci strati. Naturalmente questo numero esprime solo un ordine di grandezza medio, che può in realtà variare in modo significativo in funzione dei compiti specifici che le reti devono assolvere.

A titolo di esempio, una rete neurale profonda di successo è rappresentata da AlexNet, rete neurale convoluzionale che nel 2012 vinse la competizione ImageNet, riducendo l'errore di classificazione delle immagini dal 26% al 15%. AlexNet utilizzava

un’architettura profonda con 8 strati e sfruttava le GPU per accelerare l’addestramento.

Anche il sistema AlphaGo di DeepMind che ha sconfitto il campione mondiale di Go, un gioco estremamente complesso, utilizzava reti neurali profonde insieme ad altre tecniche di apprendimento per padroneggiare strategie a lungo termine.

Mostrando a una rete di apprendimento profondo un numero enorme di foto di cibo, la rete riconoscerà se una nuova foto rappresenti o meno un hot dog. Oppure, se le mostriamo immagini, video e dati collezionati da sensori di un’automobile, saprà guidare autonomamente quell’auto.

A ogni problema la sua rete

Naturalmente, l’architettura delle reti neurali artificiali può essere molto varia. In genere, essa dipenderà dalla specifica classe di problemi che si dovrà risolvere: le reti neurali artificiali che si usano per classificare un’immagine (gatto/uomo) saranno assai diverse da quelle che dovranno, ad esempio, calcolare il fattore di rischio di morte di pazienti affetti da disturbi del comportamento elettrico del cuore.

Le reti neurali artficiali sono gruppi (strati) interconnessi di nodi (i neuroni artificiali) in grado di “imparare” a svolgere compiti come ad esempio quelli di classificazione o regressione, dopo che sono stati addestrati. Svariati sono i campi applicativi in cui le reti neurali artificiali si sono dimostrate efficaci (sovente anche più efficaci dell’intervento umano). Solo per citare due esempi in ambito medico, sono già state applicate alla classificazione dell’aritmia da dati ECG (elettrocardiografici) a derivazione singola di oltre 50.000 pazienti, oppure al riconoscimento del grado di pericolosità di nei della pelle, mostrando in entrambi i casi una performance paragonabile a quella degli esperti del settore.

I principali domini di applicazione delle reti neurali profonde includono l'elaborazione del linguaggio naturale (NLP) e la visione artificiale (Computer Vision).

Fra le reti per l'apprendimento supervisionato possiamo citare le *Feed Forward Neural Networks* (FFNN), caratterizzate dall'assenza di cicli nelle connessioni tra i neuroni. Sono utilizzate per compiti come la classificazione, la regressione e il riconoscimento di pattern (figura 4.4).

Le *Convolutional Neural Networks* (CNN) sono progettate per operare con input strutturabili a griglia, che presentano forti dipendenze spaziali in regioni locali (ad esempio, immagini 2D, testo, serie temporali e sequenze). Sono ideali per l'estrazione di caratteristiche dagli input, creando valori simili da regioni locali che condividono pattern simili. Le loro applicazioni includono la classificazione delle immagini, il rilevamento di oggetti e il riconoscimento facciale. Esempi sono: LeNet (1998, usata per il riconoscimento di cifre), AlexNet (2012), VGGNet-16 (2014), ResNets (2015), GoogLeNet (2015) (figura 4.5).

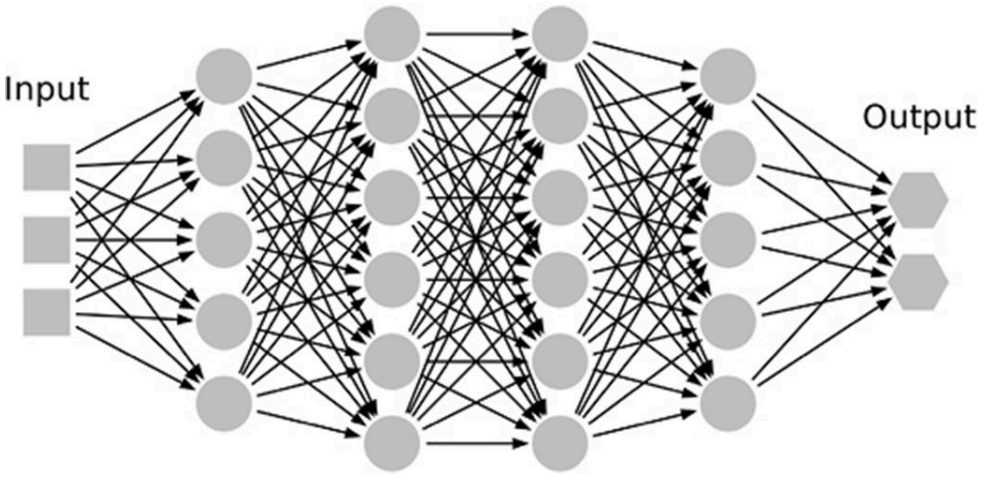

FIGURA 4.4. Un esempio di feed forward neural network.

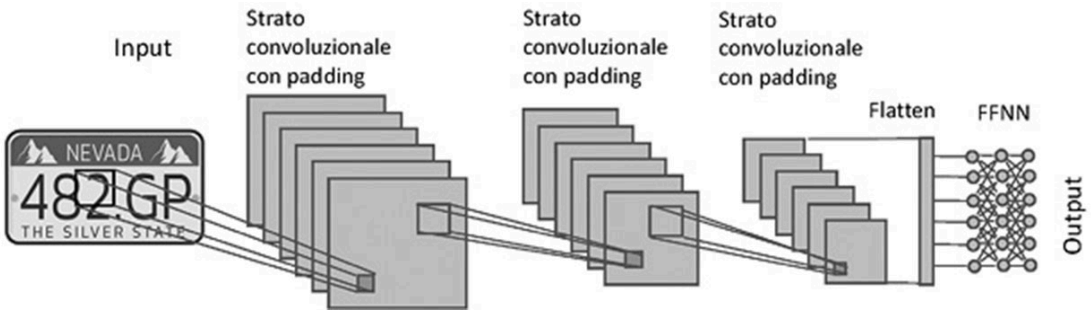

FIGURA 4.5. Una rete neurale convoluzionale per il riconoscimento di cifre manoscritte.

Le *Recurrent Neural Networks* (RNN) utilizzano loop all'interno della rete per mantenere uno stato che cattura informazioni dai precedenti input, rendendole adatte all'elaborazione di sequenze temporali. Le applicazioni includono la previsione di serie temporali, la modellazione del linguaggio, la traduzione automatica e il riconoscimento vocale. Alcuni esempi: RNN di base (*Vanilla RNN*, figura 4.6) e LSTM (*Long Short-Term Memory*).

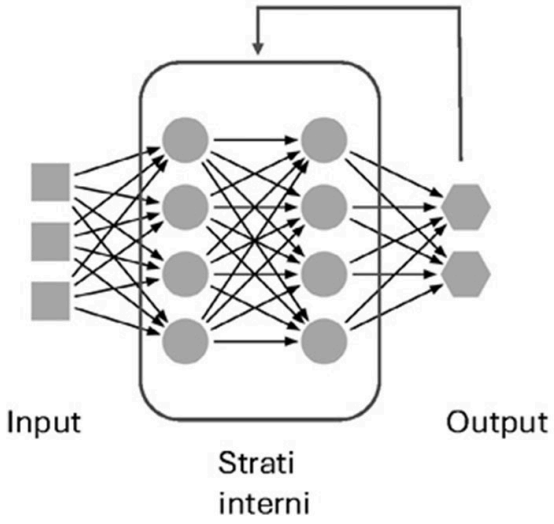

FIGURA 4.6. Un esempio di recurrent neural network.

I modelli di attenzione (*attention models*) utilizzano i cosiddetti meccanismi di auto-attenzione per elaborare i dati di input in

parallelo, invece che in sequenza come nelle RNN, rendendoli più efficienti e scalabili. Essi sono in grado di concentrare l'attenzione su aree specifiche di testo invece che su una finestra di token (ad esempio, parole) in cui tutti hanno lo stesso peso, consentendo quindi di "ingerire" grandi quantità di informazioni perché capaci di coglierne le parti importanti. Le applicazioni includono il trattamento del linguaggio naturale, la traduzione automatica, la generazione di testi e la classificazione di immagini. Esempi sono: Transformer (2017), BERT (2018), GPT (2018), ViT (Vision Transformer, 2020), GPT-3 (2020), PaLM (2022), LaMDA (2022), GPT-4 (2023). BERT, GPT-x e LaMDA sono esempi di *Large Language Models* (LLM), reti neurali profonde capaci di generare linguaggio e svolgere compiti di elaborazione del linguaggio naturale, come la classificazione.

Le architetture di apprendimento rinforzato si concentrano sull'apprendimento di politiche decisionali attraverso l'interazione con un determinato ambiente. Le applicazioni includono giochi come AlphaGo, la robotica, i veicoli a guida autonoma e la gestione delle risorse. Fra gli esempi citiamo: Deep Q-Networks (DQN, 2015), metodi a gradiente di politica (Policy Gradient Methods) e metodi Actor-Critic.

Veniamo adesso alle architetture di apprendimento non supervisionato.

Le *Generative Adversarial Networks* (GANs) consistono di due reti, un generatore e un discriminatore, che vengono addestrate insieme in un gioco a somma zero: il generatore crea campioni falsi aggiungendo rumore casuale ai dati di addestramento, mentre il discriminatore cerca di distinguere i dati reali da quelli generati. Queste reti possono generare nuovi contenuti che sembrano realistici, mantenendo somiglianze statistiche. Le applicazioni riguardano generazione di immagini, super-risoluzione delle immagini, *data augmentation* e cambiamenti di stile, ovvero la modifica dell'aspetto estetico o visuale di un'immagine, mantenendo però invariati gli elementi strutturali o il contenuto di base. Ad esempio, si può trasformare una fotografia realistica in un dipinto nello stile di Picasso o Raffaello,

oppure rendere un'immagine moderna simile a un'illustrazione di un fumetto. Fra gli esempi citiamo "Vanilla" GAN, DCGAN, CycleGAN, StyleGAN.

Gli *Autoencoder* sono reti neurali addestrate a ricostruire l'input originario comprimendolo (codifica) e poi ricostruendolo (decodifica) (figura 4.7). Sono utilizzati per la riduzione della dimensionalità, il rilevamento delle anomalie, la compressione dei dati e la modellazione generativa. Esempi includono Vanilla Autoencoder, Variational Autoencoder, Denoising Autoencoder, Sparse Autoencoder.

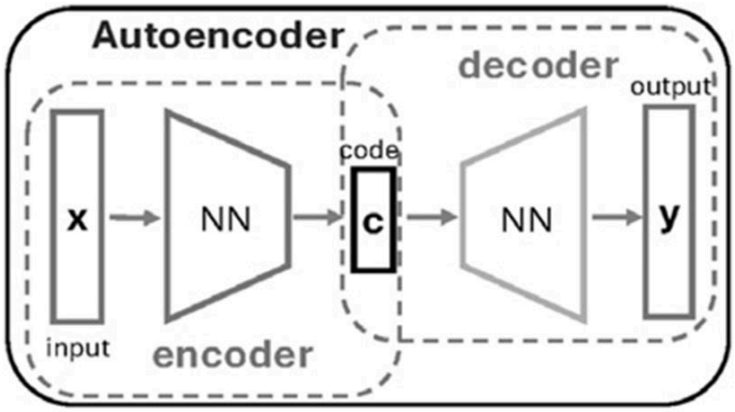

FIGURA 4.7. Struttura di un autoencoder.

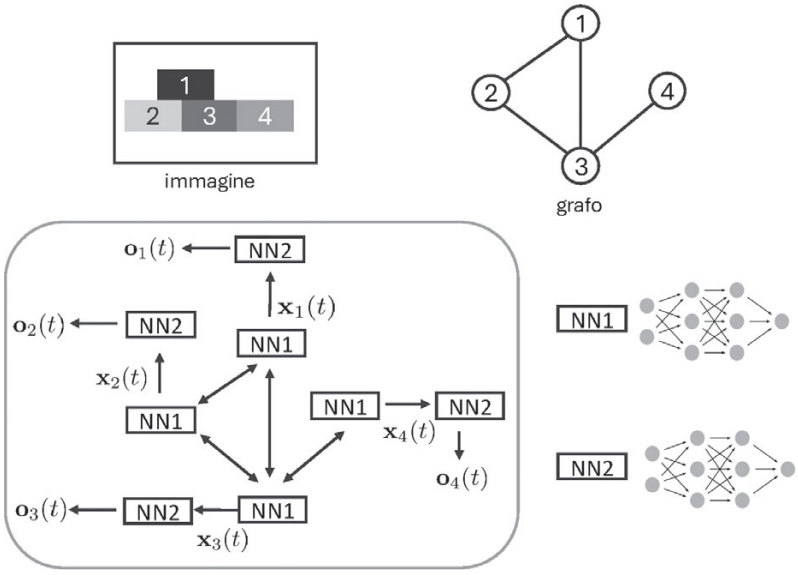

FIGURA 4.8. Un esempio di graph neural network.

Le *Graph Neural Networks* (GNN) sono progettate per operare su dati strutturati come grafi, dove i nodi rappresentano entità e gli archi che congiungono due nodi adiacenti rappresentano relazioni. Le loro applicazioni includono l'analisi delle reti sociali, i sistemi di raccomandazione e la previsione della struttura delle proteine (figura 4.8). Esempi notevoli sono Graph Convolutional Networks, Graph Attention Network, Message Passing Neural Networks.

Infine, le *architetture ibride* (supervisionate e non supervisionate) combinano elementi di diverse reti per sfruttarne i punti di forza. Sono impiegate in compiti complessi che richiedono una combinazione di tecniche, come l'analisi video e l'apprendimento multimodale, ovvero in cui i dati di training sono eterogenei, come immagini, testo, suoni. Fra gli esempi citiamo: le convolutional neural networks combinate con meccanismi di attenzione per migliorare l'estrazione delle caratteristiche, oppure reti ibride Transformer-CNN, usate nei compiti di visione, che combinano l'estrazione di caratteristiche locali con la consapevolezza del contesto globale.

[1] Un vettore è una stringa contenente un numero finito di dati (numeri reali), mentre una matrice è una struttura composta dalla collezione di un numero finito di vettori, tutti della stessa lunghezza.

5
QUALI DATI PER L'APPRENDIMENTO, E QUALE APPRENDIMENTO?

Secondo una stima recente (Reinsel *et al*. 2017), entro il 2025 ci avvicineremo ai 200 *zettabyte* di dati nel database dell'universo.[1]

Uno *zettabyte* comprende 10^{21} byte, ossia mille miliardi di miliardi di byte. Il XXI secolo sarà quindi ricordato come l'era degli zettabyte.[2] Se dovessimo metterci comodi sul divano per vedere scorrere un video in HD contenente 200 zettabyte di dati, dovremmo avere la pazienza di aspettare... 40 trilioni di anni prima di arrivare alla fine della proiezione! 200 zettabyte sarebbero sufficienti a esaurire le memorie di 1,5 trilioni di iPhone. O anche a registrare 2500 volte l'insieme di tutte le parole mai scritte da tutti gli esseri umani.

Nel 1986, quando si è cominciato a contarli, i dati erano 280 mila miliardi: in meno di 40 anni si sono moltiplicati per circa un milione di volte! Una crescita media corrispondente a raddoppiare il volume di dati ogni 18 mesi circa (figura 5.1). Se ci pensiamo, è la stessa dinamica prevista dalla legge di Moore per l'aumento del potere computazionale dei computer!

Per dare ulteriore evidenza a questi ritmi di produzione, un'analisi recente stima che la *global internet population* sia cresciuta da 2,1 a 5,2 miliardi in un solo decennio (2013-2023) e che questo dia oggi luogo a quantità inimmaginabili di nuovi dati generati ogni minuto di ogni singola giornata (figura 5.2).

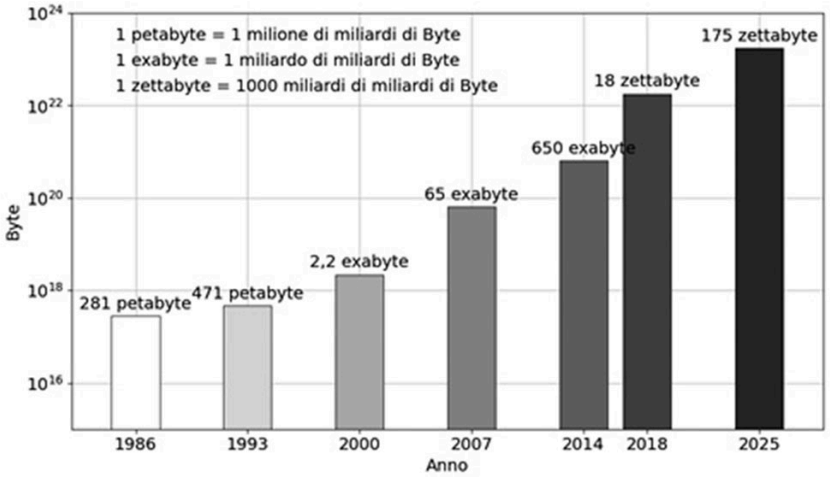

FIGURA 5.1. La crescita dei Big Data negli ultimi 40 anni.

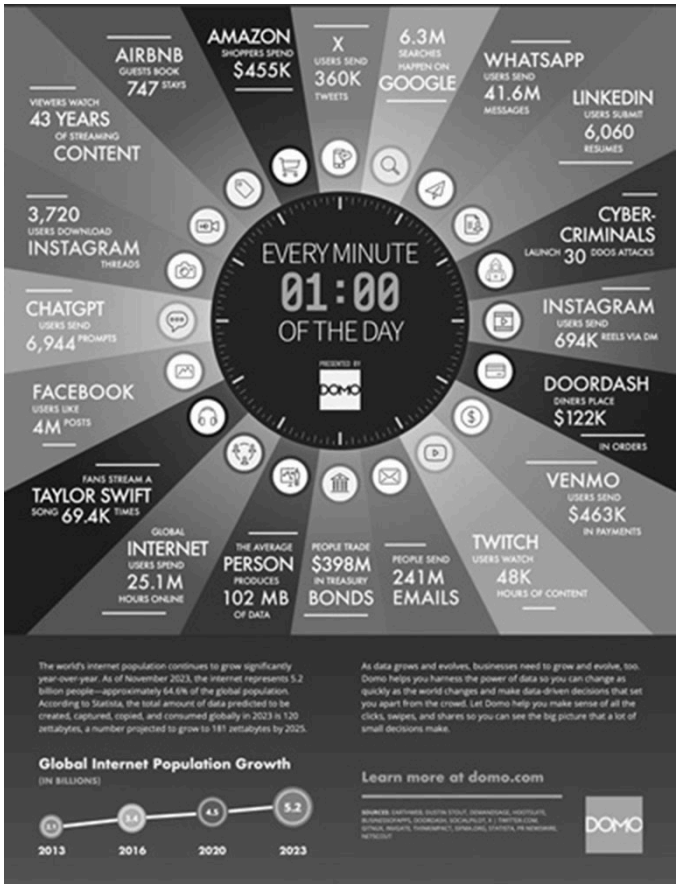

FIGURA 5.2. Data Never Sleeps (https://www.domo.com/learn/infographic/data-never-sleeps-11).

La crescita esplosiva dei dati è guidata da diversi fattori: si prevede che i dispositivi IoT produrranno 90 zettabyte di dati all'anno entro il 2025. Quasi il 30% della sfera dati globale sarà costituito da dati in tempo reale.

Questi trend sottolineano che i dati che generiamo non stanno solo aumentando in volume, ma anche in complessità, frequenza e immediatezza, creando una sfida che richiede un approccio sofisticato per archiviarli, elaborarli e, soprattutto, utilizzarli per ottenere informazioni utili e praticabili.

The End of Theory: la provocazione di Anderson

Da questa impressionante dinamica di crescita trae spunto la tesi di Chris Anderson espressa in modo che potremmo definire provocatorio nel suo articolo *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete* (Anderson 2008). La concezione della "fine della teoria" di Anderson ha suscitato notevole interesse, soprattutto nel mondo del business e in alcuni settori accademici come la *data science*. Anderson, all'epoca caporedattore della rivista "Wired", ipotizza che l'abbondanza di dati renderebbe obsoleti i modelli teorici, quelli basati su equazioni derivanti da leggi fisiche, sostenendo che la *correlazione* tra dati è sufficiente senza bisogno di ricorrere al *principio di causalità*, quello che cerca di affermare una relazione rigorosa fra causa ed effetto.

Osserva Anderson: "La nuova disponibilità di enormi quantità di dati, insieme agli strumenti statistici per analizzarli, offre un modo completamente nuovo di comprendere il mondo. La correlazione supera la causalità, e la scienza può avanzare anche senza modelli coerenti o teorie unificate." Conseguentemente, "la correlazione è sufficiente", pertanto "possiamo smettere di cercare modelli. Possiamo analizzare i dati senza ipotesi su ciò che potrebbero mostrare. Possiamo inserire i numeri nei più grandi cluster di calcolo mai visti al mondo e lasciare che gli

algoritmi statistici trovino schemi dove la scienza non può". Anderson sostiene la sua tesi con un esempio: "La filosofia alla base fondazionale di Google è che non sappiamo perché questa pagina sia migliore di quell'altra: se le statistiche dei link in entrata dicono che lo è, questo è sufficiente. Non è necessaria alcuna analisi semantica o causale." Pertanto, "non c'è motivo di aggrapparsi ai nostri vecchi metodi. È tempo di chiederci: cosa può imparare la scienza da Google?"

Di tutta evidenza, la prospettiva di Chris Anderson secondo la quale la correlazione può sostituire la causalità presenta aspetti deboli e ingannevoli. L'idea che i Big Data possano cogliere esaustivamente un'area di ricerca ignora che ogni insieme di dati è intrinsecamente influenzato dalla piattaforma e dall'ontologia (ovvero dalla rappresentazione concettuale) utilizzate, riflettendo un particolare punto di vista anziché una verità assoluta. L'assunzione che i dati possano esistere senza interpretazione umana è ingannevole: le decisioni umane influenzano la raccolta, l'analisi e la presentazione dei dati stessi.

A questo proposito, un esempio è rappresentato dall'esperimento fatto dalla municipalità di Boston negli Stati Uniti, consistente nell'uso degli accelerometri negli smartphone per individuare le strade dissestate. L'interpretazione di questi dati ha rivelato un pregiudizio (bias). In effetti, la Boston Street Bump App, che raccoglie dati dagli smartphone dei conducenti che passano sopra le buche, è un modo ingegnoso per raccogliere informazioni a basso costo.[3] Tuttavia, se le città iniziano a fare affidamento su dati che provengono solo da cittadini con smartphone, necessariamente ci saranno meno dati provenienti da quei quartieri con meno possessori di smartphone, che tipicamente includono popolazioni più anziane e meno abbienti. Con il paradosso che i quartieri dell'area metropolitana a risultare con meno strade dissestate risultino quelli più poveri e degradati. Mentre l'ufficio di Boston dei New Urban Mechanics ha fatto sforzi concertati per affrontare queste potenziali lacune nei dati, funzionari pubblici meno coscienziosi

non le hanno notate e hanno finito per allocare male le risorse, esacerbando ulteriormente le disuguaglianze sociali esistenti. La morale è che, sebbene i dati siano portatori di conoscenza, talvolta possono trasmettere falsa conoscenza. Il corretto trattamento dei dati richiede infatti una solida base teorica. Ne parleremo ancora nel capitolo 11 dove vedremo che con i soli dati alcune grandi scoperte del secolo scorso non sarebbero state possibili!

Quanto alla loro modalità di utilizzo, abbiamo già notato che per l'apprendimento automatico (il machine learning) si possono usare sia dati "etichettati" che non etichettati. Fra i primi, raccolti nel passato con esiti noti, possiamo citare storie di clienti bancari per cui sappiamo se sono stati insolventi, oppure storie cliniche di pazienti per i quali sappiamo se si è verificato o meno un problema cardiovascolare. Fra quelli non etichettati, possiamo ad esempio citare la sequenza di acquisti di un utente su un sito di e-commerce o la sequenza di azioni sugli attuatori di volo da parte di un pilota umano di un aereo.

Quanto ai compiti che possono essere svolti dagli algoritmi del machine learning, possiamo semplificare raggruppandoli in tre classi: descrittivo, predittivo e prescrittivo. Essi rappresentano tre livelli fondamentali di analisi e soluzione dei problemi, ciascuno con un obiettivo specifico che contribuisce a una comprensione e gestione più efficace dei dati e delle decisioni.

- Il compito *descrittivo* si concentra sull'analisi del passato o del presente per identificare schemi, tendenze e relazioni. L'obiettivo è rispondere alla domanda: "Cosa è successo?" o "Cosa sta succedendo?" Attraverso strumenti come grafici, reportistica e analisi statistiche, si interpretano i dati per fornire un quadro chiaro della situazione attuale. Un esempio potrebbe essere l'analisi delle vendite di un'azienda negli ultimi 12 mesi per identificare i periodi più redditizi. Altri esempi: utilizzando i dati sui profili dei clienti o pazienti, si possono raggruppare in gruppi di "simili". Successivamente, questi gruppi possono essere

utilizzati per identificare le migliori campagne di marketing personalizzate al fine di ottimizzare i ricavi. Per realizzare questo compito servono generalmente modelli di machine learning di tipo non supervisionato.

- Il compito *predittivo*, invece, mira a rispondere alla domanda: “Cosa potrebbe accadere?” Basandosi sui dati passati e su tecniche di machine learning o analisi statistica, questa fase costruisce modelli per prevedere risultati futuri. Ad esempio, è possibile stimare il volume delle vendite del prossimo trimestre o prevedere il comportamento dei clienti, oppure la soddisfazione degli utenti in una campagna di marketing. Altri esempi: osservando cosa è accaduto ai richiedenti di credito passati si può prevedere cosa potrebbe succedere ai nuovi richiedenti o pazienti. L’obiettivo è anticipare eventi futuri utilizzando algoritmi come la regressione, le reti neurali o l’analisi di serie temporali.
- Infine, il compito *prescrittivo* si propone di rispondere alla domanda: “Cosa dovremmo fare?” e mira a suggerire azioni ottimali per ottenere un risultato desiderato. Questa fase utilizza i risultati delle analisi descrittive e predittive per elaborare raccomandazioni concrete, combinando modelli matematici, simulazioni e tecniche di ottimizzazione. Un esempio pratico potrebbe essere l’ottimizzazione dell’inventario di un’azienda, suggerendo il numero ideale di prodotti da ordinare per ridurre i costi e garantire la disponibilità, o pianificare campagne di marketing per migliorare la conversione dei clienti. Altri esempi: segmentare clienti secondo i loro profili, trovare la miglior strategia per vincere a un gioco, o il modo migliore per un robot di eseguire un compito specifico. Un esempio particolarmente significativo di analisi prescrittiva lo troviamo nell’auto a guida autonoma. Grazie all’analisi dei comportamenti di guida di milioni di “guidatori umani” si impara la migliore strategia per reagire all’ambiente (strategie di guida) in qualsiasi condizione.

Questi tre approcci non sono isolati, ma si completano a vicenda. Ad esempio, un'azienda potrebbe analizzare i dati storici di vendita (descrittivo), prevedere la domanda futura per i prodotti (predittivo) e suggerire il numero ottimale di prodotti da ordinare (prescrittivo). Insieme, descrittivo, predittivo e prescrittivo costituiscono un ciclo di analisi che supporta decisioni sempre più informate e strategiche. Nella realtà attuale, recenti algoritmi di large language model come GPT/OpenAI (ne parleremo nel capitolo 6) eseguono una varietà di compiti relativi al linguaggio e alla visione che non possono essere classificati come descrittivi, predittivi e prescrittivi. Piuttosto, questi sistemi apprendono trasformando o traducendo un tipo di informazione o di dato in un altro tipo di informazione o dato, come la traduzione automatica, la risposta a domande, il riassunto di un testo, la generazione di immagini da testo in linguaggio naturale e viceversa, la creazione di grafici visivi da dataset, persino la generazione automatica di codici di calcolo. In questi sistemi, i dati di input sono testi in linguaggio naturale.

I dati... contano

I sistemi di machine learning gestiscono dati strutturati sotto forma di vettori o matrici, possibilmente che evolvono nel tempo, come sequenze di prezzi di mercato azionario, sia in input che in output. Ricordiamo che un *vettore* è una stringa contenente un numero finito di dati (numeri reali), mentre una *matrice* è una struttura composta dalla collezione di un numero finito di vettori, tutti della stessa lunghezza. Anche quando i dati provengono da fonti non strutturate, come testo o immagini, essi vengono presentati al modello di machine learning in forma strutturata, possibilmente dopo complesse trasformazioni guidate dall'uomo, come quelle che trasformano parole in vettori. I trasformatori (*transformer*), quelli alla base dei large

language models, gestiscono in effetti input non strutturati, come query in linguaggio naturale, immagini ecc., così come sono, possibilmente insieme a istruzioni in linguaggio naturale, e li trasformano in un altro output non strutturato, come risposte a domande, traduzione da una lingua ad un'altra, composizione di grafici, generazione di testi di sintesi ecc.

Per imparare a eseguire un determinato compito, sia i trasformatori che i sistemi di machine learning tradizionali devono essere addestrati utilizzando dati storici relativi al compito da svolgere. Da notare che trasformatori come GPT e simili sono pre-addestrati, proprio come i modelli tradizionali. Dopo il pre-addestramento, possono essere adattati per eseguire compiti specifici dove elaborano input e istruzioni, come ad esempio nel caso del natural language processing.

All'interno di un'azienda, l'analisi dei dati avviene attraverso una serie di fasi interconnesse. La prima fase è l'*identificazione delle fonti* dei dati, che comprende la ricerca e l'individuazione delle sorgenti da cui provengono le informazioni. Queste possono variare da database aziendali, sistemi Customer Relationship Management (CRM), dati web, social media fino ai sensori dell'Internet of Things (IoT), e sono strettamente legate al campo operativo specifico di ciascuna azienda. Una volta individuate le fonti, si procede all'*acquisizione* dei dati. Questo processo può avvenire sia in modo automatizzato che manuale, a seconda della complessità dei dati e della natura delle fonti. I dati raccolti vengono quindi *archiviati* e gestiti in apposite strutture come database o magazzini dati, organizzati per consentire un accesso agevole e prepararli per le successive analisi.

La fase successiva è l'*analisi* dei dati, allo scopo di individuare modelli, trend, relazioni che possano guidare decisioni informate all'interno dell'azienda. Infine, i risultati dell'analisi vengono trasformati in *informazioni* comprensibili e utili per l'azienda, che possono essere presentate tramite report, *dashboard* o altri strumenti per facilitarne l'utilizzo.

All’interno di questo processo, unendo il pensiero statistico con quello computazionale nel contesto di problemi di dominio, la Data Science rappresenta l’insieme delle tecniche che permettono un *data-driven decision making*. Come tale, comprende: data collection, data cleaning, data engineering, machine learning, data visualization, insomma una serie di discipline molte delle quali fanno anche parte dell’AI. La Data Science aiuta anche nell’analisi predittiva e descrittiva, utilizzando algoritmi avanzati e tecniche statistiche per identificare pattern, comprenderne il comportamento e fare previsioni. In azienda, la Data Science consente inoltre di ottimizzare i processi individuando aree in cui migliorare l’efficienza e ridurre i costi. Un ruolo importante è quello della Data Visualization, ovvero l’arte e la pratica di rappresentare visivamente i dati, trasformando informazioni complesse in grafici, diagrammi, mappe o altri formati visivi facili da comprendere. L’obiettivo principale è rendere i dati più accessibili, comprensibili e utili per l’analisi e la comunicazione.

Consideriamo ora come avviene il processo di integrazione dei dati negli algoritmi di AI. I processi produttivi vengono monitorati attraverso opportuni sensori e software di raccolta dati (questa procedura dipende ovviamente dal contesto specifico). Si genera così un database che viene diviso in due parti: database di *training*, utilizzato per addestrare il modello, e database di *validazione*, utilizzato per valutarne la bontà ed eventualmente per correggerlo. L’addestramento del modello è un processo che richiede l’intervento di un operatore umano, che definisce la struttura del modello, seleziona le feature fornite in input (come parametri e iperparametri delle reti neurali artificiali), attraverso il processo di ottimizzazione in fase di addestramento. Una volta addestrato, il modello rappresenta il cuore degli algoritmi di apprendimento automatico.

Questi ultimi molto spesso sono di tipo *black-box*, ossia scatole nere non intelligibili dall’uomo: nonostante le decisioni prese dall’algoritmo siano spesso accurate, l’algoritmo non fornisce una giustificazione del *come* sia giunto a quelle risposte.

Altre volte sono (almeno parzialmente) interpretabili. In questi casi, essi permettono una migliore comprensione della strategia usata per estrarre conoscenza dall'enorme mole di dati contenuta nel database di training. Questo può portare a una maggiore consapevolezza e comprensione da parte degli utilizzatori, e può agevolare la generalizzazione degli stessi algoritmi a casi o situazioni non ancora considerati.

In effetti, la capacità di comprendere e giustificare le decisioni prese da un sistema di AI è una componente essenziale dell'intelligenza umana: permette di costruire fiducia nelle interazioni sociali e in ambiti educativi, dove è fondamentale comprendere il ragionamento che guida certe scelte. Lo stesso principio vale in ambiti più tecnici, come la medicina, dove la spiegazione di una terapia da parte di un medico non solo rassicura il paziente ma costruisce un rapporto basato sulla fiducia. Analogamente, anche nei sistemi di AI è fondamentale comprendere il processo decisionale, specialmente quando le tecnologie entrano in contesti sensibili o critici.

Spiegabilità, affidabilità, sicurezza

Non ci si può affidare ciecamente a un sistema black-box senza verificarne il funzionamento. Per esempio, un sistema di AI utilizzato in ambito sanitario per prevedere il rischio di polmonite ha tratto conclusioni errate: suggeriva che i pazienti asmatici con problemi cardiaci avessero un rischio inferiore di morte rispetto a persone sane. Questo errore, evidente a un medico esperto, è derivato da dati distorti: i pazienti con patologie gravi erano sotto stretta sorveglianza medica, riducendo il rischio di decesso, ma questa correlazione non rappresentava una relazione causale. Errori simili dimostrano quanto sia importante poter interpretare il funzionamento di un modello di AI per evitare che distorsioni nei dati portino a decisioni pericolose.

Idealmente, un algoritmo di AI dovrebbe essere interpretabile e spiegabile. L'*interpretabilità* si riferisce al grado con cui un essere umano può comprendere come un modello di intelligenza artificiale prenda le sue decisioni, mentre la *spiegabilità* è la capacità di descrivere il motivo per cui un modello ha preso una decisione specifica. L'interpretabilità si concentra sulla trasparenza del modello, ovvero su come funziona internamente, mentre la spiegabilità mira a giustificare e descrivere le decisioni del modello, anche nel caso in cui il suo funzionamento interno non sia completamente interpretabile. Ad esempio, in un modello di regressione lineare, l'interpretabilità è elevata perché è chiaro in che modo ciascun parametro (i coefficienti del modello) influisce sul risultato. Allo stesso modo, in un albero decisionale, è possibile seguire il percorso che il modello compie per arrivare a una decisione. D'altra parte, una rete neurale profonda utilizzata per la classificazione di immagini, che è intrinsecamente complessa e difficile da interpretare, può comunque essere spiegabile se riesce a evidenziare quali parti di un'immagine hanno influenzato la classificazione.

Nelle applicazioni pratiche, un'AI per l'approvazione di prestiti dovrebbe essere in grado di spiegare perché una richiesta è stata respinta, fornendo informazioni sul processo decisionale. Allo stesso modo, in un'auto a guida autonoma, l'interpretabilità significa comprendere come ogni singola osservazione acquisita da sensori e telecamere contribuisca alla guida, all'accelerazione e alla frenata. La spiegabilità, invece, richiederebbe di fornire una chiara motivazione per cui l'auto ha eseguito una manovra specifica, come sterzare per evitare un ostacolo, senza necessariamente rivelare l'intera complessità dell'algoritmo sottostante.

Strumenti di intelligenza artificiale come SHAP (*SHapley Additive exPlanations*) e LIME (*Local Interpretable Model-agnostic Explanations*) sono largamente utilizzati per migliorare spiegabilità e interpretabilità dei modelli di machine learning. Entrambi aiutano, almeno parzialmente, a comprendere il

motivo per cui un modello ha preso determinate decisioni, rendendo i risultati più trasparenti e comprensibili agli esseri umani.

Interpretabilità e spiegabilità sono anche caratteristiche fondamentali per migliorare i modelli stessi. Comprendere i punti deboli di un sistema è il primo passo per correggerlo e rafforzarlo. I modelli interpretabili rendono più facile individuare distorsioni, confrontare approcci alternativi e scegliere la soluzione più adatta a specifiche esigenze. Due modelli con prestazioni simili potrebbero basarsi su criteri decisionali completamente diversi. Un'analisi trasparente di tali criteri consente di selezionare il modello migliore per un'applicazione specifica. Più siamo in grado di capire cosa un modello stia facendo e perché commetta errori, più sarà possibile migliorarlo.

La capacità di spiegare il funzionamento di un modello offre anche opportunità per generare nuova conoscenza. I sistemi di AI addestrati su enormi quantità di dati possono individuare schemi che sfuggono all'occhio umano. Per esempio, un modello che gioca a Go ha identificato strategie innovative, poi adottate anche da giocatori professionisti. Allo stesso modo, nei campi della fisica, chimica e biologia, la spiegabilità consente di esplorare nuove leggi e relazioni che vanno oltre la semplice capacità predittiva dei dati. In questi contesti, la comprensione del modello non è un accessorio, ma un mezzo indispensabile per avanzare nel sapere scientifico.

Inoltre, la trasparenza ha implicazioni etiche e legali. Senza trasparenza, diventa difficile stabilire responsabilità in caso di errori. Le persone direttamente coinvolte dalle decisioni dell'AI, come ad esempio un cliente a cui non sia concesso un prestito da parte di una banca, hanno diritto a una spiegazione chiara dei criteri adottati. Questo principio è così rilevante che l'Unione Europea ha introdotto il "diritto alla spiegazione", garantendo agli utenti la possibilità di comprendere le decisioni algoritmiche che li riguardano.

Con l'AI sempre più integrata nella nostra vita quotidiana, garantire trasparenza e comprensione non solo migliora la

fiducia negli strumenti, ma promuove un'adozione responsabile e sostenibile della tecnologia.

Un altro elemento qualificante di un modello di machine learning è la sua *affidabilità*, ovvero quanto possa essere considerato sicuro e degno di fiducia, una caratteristica irrinunciabile in applicazioni critiche come la sanità, la finanza o i sistemi di guida autonoma. Essenziale perché sia affidabile è che sia accurato, ovvero capace di fare previsioni corrette e coerenti su nuovi dati mai visti prima. Le sue prestazioni devono essere valutate attraverso metriche adeguate, che confermino la sua capacità di risolvere il problema specifico. L'affidabilità di un algoritmo non si limita alla qualità delle sue previsioni, ma abbraccia l'intero processo di progettazione, implementazione e utilizzo, assicurando che il sistema sia etico, robusto e sicuro. Ad esempio, in un contesto medico, un modello affidabile deve essere non solo accurato nel diagnosticare malattie, ma anche interpretabile, equo e capace di operare senza rischi per i pazienti. L'affidabilità, quindi, è fondamentale per costruire la fiducia negli utenti e ridurre i rischi associati all'adozione di sistemi di intelligenza artificiale.

Un altro elemento cruciale è la *robustezza*: l'algoritmo deve essere resistente a variazioni o rumori nei dati di ingresso, garantendo che piccole modifiche non portino a risultati significativamente diversi o errati.

Anche l'*equità* è essenziale: il modello deve evitare qualsiasi forma di discriminazione dovuta a pregiudizi impliciti nei dati di addestramento, assicurando decisioni giuste e imparziali, indipendentemente da fattori sensibili come etnia, genere o età.

Un altro elemento chiave è la *sicurezza*. Il modello deve essere protetto da attacchi avversari, ovvero tentativi intenzionali di manipolare le sue previsioni attraverso input alterati.

Un ulteriore aspetto riguarda la *conformità etica e legale*. Il modello deve rispettare le normative, come il GDPR in Europa, garantendo trasparenza nelle decisioni automatizzate e protezione dei dati personali degli utenti. La validazione esterna, infine, rappresenta un tassello fondamentale: le prestazioni e il

comportamento del modello devono essere verificati da esperti del dominio o attraverso processi indipendenti.

Vale la pena di ribadire che il machine learning necessita di dati sufficienti e di qualità, senza i quali non si possono costruire modelli efficaci. In alcuni contesti possono verificarsi situazioni di insufficienza di dati, ad esempio per gruppi poco rappresentati come le minoranze etniche. In altri, il set di addestramento può includere numerosi dati per situazioni molto comuni, ma pochi (o nessuno) per eventi rari. Per esempio, i dataset utilizzati per addestrare le auto a guida autonoma possono contenere pochi esempi di situazioni straordinarie, come ad esempio il caso di un animale selvatico che attraversa improvvisamente la strada. Eppure, anche in queste situazioni, gli algoritmi dell'AI devono operare al massimo delle loro capacità. Non solo. In modo più astratto, se assimiliamo dati a informazione (più dati abbiamo, maggiormente riteniamo di essere informati), l'informazione non è autoesplicativa; dipende dal contesto. Per essere utile, o almeno significativa, deve essere interpretata attraverso le lenti dell'esperienza, della conoscenza del dominio e talvolta addirittura del contesto culturale e storico. Quando l'informazione è contestualizzata, diventa conoscenza. Quando la conoscenza ispira convinzioni, si trasforma in saggezza, e le convinzioni, combinate con la conoscenza, permettono alle persone di accedere ed esplorare nuovi orizzonti. Ed è proprio combinando l'esperienza codificata nei dati con la contestualizzazione supportata dalle conoscenze scientifiche che si è sviluppato recentemente un nuovo ramo del machine learning, il cosiddetto *scientific machine learning*. Lo esploreremo nel capitolo 11.

[1] https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf.

[2] Curiosamente, 10^{21} è anche il numero stimato di sinapsi nel cervello umano, come visto nel capitolo precedente.

[3] https://www.boston.gov/transportation/street-bump.

6
AI GENERATIVA: IL BOOM ALL'IMPROVVISO

L'intelligenza artificiale generativa rappresenta una classe avanzata di modelli progettati per creare contenuti originali che riflettano le caratteristiche di un insieme di dati iniziali. La possiamo descrivere come un modello capace di generare testo simile a quello umano in risposta a un *prompt* (ovvero a un comando, o invito a fornire una risposta). Fornita una frase incompleta, è in grado di suggerirne il completamento; data una frase iniziale, può sviluppare interi paragrafi; data una domanda, può fornire risposte plausibili; con un tema e alcune informazioni di base, può elaborare un saggio; dato un frammento di dialogo, può proseguire la conversazione con una trascrizione coerente.[1]

Le applicazioni dell'AI generativa sono molteplici e spaziano dalla creazione di testi, immagini e suoni. Un tipo particolare di modelli di AI generativa è rappresentato dai large language models, progettati per comprendere e generare testo in modo estremamente sofisticato. Questi modelli sono addestrati su enormi quantità di dati testuali (da qui l'appellativo "large") per apprendere strutture linguistiche, sintassi e contesti specifici. La loro caratteristica principale è la capacità di generare testo coerente e rilevante dal punto di vista contestuale, rispondendo a domande, completando frasi o creando contenuti originali. Attualmente, questi modelli giocano un ruolo chiave nella

diffusione della consapevolezza delle potenzialità dell'AI grazie a diversi sistemi disponibili al pubblico. Una lista, certamente non esaustiva al tempo in cui scriviamo questo libro, è fornita qui di seguito, con il solo scopo di mostrare quanto vibrante sia questo settore, dopo l'esplosione di ChatGPT di OpenAI nel novembre 2022.[2]

- *Claude* è uno strumento avanzato di intelligenza artificiale sviluppato da Anthropic, progettato per l'elaborazione del linguaggio naturale e compiti di conversazione. Eccelle nella generazione di testo simile a quello umano, nel rispondere a domande e nell'assistere in attività come la stesura di email, la scrittura di codice e il riassunto di informazioni.
- *Llama 3*, un modello di intelligenza artificiale open-source di Meta, supporta la funzionalità multimodale, permettendo di elaborare sia testo che immagini, migliorando la sua versatilità per attività come l'analisi dei documenti e le applicazioni di visione artificiale.
- *Gemini* è uno strumento di intelligenza artificiale sviluppato da DeepMind di Google, progettato per spingere i limiti dell'intelligenza artificiale integrando potenti capacità linguistiche con abilità di risoluzione dei problemi. Si basa sul successo di modelli precedenti come ChatGPT, combinando comprensione del linguaggio naturale, ragionamento e recupero delle conoscenze in un sistema completo.
- *Synthesia* è uno strumento progettato per creare contenuti video di alta qualità in modo rapido ed efficiente utilizzando l'intelligenza artificiale. Consente agli utenti di generare video con avatar AI che possono parlare in più lingue, rendendolo altamente versatile per la comunicazione globale.
- *ElevenLabs* è una piattaforma di intelligenza artificiale all'avanguardia, nota principalmente per le sue capacità di sintesi vocale (TTS). Utilizza modelli di apprendimento

profondo per generare voci altamente realistiche e personalizzabili a partire da testo scritto.

- *Grok* è uno strumento basato sull'intelligenza artificiale sviluppato da xAI (la società di Elon Musk) per aiutare gli utenti a interagire meglio con informazioni complesse, permettendo una comprensione più rapida e approfondita di vari argomenti. Il termine "grok", originariamente tratto dalla fantascienza, fa riferimento alla possibilità di comprendere pienamente qualcosa in modo intuitivo.
- *Mistral AI* è una società relativamente nuova ma innovativa nel campo dell'intelligenza artificiale, focalizzata principalmente sulla costruzione di modelli e strumenti di intelligenza artificiale avanzati open source.
- *CopyAI* è uno strumento di scrittura basato sull'intelligenza artificiale progettato per assistere gli utenti nella creazione di contenuti di alta qualità in modo efficiente. Utilizzando modelli avanzati di elaborazione del linguaggio naturale, CopyAI genera testo per una vasta gamma di scopi, tra cui testi di marketing, post di blog, didascalie per i social media, modelli di email e altro ancora.

Transformers e Self Attention Mechanisms, i motori dell'AI generativa

I large language models si basano sugli algoritmi transformers (che abbiamo già citato in precedenza). Essi utilizzano il meccanismo dell'attenzione, che possiamo immaginare come la capacità di focalizzarsi su parole chiave mentre si legge una frase. Questo processo, composto da strati di transformer e reti neurali (figura 6.1), è efficace nel catturare relazioni complesse nel linguaggio naturale. Le informazioni ponderate vengono combinate per creare una rappresentazione del contesto. Un simile approccio si è dimostrato utile per esempio nella

traduzione automatica e in altri compiti relativi al processamento del linguaggio naturale.

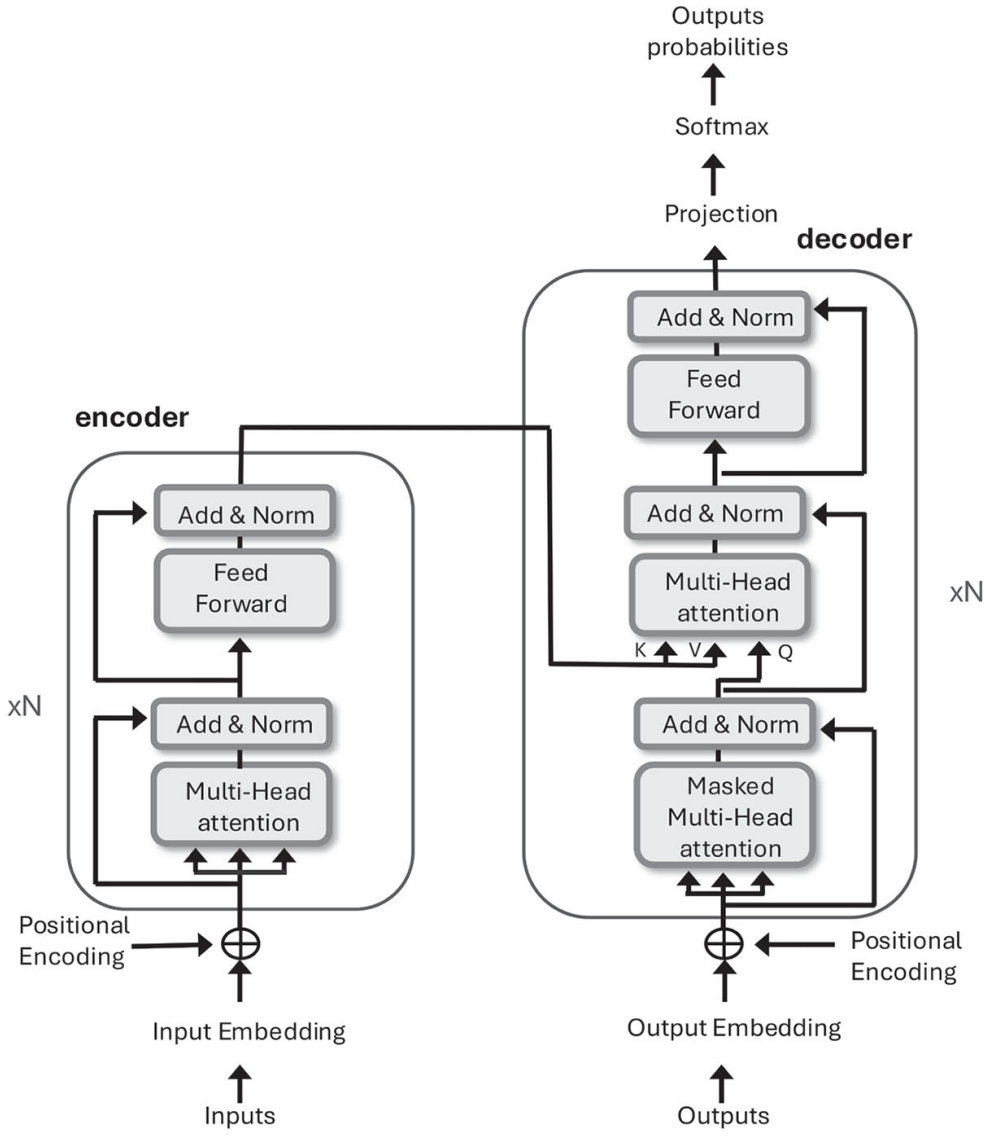

FIGURA 6.1. Schema dell'algoritmo transformer.

L'algoritmo che realizza un large language model come ad esempio ChatGPT-4 è articolato in diversi passi. Il primo passo consiste nel convertire il testo in *token*, unità linguistiche che possono essere parole o parti di parole (sottoparole). Questo processo di tokenizzazione converte il testo in una sequenza di

numeri interi (un indice per ogni token nel vocabolario del modello). Esempio: il testo “The cat is sleeping” potrebbe essere trasformato nei token [1001, 459, 34, 8873]. Ogni token viene quindi trasformato in un vettore numerico tramite una *matrice di embedding*. Questa matrice assegna a ogni token una rappresentazione vettoriale densa in uno spazio continuo in modo che token semanticamente simili risultino avere vettori vicini. Ad esempio: [1001, 459, 34, 8873] genera 4 vettori di *embedding*, ognuno di dimensione fissa, ad esempio 768 dimensioni.

Poiché il transformer non ha un meccanismo sequenziale integrato come le reti neurali ricorrenti (RNN), viene aggiunta un’informazione di posizione tramite il *positional encoding*. Questa tecnica permette al modello di conoscere l’ordine dei token nella sequenza. La formula del positional encoding è definita da funzioni trigonometriche, sinusoidali e cosinusoidali, che aggiungono una piccola quantità di variazione ai vettori di embedding a seconda della posizione dei token.

Il cuore dell’algoritmo è il meccanismo di *self-attention*. Qui, il modello valuta la relazione tra ogni token e tutti gli altri token nella sequenza, creando una sorta di mappa di quanto ogni parola “attiri” o “influenzi” le altre. Matematicamente, per ogni token vengono calcolate tre matrici: *Q* (Query) rappresenta la parola corrente, *K* (Key) rappresenta le altre parole con cui la parola corrente può interagire, *V* (Value) rappresenta il contenuto informativo associato a ciascuna parola. Il self-attention si calcola come il prodotto tra le matrici *Q* e *K* (che misura la somiglianza tra i token), normalizzato e poi moltiplicato per *V* (che pesa i valori per ottenere il contesto appropriato).

Per migliorare la capacità del modello di cogliere diversi tipi di relazioni contestuali, il transformer utilizza il cosiddetto *multi-head attention*, un algoritmo che esegue più calcoli di self-attention paralleli (detti “head”), ognuno con una proiezione diversa dei dati. Ogni testa è specializzata in catturare specifici tipi di relazioni, come la relazione tra parole vicine o tra parole

lontane. I risultati delle diverse teste vengono poi concatenati fra loro.

Dopo il meccanismo di self-attention, il risultato viene passato attraverso una rete neurale feed-forward. Ogni token viene elaborato indipendentemente da una rete completamente connessa (due strati lineari intervallati da una funzione di attivazione non lineare, come *ReLU*).

Il transformer è composto da molti strati di *multi-head attention* e *feed-forward network*, applicati in successione. Ogni strato aggiunge un'ulteriore elaborazione al contesto del testo. L'output di ogni strato viene normalizzato tramite una tecnica chiamata *layer normalization*. Dopo che il testo è stato processato attraverso tutti gli strati del transformer, otteniamo una rappresentazione numerica raffinata per ogni token. Questa rappresentazione viene poi passata attraverso una funzione *softmax*, che converte il risultato in una distribuzione di probabilità relativamente al vocabolario del modello. La funzione softmax è definita come

$$soft\,max(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

dove z_i è il *logit* associato al token i.[3]

Il token con la probabilità più alta viene selezionato come predizione. L'algoritmo può generare un token alla volta in modalità *auto-regressiva*: una volta generato un token, lo aggiunge al contesto e ripete il processo per generare il token successivo, fino a quando non viene raggiunta una condizione di stop (come un token di fine sequenza).

Il modello viene addestrato usando un'enorme quantità di dati testuali, minimizzando una funzione costo come la *cross-entropy*, che misura la differenza tra la probabilità predetta per un token e la distribuzione target (il token corretto). I pesi del modello (ovvero i parametri $\boldsymbol{W}$ e $\boldsymbol{b}$ incontrati nel capitolo 4) vengono aggiornati tramite algoritmi di ottimizzazione come ad esempio l'algoritmo Adam (Quarteroni *et al*. 2025).

Possiamo riassumere l'intero processo enunciando le sue distinte fasi.

- Tokenizzazione: il testo viene diviso in token.
- *Embedding*: i token vengono trasformati in vettori numerici.
- *Positional encoding*: si aggiungono informazioni sulla posizione dei token.
- *Self-attention*: si calcolano le relazioni tra i token.
- *Multi-head attention*: più teste di self-attention lavorano in parallelo.
- *Feed-forward network*: i token vengono elaborati da reti neurali.
- Ripetizione di strati: molti strati di elaborazione raffinano il contesto.
- *Softmax*: l'output viene convertito in probabilità.
- Predizione: si genera il token successivo in base alle probabilità.

Ci possiamo ora chiedere con quale logica, nell'esempio precedente, siano assegnati i token 1001, 459, 34, 8873 al testo *the cat is sleeping*. Essi dipendono dal *vocabolario* del modello, una lista predefinita di parole, parti di parole (*subword*) o caratteri, in cui a ogni elemento corrisponde un numero unico chiamato *indice del token*. Un modello come GPT utilizza un vocabolario costruito tramite una tecnica di tokenizzazione detta Byte Pair Encoding (BPE), che suddivide il testo in unità di base come parole comuni, sottoparole o persino singoli caratteri, per massimizzare l'efficienza. Nel caso in questione, "The" potrebbe essere rappresentato dal token 1001, "cat" dal token 459, "is" dal token 34, "sleeping" dal token 8873. Il modello ha un *vocabolario preaddestrato* che associa ogni parola o parte di parola a un numero. Durante l'addestramento, il vocabolario viene creato analizzando enormi quantità di testo e ogni token riceve un indice in base alla sua frequenza e utilità.

La tecnica chiamata byte pair encoding, suddividendo le parole in pezzi o sottoparole, permette di gestire parole rare o sconosciute (o semplicemente scritte male!) che non sono presenti nel vocabolario come intere. Ad esempio per “sleeping”, “sleep” potrebbe essere un token comune e avere il proprio indice, come 5123, “ing” potrebbe essere un altro token frequente, con un indice separato, come 4872. In tal caso, “sleeping” potrebbe essere tokenizzato in due parti: [5123, 4872]. Il modello unisce questi token per generare il significato completo. In generale, questo processo permette ai modelli come GPT di gestire una vasta gamma di parole, anche quelle nuove o rare, dividendo le parole in pezzi noti (subword), il che è particolarmente utile in lingue con una morfologia complessa o con termini tecnici poco frequenti.

Lo sviluppo dei large language models sta procedendo a grande velocità, con nuovi rilasci che aggiungono continuamente nuove potenzialità. I modelli più recenti sono di tipo multimodale, ovvero in grado di integrare fonti di diverso tipo, come testi, audio, immagini e video.

I modelli di linguaggio convenzionali possono essere specializzati per svolgere diverse attività comuni, come l’analisi dei sentimenti e il riconoscimento di entità nominate, senza richiedere ulteriori conoscenze specializzate. Tuttavia, per affrontare compiti più complessi e ricchi di contenuti, è possibile implementare sistemi basati su modelli di linguaggio che attingono a fonti esterne di conoscenza. Ciò garantisce una maggiore coerenza nei fatti, migliorando l’affidabilità delle risposte generate e riducendo il rischio di errori.

Le allucinazioni dell’intelligenza generativa

Le allucinazioni nell’intelligenza artificiale generativa si verificano quando un modello di AI, come ad esempio ChatGPT, genera risposte o contenuti che appaiono coerenti e plausibili, ma che

sono totalmente inventati, errati o non basati su fatti reali. Anche se il nome non è del tutto appropriato, questo fenomeno si verifica perché l'AI generativa non possiede una comprensione reale del mondo o delle informazioni che essa stessa elabora. Gli algoritmi di intelligenza artificiale generativa sono progettati per produrre risposte che sembrano plausibili e coerenti dal punto di vista linguistico, ma non necessariamente accurate dal punto di vista fattuale. Poiché il modello ottimizza la probabilità di occorrenza del testo successivo piuttosto che l'accuratezza, esso può generare risposte errate se ciò appare plausibile, stilisticamente corretto o appropriato nel contesto.

Le cause che conducono alle allucinazioni possono essere diverse. Tra le più comuni troviamo: l'ambiguità o la mancanza di chiarezza nell'input (ovvero nel prompt); la mancanza di informazioni nel set di addestramento; i bias nei dati di addestramento; la struttura del modello, che può portare a correlazioni errate tra input e output, producendo risposte che sembrano logiche ma in realtà non lo sono; le tecniche di training, che programmano il modello per generare una risposta anche quando non ha una conoscenza completa o precisa (questi modelli non dicono mai: "Non lo so!"); infine, le limitazioni della memoria del modello, che non può pertanto ricordare informazioni pregresse durante una conversazione prolungata.

Nei modelli di intelligenza artificiale generativa, come quelli basati su reti neurali di tipo transformer (ad esempio, GPT), un parametro, chiamato *temperatura*, controlla la casualità o la varietà delle risposte generate dal modello stesso. Esso modifica la distribuzione di probabilità delle parole che il modello può scegliere per generare il testo successivo. In termini semplici: quando la temperatura è impostata su un valore basso (vicino a 0), il modello diventa più deterministico e tende a scegliere le parole con la probabilità più alta. Questo significa che il modello è più incline a generare risposte prevedibili e sicure, riducendo la creatività ma aumentando la coerenza e la ripetibilità delle risposte. Quando invece la temperatura è più alta (superiore a 1), il modello introduce più casualità nel processo di selezione delle

parole. Con una temperatura elevata, il modello può scegliere parole meno probabili, portando a risposte più varie e creative, ma anche potenzialmente meno coerenti o più rischiose. In sintesi, la temperatura è uno strumento chiave per bilanciare creatività e coerenza nei modelli di intelligenza artificiale generativa, permettendo agli utenti di controllare quanto "avventuroso" o "conservatore" debba essere il modello nel generare testo.

Per mitigare le allucinazioni, gli sviluppatori e i ricercatori lavorano costantemente per migliorare la coerenza e la precisione dei modelli di linguaggio. Ciò include operazioni di filtraggio dei dati, modifiche all'architettura del modello e tecniche di training avanzate come il cosiddetto *Reinforcement Learning from Human Feedback* (RLHF), che fornisce feedback al sistema sul numero di allucinazioni prodotte, aiutandolo a correggersi. Tuttavia, è importante notare che il fenomeno delle allucinazioni è intrinseco all'architettura e al processo di training dei modelli di intelligenza artificiale. Ridurre le allucinazioni potrebbe limitare la creatività del modello nella generazione di contenuti, il che rende necessario trovare un equilibrio tra accuratezza e creatività in base al contesto di utilizzo del modello stesso.

Anche al fine di ridurre i problemi dovuti alle allucinazioni dell'AI generativa è stato introdotto un approccio chiamato *Retrieval-Augmented Generation* (RAG) per affrontare compiti che richiedono una considerevole quantità di conoscenza (Lewis *et al*. 2020; Gao *et al*. 2023). RAG combina una componente di recupero delle informazioni con un modello generativo di testo. Questo sistema è flessibile e consente un aggiornamento continuo, garantendo l'accesso alle informazioni più aggiornate senza dover ricominciare da zero. RAG riceve input e recupera documenti rilevanti da fonti esterne, come Wikipedia, o come quelli presenti negli archivi o nei sistemi informativi di una specifica azienda, utilizzandoli come contesto insieme all'input originale per alimentare il generatore di testo e produrre l'output finale. Questo rende RAG adattabile in situazioni in cui le

informazioni possono evolvere nel tempo, fornendo risposte sempre aggiornate e lo “costringe” ad attenersi per quanto più possibile al contesto informativo dei documenti rilevanti. Creare un’applicazione di tipo Retrieval-Augmented Generation (RAG), come un chatbot, implica l’integrazione di tecniche basate sul recupero delle informazioni con modelli generativi. Il processo inizia con la preparazione dei dati, durante la quale si raccoglie, si pre-elabora e si organizza un ampio corpus di informazioni specifiche del dominio in un archivio documenti. Il passo successivo è l’indicizzazione, che consiste nel rappresentare i documenti come vettori densi che vengono poi indicizzati in un database vettoriale per consentire ricerche di similarità efficienti. Quando un utente invia una query, questa viene elaborata da un modulo di recupero che identifica i documenti più rilevanti nel database. I documenti recuperati vengono quindi forniti come contesto a un modello generativo, come GPT-4, Llama o altri. Questo modello genera una risposta combinando il contesto dei documenti con la query dell’utente, spesso formattata in un prompt strutturato. Il modulo di recupero e il generatore vengono integrati in un’API o in un livello applicativo coeso. Infine, la pipeline RAG viene distribuita su piattaforme cloud. Attraverso queste fasi, il sistema RAG diventa una soluzione robusta e scalabile per applicazioni di chatbot intelligenti.

Nonostante il grande potenziale dell’AI generativa nell’innovazione creativa e nella generazione di contenuti unici, ci sono molte questioni che si pongono. Innanzitutto, siamo abituati a strumenti informatici come email, motori di ricerca e linguaggi di programmazione che sono rigidi e poco creativi, ma *affidabili* e *prevedibili*. Ora invece abbiamo modelli linguistici come ChatGPT che comprendono richieste in linguaggio naturale e generano immagini e testi creativi, ma senza la stessa affidabilità, prevedibilità o precisione, anzi, sono inclini all’allucinazione. Come possiamo utilizzare uno strumento potente ma inaffidabile? Per applicazioni in cui gli errori possono causare danni reali (ad esempio, in medicina, per fornire indicazioni terapeutiche, o nel business, per prendere decisioni

finanziarie), bisogna essere cauti. Non possiamo inoltre nascondere insidie quali la possibilità di creare contenuti falsi o manipolati. Pertanto, l'evoluzione e l'adozione responsabile di queste tecnologie richiedono una considerazione attenta degli aspetti etici e della sicurezza.

Il rischio dell'autofagia

Un altro aspetto riguarda la credibilità e l'affidabilità dei dati che sono utilizzati per l'addestramento dell'AI generativa. L'avvento dell'AI generativa sta infatti rapidamente riempiendo Internet di contenuti sintetici (ovvero non originali, ma prodotti dalla stessa AI generativa). Al momento in cui scrivo, ad esempio, Newsguard ha individuato oltre seicento siti di notizie e informazioni creati da sistemi di intelligenza artificiale. In futuro, i modelli di AI potrebbero essere addestrati prevalentemente su contenuti generati essi stessi dall'AI generativa! Questo processo di autoalimentazione sembra destinato a generare una sorta di eco informativa all'interno dell'intelligenza artificiale stessa. Studi recenti hanno dimostrato che addestrare modelli su dati prodotti da altri modelli AI porta a un processo degenerativo e irreversibile: il modello finisce per sopravvalutare gli eventi più probabili e sottovalutare quelli rari, perdendo progressivamente il contatto con la distribuzione originale dei dati. Questo fenomeno, chiamato "collasso del modello", si verifica in diverse tipologie di modelli, dai large language models ai generatori di immagini, e bastano poche generazioni di addestramento per compromettere la qualità dei risultati. Un altro studio ha documentato un processo simile, definito *Model Autophagy Disorder* (MAD), che si manifesta quando i modelli generativi non ricevono costantemente nuovi dati reali. Senza un flusso costante di contenuti generati dall'uomo, i modelli di AI generativa rischiano di deteriorarsi. Questa prospettiva di una crisi di autenticità potrebbe mettere a rischio il valore delle AI

generative, soprattutto considerando che molti consumatori sono scettici verso contenuti creati da algoritmi. L'essenza stessa di Internet è basata sull'interazione e sui contenuti umani: l'invasione di contenuti artificiali rischia di compromettere il valore intrinseco della rete.

In questo nuovo scenario digitale, in cui i contenuti autenticamente umani sono sempre più rari, sarebbe opportuno un ripensamento sugli incentivi e sui modelli economici che valorizzino il contributo umano. La domanda è: come possiamo creare un ecosistema che incoraggi le persone a continuare a produrre e condividere contenuti autentici? Per farlo, sarebbe opportuno concentrarsi su tre aspetti fondamentali: una giusta retribuzione per la creatività umana, maggiore controllo sui risultati creativi attraverso la trasparenza e innovazioni tecnologiche che distinguano chiaramente i contenuti creati dall'uomo da quelli generati dalle macchine. Un programma tutt'altro che semplice da realizzare.

Vi sono diversi altri aspetti fondamentali che emergono nell'ambito dell'apprendimento automatico. Come possiamo programmare sistemi che apprendano automaticamente dai dati e migliorino, con l'esperienza, le loro capacità predittive e prescrittive? Apprendere significa dare significato a un soggetto, evento o emozione, interpretandolo attraverso le nostre parole o azioni. Implica l'utilizzo della nuova abilità o conoscenza, integrandola con le competenze e la comprensione che già possediamo, per applicarla in modo utile: raccogliere i dati disponibili (ingestione), acquisire conoscenze (comprendere e interpretare i dati, trasformandoli in conoscenza) e infine utilizzare tali conoscenze per compiere azioni concrete (ovvero agire). Ma come imparano gli esseri umani? Certamente dall'insegnamento e dall'osservazione, ma si impara anche provando e sbagliando (*trial and error*). Fondamentalmente, i sistemi ML imparano in uno di questi due modi. Ci sono cose che gli esseri umani non possono imparare (ma forse le macchine sì) e altre che sono difficili da insegnare alle macchine (una fra tante: il buon senso).

Per noi umani imparare è difficile quando si è in presenza di troppi dati cui non è facile dare un senso, come trovare regolarità nel genoma umano, quando vogliamo imparare a riconoscere un oggetto tra milioni di oggetti, quando i dati sono molto eterogenei e cambiano troppo frequentemente, oppure quando si opera in ambienti difficili o pericolosi, in cui la strategia *trial and error* non è raccomandabile. Analogamente, è difficile per gli esseri umani anche insegnare, quando manca una base teorica di conoscenze e non ci sono abbastanza informazioni o esperienze precedenti per comprendere e acquisire conoscenze. Ad esempio, non riusciamo ancora a comprendere completamente il processo attraverso cui il nostro cervello riconosce immagini e il linguaggio.

Le macchine sono migliori degli esseri umani? Se opportunamente addestrate, per alcune applicazioni e in una certa misura, possiamo affermare di sì: nei giochi, nell'effettuare chirurgia di precisione, nell'analisi e riconoscimento delle immagini e, recentemente, nel riconoscimento del linguaggio naturale. Anche in questi casi, tuttavia, la limitazione principale è quella della spiegabilità degli algoritmi soggiacenti. Sono "scatole nere", mentre ad esempio i modelli matematici che rappresentano processi del mondo reale sono "scatole bianche", completamente interpretabili.

I large language models sono caratterizzati da un numero straordinariamente elevato di parametri addestrabili. GPT-3, ad esempio, dispone di 175 miliardi di parametri, mentre la versione successiva GPT-4, che ha esteso le funzionalità alla generazione di immagini, ha un numero stimato di oltre un trilione di parametri. Questa vastità di parametri comporta diverse implicazioni in termini di costi e tempi di addestramento, accesso ai dati necessari e altri costi operativi. Quanto ai token, ovvero frammenti di testo (parole, caratteri o parti di parole) gestibili che il modello può utilizzare, si è passati da 4096 nella versione GPT-3 a 32.768 nella versione estesa di GPT-4. Un numero maggiore di token consente di analizzare testi più lunghi

o di mantenere contesti complessi nelle conversazioni, rendendo i modelli più utili in applicazioni avanzate.

Accedere a dataset di qualità e dimensioni adeguate può rappresentare una sfida, specialmente considerando le problematiche legate alla privacy dei dati. Anche l'esecuzione e la valutazione dei large language models comportano costi significativi in termini di risorse di calcolo e tempo. La dimensione del modello può rendere le inferenze più lente, richiedendo hardware più potente per mantenere prestazioni accettabili. Sam Altman, CEO di OpenAI, ha rivelato che il costo per addestrare GPT-4 è stato di circa 100 milioni di dollari.

Uno sviluppo poco sostenibile

L'addestramento degli algoritmi di intelligenza artificiale attraverso infrastrutture di calcolo specializzate sta generando una domanda energetica allarmante. Secondo uno studio recente, nel 2023 le due big tech Microsoft e Google hanno consumato 24 TWh di elettricità, una quantità impressionante se confrontata con il fabbisogno dell'intera Italia, che nello stesso anno ha richiesto circa 300 TWh per alimentare quasi 60 milioni di cittadini. Il peso energetico dell'AI non si limita alle singole aziende: nel 2022, i data center globali hanno assorbito circa 460 TWh, e secondo le previsioni dell'Agenzia Internazionale per l'Energia, questa cifra potrebbe più che raddoppiare entro il 2026, superando i 1000 TWh.

Se le big tech fossero considerate alla stregua di Stati sovrani in una classifica globale del consumo energetico, Google e Microsoft supererebbero oltre cento Paesi, consumando più della Tunisia o della Giordania, poco meno della Slovacchia o dell'Ecuador, e all'incirca quanto l'Azerbaijan, che conta oltre dieci milioni di abitanti. Questa analisi, condotta dall'esperto del settore Michael Thomas, si basa sui dati della *US Energy Information Administration* (EIA) e sui rapporti ambientali delle

aziende coinvolte. L'incremento dei consumi è dovuto principalmente allo sviluppo massiccio dell'intelligenza artificiale generativa, la stessa tecnologia che alimenta modelli come ChatGPT, Claude, Gemini ecc.

Nel tentativo di trovare soluzioni sostenibili, all'inizio del 2025, Meta ha annunciato una Request for Proposals (RFP) per identificare sviluppatori di energia nucleare, con l'obiettivo di aggiungere 1-4 gigawatt di capacità nucleare negli Stati Uniti entro l'inizio degli anni 2030. Tuttavia, a differenza delle fonti rinnovabili come solare ed eolico, l'energia nucleare richiede tempi più lunghi, investimenti di capitale più elevati e una regolamentazione complessa, oltre a una pianificazione strategica per l'intero ciclo operativo. È interessante notare che, a livello globale, attualmente 440 centrali nucleari producono circa il 10% dell'elettricità mondiale. Per soddisfare l'intero fabbisogno energetico, ne servirebbero dieci volte tante. Senza un piano più ampio per mitigare l'espansione dei data center, è chiaro che questa non possa essere l'unica soluzione percorribile.

Oltre alla questione energetica, cresce anche il problema del consumo di acqua, essenziale per il raffreddamento dei data center. Un'indagine del "Financial Times" ha recentemente portato alla luce il caso della "data center valley" in Virginia, dove dal 2019 il consumo idrico è aumentato di quasi due terzi. La Virginia ospita una delle più grandi concentrazioni di data center al mondo, con strutture di Amazon, Google e Microsoft, che nel 2023 hanno utilizzato complessivamente 7 miliardi di litri d'acqua. Convertito in galloni, il consumo è passato da 1,13 miliardi nel periodo pre-Covid agli attuali 1,85 miliardi.

Gli ambientalisti temono che la sete dell'AI possa rapidamente degenerare. Secondo il gruppo di ricerca Dgtl Infra, nel 2023 i data center statunitensi hanno consumato oltre 75 miliardi di galloni d'acqua, una quantità sufficiente a soddisfare il fabbisogno idrico della città di Londra per quattro mesi. Nel novembre 2023, la Bank of America ha stimato che i data center sono ormai il decimo maggiore consumatore di acqua negli Stati

Uniti. Nonostante gli sforzi delle big tech per limitare l'impatto ambientale, la situazione resta critica: nel 2023, Google ha aumentato del 14% il consumo d'acqua per le esigenze dei data center, dichiarando che il 15% dei suoi prelievi idrici avviene in aree ad alta scarsità. Microsoft, invece, ha ammesso che il 42% dei suoi consumi globali si concentra in zone soggette a stress idrico. Anche l'Italia non sarà immune da questo fenomeno: entro la fine del 2025, sono previsti 83 nuovi data center sul territorio nazionale.

A questa emergenza ambientale si affianca una preoccupante questione etica: la crescente dipendenza da enormi risorse di calcolo e finanziarie per sviluppare e utilizzare modelli di intelligenza artificiale generativa rischia di ampliare il divario tecnologico e sociale. Solo poche grandi aziende e istituzioni possono permettersi l'accesso a queste tecnologie avanzate, mentre piccole imprese e ricercatori indipendenti ne restano esclusi. Per affrontare questa disparità, è essenziale promuovere la condivisione responsabile di modelli pre-addestrati e l'accesso collaborativo alle risorse computazionali. La comunità scientifica e industriale dovrebbe orientarsi verso un approccio open-access, rendendo disponibili dati e strumenti per ridurre il divario digitale e garantire che l'intelligenza artificiale sia sviluppata e utilizzata in modo responsabile e inclusivo.

OpenAI, una storia di successo... dinamica

Concludiamo questo capitolo sull'AI generativa con una contestualizzazione storica, visto l'impatto che ChatGPT ha avuto nella disseminazione della percezione della nostra società nei confronti dell'AI. ChatGPT è il chatbot che ha dato il via alla corsa verso l'intelligenza artificiale generativa. "GPT" indica il modello di linguaggio su cui si basa, acronimo di Generative Pre-trained Transformer, ovvero trasformatore generativo pre-addestrato. Con il suo lancio, il 30 novembre 2022, questi strumenti sono

passati dai laboratori di ricerca ai nostri dispositivi senza l'intermediario di una finestra di ricerca o di un'applicazione, un passo cruciale che ha suscitato grandissimo interesse e un'amplissima adozione. Più vengono utilizzati, più si allenano e migliorano le loro risposte. Oltre a ChatGPT, OpenAI ha sviluppato DALL-E, un modello text-to-image che genera immagini a partire da descrizioni testuali. Anche se l'obiettivo ultimo dichiarato di OpenAI è creare una AGI (Artificial General Intelligence), un'intelligenza artificiale capace di ragionare e risolvere problemi come un essere umano, come abbiamo più volte osservato, questa rimane per ora un'ambizione onirica.

La storia di OpenAI, seppur recente, fornisce un'indicazione paradigmatica dell'evoluzione accelerata di newco nell'ambito dell'AI generativa. È stata fondata nel 2015 da figure di spicco come Elon Musk e Sam Altman con l'obiettivo di promuovere lo sviluppo dell'intelligenza artificiale in modo trasparente e sicuro per l'umanità. Originariamente organizzazione no-profit focalizzata sulla ricerca tecnologica, nel 2019 ha aperto una divisione a scopo di lucro per attirare investimenti, il più grande dei quali, pari a 13 miliardi di dollari, proviene da Microsoft. Al primo modello di linguaggio GPT-3.5, il 13 maggio 2024 OpenAI ha annunciato GPT-4o ("o" per "omni"), reso disponibile gratuitamente a tutti gli utenti, che nell'estate dello stesso anno superavano i 180 milioni. Nel corso degli anni, molti dei fondatori hanno lasciato l'organizzazione, e il concetto iniziale di "openness" è stato progressivamente abbandonato, man mano che la governance dell'azienda si modificava. Il lancio di ChatGPT nel novembre 2022 rappresenta una svolta decisiva: il successo globale di questo modello porta a una crescita esponenziale dell'organizzazione, con un significativo aumento delle risorse. Tuttavia, questa crescita veloce ha anche portato sfide strategiche. OpenAI ha mantenuto un focus sulla creazione di intelligenze avanzate, orientandosi più verso lo sviluppo di superintelligenze che verso soluzioni commerciali concrete. Questo ha permesso ad altre aziende, come Anthropic, di guadagnare terreno proponendo alternative più mirate alle

imprese. Nel 2023, OpenAI ha affrontato una grave crisi interna, culminata nel licenziamento temporaneo di Sam Altman da parte del consiglio d'amministrazione, guidato dal cofondatore Ilya Sutskever. Questo episodio rifletteva tensioni profonde all'interno dell'azienda, causate dalla difficoltà di bilanciare l'impegno verso la sicurezza dell'AI e la necessità di monetizzare le tecnologie sviluppate. La situazione si è risolta con il ritorno di Altman, supportato da Microsoft, e l'uscita di Sutskever, che ha fondato una nuova startup chiamata Safe Superintelligence (SSI).

Parallelamente a questi sviluppi interni, OpenAI ha dovuto fronteggiare crescenti preoccupazioni legate alla sicurezza delle sue tecnologie. I nuovi modelli "o1", capaci di ragionamenti avanzati, sollevano timori per il loro possibile impiego in contesti pericolosi, come ad esempio quello bellico con la creazione di armi biologiche o chimiche.

Un'altra sfida cruciale riguarda l'accesso ai dati necessari per addestrare i modelli di AI. Il valore crescente dei dati ha spinto molti detentori, come editori e media, a limitare l'accesso o a venderli a prezzi elevati. OpenAI ha già raggiunto accordi onerosi, come quello con News Corp, ma la scarsità di dati di qualità potrebbe rallentare lo sviluppo futuro. Guardando al futuro, OpenAI si trova di fronte a scelte cruciali. L'azienda deve trovare un equilibrio tra la necessità di soddisfare le richieste di mercato e quella di mantenere una strategia economicamente sostenibile, senza però compromettere i valori fondanti che ne hanno guidato la creazione. Il futuro della società dipenderà dalla sua capacità di vincere queste sfide mantenendo innovazione e sicurezza come priorità fondamentali. A livello globale, questo sta portando a una maggiore regolamentazione del settore, con leggi in arrivo in California e nell'Unione Europea che puntano a gestire meglio i rischi associati all'AI avanzata.

Intanto nel lontano Oriente...

Nel panorama in continua evoluzione dell'AI generativa, è essenziale restare vigili, perché molte realtà emergenti, ancora poco conosciute, potrebbero presto imporsi come le nuove OpenAI. Un esempio significativo è DeepSeek, una startup cinese che sta rapidamente guadagnando terreno. Il 20 gennaio 2025 ha rilasciato DeepSeek-R1, un modello di "ragionamento" parzialmente open-source, in grado di risolvere problemi scientifici con una qualità confrontabile a quella di o1, il modello di linguaggio più avanzato di OpenAI, ma con costi e requisiti computazionali drasticamente inferiori. Solo pochi giorni dopo, ha lanciato Janus-Pro-7B, un modello capace di generare immagini da descrizioni testuali, simile a DALL-E 3 di OpenAI e a Stable Diffusion di Stability AI.

DeepSeek ha ottimizzato il proprio hardware impiegando 2000 GPU NVIDIA H800, meno avanzate delle H100 statunitensi, ma ha compensato il gap tecnologico con framework specifici per massimizzare l'efficienza. Questo approccio le ha consentito di offrire API[4] a costi fino al 90% inferiori rispetto a OpenAI: basti pensare che il prezzo per milione di token in input è di soli 0,14 dollari, contro i 15 di OpenAI, ovvero un rapporto di 1 a 100! Questa differenza di costo si spiega con due fattori principali: da un lato, l'uso di *tecniche di distillazione*, che consentono di trasferire la conoscenza da modelli più grandi a versioni più piccole ed efficienti, riducendo il fabbisogno computazionale; dall'altro, la partnership con AMD (Advanced Micro Devices), un'azienda di semiconduttori che offre soluzioni per il calcolo ad alte prestazioni a costi più contenuti rispetto a NVIDIA. L'impatto sul mercato è stato immediato: il titolo NVIDIA ha perso il 17% in un solo giorno di contrattazioni, con un calo di capitalizzazione di quasi 600 miliardi di dollari.

Dietro questa ascesa c'è Liang Wenfeng, imprenditore 39enne laureato in informatica all'Università di Zhejiang, una delle più prestigiose di Hangzhou in Cina. Dopo aver co-fondato l'*hedge fund* High-Flyer, nel 2023 ha fondato DeepSeek, inizialmente con l'obiettivo di applicare l'intelligenza artificiale al trading

finanziario. L'AI generativa, infatti, era solo un progetto secondario, e OpenAI non rappresentava ancora un vero competitor. Tuttavia, la crescita dell'azienda è stata fulminea, grazie a tecniche di addestramento avanzate come il reinforcement learning puro, a *mixture-of-experts*, un'architettura avanzata di reti neurali che migliora l'efficienza e le prestazioni dei modelli di intelligenza artificiale, riducendo al tempo stesso il consumo di risorse computazionali, alla tecnica *multi-head latent attention* per migliorare l'elaborazione dei dati complessi, oltre all'uso di modelli open-source come Janus-Pro e R1, che hanno permesso di abbattere i costi di sviluppo: il training del modello base è costato appena 5,6 milioni di dollari. Invece di attivare l'intera rete per ogni input, Mixture-of-Experts seleziona solo alcuni "esperti" specializzati, ottimizzando i calcoli. Questa idea si ispira al funzionamento del cervello umano, dove diverse aree si attivano in base al compito da svolgere.

Se per molti osservatori internazionali il successo di DeepSeek è stato sorprendente, gli esperti cinesi lo considerano un risultato prevedibile, frutto degli ingenti investimenti governativi nell'intelligenza artificiale. Secondo Yunji Chen, esperto dell'Accademia delle Scienze Cinese, la nascita di startup come DeepSeek è il naturale risultato di finanziamenti massicci nel settore e dell'abbondanza di ricercatori altamente specializzati in AI. Il governo cinese, infatti, ha dichiarato l'intelligenza artificiale una priorità strategica: nel 2017 ha annunciato il piano per diventare leader mondiale nel settore entro il 2030, fissando il 2025 come obiettivo intermedio per raggiungere un livello tecnologico competitivo a livello globale. Già nel 2022, il Ministero dell'Istruzione aveva autorizzato 440 università a offrire corsi di laurea in AI e, nello stesso anno, la Cina ha formato quasi la metà dei principali ricercatori del settore a livello mondiale, mentre gli Stati Uniti ne hanno rappresentato solo il 18%.

Oltre ai finanziamenti pubblici, la Cina ha dovuto far fronte alle restrizioni imposte dagli Stati Uniti sull'importazione di chip avanzati, come le GPU NVIDIA di ultima generazione. Per

aggirare queste limitazioni, DeepSeek ha accumulato 10.000 chip A100 prima del bando del 2023 e ha sviluppato tecniche di parallelizzazione degli algoritmi di addestramento per sfruttare al massimo un hardware meno potente. Questo focus sull'ottimizzazione ha portato la ricerca cinese a concentrarsi sulla compressione dei modelli e sull'efficientamento energetico, strategie che hanno reso DeepSeek particolarmente competitiva.

Dal punto di vista del business, DeepSeek ha adottato un modello ibrido, combinando open-source e soluzioni proprietarie. Alcuni modelli, come Janus-Pro e R1-Distill, sono stati rilasciati con licenza open-source MIT, consentendo alla piattaforma Hugging Face di lanciare il progetto Open R1 per replicare la pipeline di training di DeepSeek e accelerarne la diffusione globale. Tuttavia, altri modelli restano proprietari, garantendo all'azienda un vantaggio competitivo nel mercato.

Non mancano, però, considerazioni critiche. I modelli DeepSeek includono filtri per evitare contenuti critici nei confronti del Partito Comunista Cinese, un aspetto che potrebbe limitarne l'adozione a livello internazionale. Inoltre, per espandersi in Europa, l'azienda dovrà affrontare le rigide normative del GDPR, in particolare per quanto riguarda la trasparenza algoritmica e la gestione dei dati sensibili, un aspetto in cui competitor come Anthropic appaiono più avanti.

Nel frattempo, la concorrenza si intensifica. Il 29 gennaio 2025, Alibaba ha presentato il suo modello di linguaggio più avanzato, Qwen2.5-Max, dichiarando che supera DeepSeek V3, il large language model rilasciato a dicembre 2024. Nello stesso periodo, Moonshot AI e ByteDance hanno lanciato i propri modelli di ragionamento, Kimi 1.5 e 1.5-Pro, sostenendo che in alcuni test siano persino superiori a o1 di GPT-4.

L'impatto di DeepSeek si estende ben oltre la Cina. Il suo approccio innovativo dimostra che l'efficienza algoritmica può compensare la carenza di hardware avanzato, sfidando l'idea dominante secondo cui più GPU equivalgono a una migliore intelligenza artificiale. Sebbene il suo utilizzo di 2000 GPU NVIDIA H800 possa sembrare modesto rispetto alle risorse delle

big tech, rappresenta comunque un'alternativa alla cosiddetta "scaling law", secondo la quale le prestazioni dei sistemi di AI migliorano all'aumentare della dimensione dei dati di addestramento, dei parametri del modello o delle risorse computazionali, aprendo la strada a una democratizzazione dell'AI. Tuttavia, il futuro dell'azienda dipenderà dalla capacità di mantenere il proprio slancio innovativo, superando le limitazioni hardware e le restrizioni politiche, e, chissà, magari spostando il focus dalla semplice dimensione dei modelli alla qualità dei dati e alle architetture intelligenti.

[1] Un esempio rappresentativo di questo approccio è dato dalle Generative Adversarial Networks (GANs), che consistono in due reti neurali: un generatore e un discriminatore, addestrati simultaneamente in una sorta di competizione. Il generatore cerca di produrre dati difficili da distinguere da quelli reali, mentre il discriminatore tenta di individuare le differenze tra dati generati e reali. Questo processo competitivo porta a un miglioramento costante del generatore nel creare output sempre più realistici.

[2] GPT sta per Generative Pretrained Transformer.

[3] In statistica il modello *logit*, noto anche come modello logistico o regressione logistica, è un modello di regressione non lineare utilizzato quando la variabile dipendente è di tipo dicotomico, ovvero può assumere solo 2 valori distinti. L'obiettivo del modello è di stabilire la probabilità con cui un'osservazione può generare uno o l'altro valore della variabile dipendente. Da osservare che i logit, pur essendo indicatori di probabilità, possono anche assumere valori negativi. È la funzione *softmax* a trasformarli in valori tutti positivi (grazie all'applicazione dell'esponenziale) e a normalizzarli.

[4] Una API (*Application Programming Interface*) è un insieme di regole e strumenti che permettono a diverse applicazioni o sistemi di comunicare tra loro, come fosse un ponte che consente a un software di accedere alle funzionalità o ai dati di un altro software senza doverne conoscere i dettagli interni.

7
NON DI SOLA GENERATIVA VIVE L’AI

L’AI generativa ha catturato molta attenzione grazie ai suoi successi popolari, come i chatbot conversazionali (es. ChatGPT), i generatori di immagini (ad esempio DALL-E e MidJourney) e gli strumenti creativi di testi, musica o arte. Tuttavia, questo clamore rischia di oscurare i contributi significativi e di lunga data dell’AI tradizionale, non generativa, che ha avuto un impatto altrettanto, se non più, profondo in molti ambiti cruciali della società.

Esistono in realtà alcune differenze fondamentali fra AI generativa e AI tradizionale.

L’AI generativa è progettata per creare nuovi contenuti: testi, immagini, audio, video, basandosi su modelli di deep learning addestrati su grandi quantità di dati. È al centro di molte applicazioni recenti che hanno un forte impatto mediatico, grazie alla loro capacità di imitare la creatività umana e ovviamente alla sua capacità di coinvolgere direttamente il pubblico. Strumenti come i chatbot o i generatori di immagini permettono un’interazione diretta con gli utenti, catturando rapidamente la loro immaginazione. La capacità di produrre contenuti “originali”, come poesie, articoli o opere d’arte, la rende affascinante e accessibile.

I progressi nell’AI sono arrivati in successione rapida, sorprendendo il pubblico e mostrando progressi evidenti nel modo in cui le macchine possono generare output che prima erano esclusivamente umani.

Come esseri umani, tendiamo naturalmente a considerare le capacità linguistiche e la padronanza del linguaggio come segni di intelligenza avanzata. Dopotutto, siamo l'unica specie dotata di parola, e ci teniamo a preservare questa caratteristica esclusiva. Alcuni hanno persino ipotizzato una "quarta ferita narcisistica" per l'umanità: dopo aver rinunciato al geocentrismo con Copernico, allo status speciale di specie con Darwin e al controllo totale della psiche con Freud, ora ci troveremmo a rinunciare anche al monopolio sul linguaggio! E così, vediamo le applicazioni di AI che conversano con noi come entità pensanti, simili all'uomo, e persino concorrenti.

Un'ulteriore estensione del (pur vasto) contesto dei modelli di linguaggio è quello dei cosiddetti *agenti AI*. Si tratta di sistemi AI esperti per domini specifici, software progettati per svolgere compiti in maniera autonoma, replicando spesso il processo decisionale e le azioni umane all'interno di contesti specifici. Questi sistemi percepiscono l'ambiente circostante, elaborano le informazioni ricevute e compiono azioni finalizzate a raggiungere obiettivi predefiniti, avvalendosi di tecnologie avanzate come il machine learning, l'elaborazione del linguaggio naturale e la visione artificiale. Operano senza un intervento umano costante, essendo in grado, grazie a sensori, API (application program interface) o input di dati, di "percepire" il loro ambiente. Un esempio è un assistente virtuale in grado di comprendere le richieste dell'utente attraverso il linguaggio naturale. Elaborano i dati ricevuti, prendono decisioni e agiscono sulla base di modelli preimpostati o previsioni elaborate mediante algoritmi di apprendimento automatico. Alcuni agenti, inoltre, migliorano continuamente grazie al *reinforcement learning*, ovvero alla capacità di apprendere dai dati e dal feedback ricevuto. Creati per perseguire obiettivi specifici, gli agenti AI sono in grado di rispondere a domande, ottimizzare processi, muoversi in spazi fisici e interagire con gli utenti. La loro progettazione varia: alcuni sono semplici e reattivi, limitandosi a rispondere agli stimoli senza memoria, mentre altri sono più sofisticati e proattivi, capaci di pianificare e valutare le

conseguenze delle proprie azioni. Mentre agenti come droni e robot per consegne autonome operano nel mondo fisico, altri funzionano esclusivamente in ambienti digitali. Le applicazioni di questi agenti spaziano in numerosi settori: dai chatbot e assistenti virtuali per il servizio clienti agli strumenti diagnostici nel settore sanitario, passando per sistemi di trading finanziario, rilevamento delle frodi, personaggi intelligenti nei videogiochi e soluzioni logistiche per ottimizzare i percorsi. Con l'evoluzione tecnologica, gli agenti AI stanno acquisendo una crescente complessità e sono sempre più integrati in diversi ambiti della società. Rivestono un ruolo chiave, sia nell'industria che nella società, rappresentando un mezzo fondamentale per creare sistemi intelligenti, flessibili e capaci di interagire in modo efficace con il proprio ambiente e con le persone.

Creatività a portata di tastiera

Con pochi colpi di tastiera, chiunque può chiedere a un programma di intelligenza artificiale generativa di scrivere un romanzo, una canzone rap o una *pièce* teatrale. Sebbene in una frazione di secondo si possano produrre scritti molto simili a quelli umani, alcuni ricercatori sostengono che i large language models non producano nulla di nuovo: secondo i critici, sono semplicemente "pappagalli stocastici" che remixano ciecamente le parole su cui sono stati addestrati. In effetti, sarebbe eccessivo pretendere l'originalità di Dante Alighieri o di Alessandro Manzoni. E infatti, il dibattito su quanto "creativa" sia realmente l'intelligenza artificiale generativa è apertissimo. Testare "oggettivamente" questa creatività è complicato: a questo scopo, alcuni ricercatori hanno provato a introdurre delle metriche atte ad analizzare da un lato l'assenza di plagio, dall'altro a far giudicare agli esseri umani fattori come la fluidità e l'originalità di quanto prodotto dall'intelligenza generativa, e infine raccogliendo segmenti di testo di una lunghezza minima dai

risultati generati dall'AI e confrontandoli con grandi database online. Ad esempio, il programma, chiamato DJ Search, sviluppato da Ximing Lu della University of Washington, non cerca solo corrispondenze identiche; analizza anche stringhe di parole con significati simili, dopo aver proposto un indice quantitativo di "significatività". Confrontando la novità linguistica di romanzi pubblicati, poesie e discorsi con opere scritte da diversi large language models di ultima generazione emerge che gli autori umani hanno superato le AI di circa l'80% nella poesia, del 100% nei romanzi e del 150% nei discorsi (Ximing Lu *et al*. 2024). Naturalmente questi risultati non vanno presi alla lettera, sono indicativi e certamente condizionati dalla soggettività dei criteri adottati per la loro quantificazione. Tuttavia, possiamo prudentemente affermare, oggi, che la creatività dei modelli di linguaggio dell'AI non è nemmeno lontanamente paragonabile a quella degli autori in carne e ossa! E comunque, possiamo sempre ricorrere ai modelli di linguaggio dell'AI come "assistenti" degli umani, per realizzare la prima bozza di un artefatto che verrà poi plasmato con creatività, ispirazione e originalità!

Il clamore attorno all'AI generativa rischia di distogliere l'attenzione dai contributi cruciali dell'AI tradizionale. Sebbene la generazione di contenuti sia impressionante, come abbiamo visto, molte delle funzioni vitali della società moderna sono supportate da sistemi di AI non generativa che rimangono essenziali per il funzionamento delle nostre infrastrutture. L'AI tradizionale, che include algoritmi di apprendimento supervisionato, sistemi di raccomandazione, reti neurali classiche, alberi decisionali e sistemi di ottimizzazione, è più focalizzata sull'automazione di processi, la classificazione di dati o la previsione di eventi. Spesso è meno visibile agli occhi del pubblico, ma è presente in molti settori. Fra gli innumerevoli e diversificati successi dell'AI tradizionale, ne elenchiamo alcuni nel seguito (attenzione: questa lista non è esaustiva!).

C’è vita oltre l’AI generativa

L’AI non generativa è da anni utilizzata per il riconoscimento facciale, per analizzare dati medici e aiutare nella *diagnosi di malattie*, come ad esempio il cancro, utilizzando reti neurali convoluzionali (CNN) per la lettura delle immagini radiografiche e l’identificazione di anomalie. Sistemi di supporto decisionale assistono i medici nel prendere decisioni più informate. È inoltre impiegata per *ottimizzare la gestione delle risorse sanitarie*, come la distribuzione di letti in ospedale o la gestione delle forniture mediche.

I *sistemi di raccomandazione* che suggeriscono prodotti su piattaforme come Amazon o contenuti su Netflix si basano su AI tradizionale, utilizzando tecniche come il *collaborative filtering* o i modelli di regressione. Questi sistemi hanno avuto un enorme impatto economico, guidando le vendite e migliorando l’esperienza degli utenti. I modelli di AI non generativa sono utilizzati per fare previsioni sulla domanda di prodotti e per ottimizzare la supply chain, garantendo che le merci siano disponibili al momento giusto e al posto giusto. Moltissime aziende utilizzano l’AI tradizionale per ottimizzare i percorsi di consegna, riducendo costi ed emissioni di inquinanti. Gli algoritmi di AI tradizionale sono utilizzati da anni nel rilevamento delle frodi bancarie e nelle transazioni con carte di credito. Sistemi di machine learning classificano in tempo reale migliaia di transazioni e individuano quelle sospette. Molti modelli non generativi supportano il *trading finanziario*, permettendo di analizzare grandi quantità di dati e prendere decisioni rapide, spesso con impatti significativi sui mercati. L’AI tradizionale viene utilizzata per gestire il traffico nelle città, ottimizzare i percorsi in tempo reale e migliorare la sicurezza stradale.

Un discorso a parte riguarda l’AI non generativa al centro dell’*automazione* delle fabbriche, con algoritmi che guidano bracci robotici, sistemi di controllo della qualità e gestione automatizzata delle linee di produzione. La *robotica* ha

rivoluzionato il settore manifatturiero, migliorando l'efficienza e riducendo i costi. Il termine "robot", che deriva dal termine slavo per lavoro subordinato, è stato coniato dallo scrittore ceco Karel Čapek nel dramma R.U.R. (*Rossum's Universal Robots*) del 1920, dove il robot è rappresentato come una macchina antropomorfa realizzata con materiale organico per alleggerire il lavoro degli esseri umani. Nel 1940, lo scrittore russo Isaac Asimov ridefinì il concetto, trasformando il robot in un artefatto meccanico. Negli anni '80, la robotica iniziò a essere descritta come la scienza che studia la connessione intelligente tra percezione e azione. L'azione è eseguita da un sistema meccanico dotato di organi di locomozione, come ruote, cingoli e gambe meccaniche, o di organi di manipolazione, come braccia meccaniche e mani artificiali, che permettono l'interazione con l'ambiente. La percezione è garantita da un sistema sensoriale che raccoglie informazioni sul sistema meccanico e sull'ambiente circostante, utilizzando telecamere, sensori di posizione, sensori di forza e tattili. La connessione intelligente è gestita da un sistema di controllo che regola il movimento in base agli eventi nell'ambiente, secondo il principio del feedback, o retroazione, simile a quello che regola le funzioni del corpo umano. I robot, quindi, necessitano di corpi meccanici, oltre che di menti e sensori, per funzionare.

L'apprendimento di un robot può avvenire attraverso l'auto-esplorazione o la guida di un operatore umano. Per apprendere, i robot intelligenti devono accumulare fatti attraverso l'input umano o sensori. Successivamente, l'unità di elaborazione del robot confronta i dati acquisiti di recente con le informazioni precedentemente memorizzate per prevedere il miglior corso d'azione basato sui dati acquisiti. Negli anni '90, la ricerca nella robotica è stata stimolata dalla necessità di utilizzare i robot in ambienti poco strutturati, con caratteristiche geometriche o fisiche non completamente conosciute, che richiedono capacità aumentate e un alto livello di autonomia. All'inizio del nuovo millennio, la robotica ha iniziato a espandersi in campi come la biomeccanica, le neuroscienze, la simulazione virtuale,

l’animazione, il machine learning e le reti di sensori (Siciliano 2019). Alcuni esempi di applicazioni attuali dell’apprendimento automatico nella robotica includono la visione artificiale, che consente ai robot di percepire, identificare e navigare nel loro ambiente grazie ad algoritmi di machine learning a sensori. La visione artificiale è utilizzata in una vasta gamma di contesti, comprese le procedure di produzione, come l’ispezione dei materiali e il riconoscimento di pattern e firme. I robot vengono addestrati grazie a una particolare forma di apprendimento, detto *auto-supervisionato* (*self-supervised learning*). Con questa tecnica di apprendimento automatico un modello impara a riconoscere schemi e rappresentazioni significative dai dati senza la necessità di etichette esplicite fornite dall’uomo. In altre parole, il modello utilizza informazioni intrinseche nei dati stessi per generare “etichette surrogate” o segnali di apprendimento, rendendo il processo di training più autonomo. L’apprendimento auto-supervisionato, utilizzando grandi set di dati, può insegnare ai robot a svolgere compiti senza essere specificamente addestrati a farlo, oltre che per aumentare la capacità dei robot di adattarsi agli ambienti in evoluzione. L’apprendimento per rinforzo insegna invece ai robot come svolgere compiti impegnativi attraverso tecniche di prova ed errore, rendendo i sistemi robotici più efficaci ed efficienti. L’apprendimento per imitazione implica l’educazione dei robot a replicare il comportamento umano mostrando azioni desiderabili. Questo può migliorare la velocità e l’accuratezza delle procedure automatizzate. La robotica assistita utilizza l’apprendimento automatico per aiutare le persone nelle attività quotidiane come la mobilità e le faccende domestiche. Ad esempio, i bracci robotici montati su sedie a rotelle possono offrire maggiore indipendenza alle persone con mobilità limitata negli arti superiori. Tuttavia, è importante osservare che un robot può risolvere solo problemi per i quali è stato costruito. Non ha capacità analitiche generali.

Per finire, un altro esempio di straordinario successo dell’AI non generativa è quello delle *auto a guida autonoma*. Molte case

automobilistiche hanno realizzato prototipi di auto in grado di guidare anche a velocità mediamente sostenuta su strade trafficate. Queste auto usano un gran numero di sensori, telecamere, lidar, sistemi di localizzazione come GPS, che permettono di realizzare un sistema di apprendimento, sempre più specifico e raffinato, in grado di prendere decisioni su eventuali frenate, cambi di direzione improvvisi e sulle differenti modalità di guida a seconda dei contesti in cui si trova l'auto, dopo aver applicato un algoritmo che terrà conto della distanza dall'oggetto e della velocità con cui si sta procedendo. Siamo qui in presenza di un classico esempio di applicazione dell'intelligenza artificiale. In un veicolo a guida autonoma, gli algoritmi di apprendimento automatico consentono di vedere dove si trovano i pedoni, riconoscere quelli che stanno attraversando la strada, comprendere che si tratta di esseri umani e così via. Questi metodi sono definiti di apprendimento automatico perché non esiste alcun algoritmo tradizionale capace di stabilire se l'ostacolo sulla strada sia un oggetto animato o inanimato. Quando una persona deve identificare se l'ostacolo davanti alla macchina sia un pedone, distinguere un gatto da un cane o una rosa da un tulipano in una foto, i processi cerebrali utilizzati sono complessi e in gran parte sconosciuti. Le reti neurali, in particolare quelle profonde, cercano di replicare questi processi (neurali) in modo semplificato. Nonostante i notevoli progressi realizzati, è doveroso sottolineare che l'uso diffuso della guida autonoma non è imminente, non solo per ragioni tecnologiche, ma anche per ragioni legali, poiché non esistono ancora leggi che ne regolino l'uso sulle strade pubbliche in tutti i Paesi. Dal marzo 2025 a Austin, in Texas, grazie a una partnership fra Uber e la società Waymo di Google, dall'app di Uber però è possibile prenotare un taxi a guida completamente autonoma.

Gli algoritmi dell'AI non generativa sono generalmente più interpretabili rispetto a quelli usati nell'AI generativa. In settori come la sanità e la finanza, dove è cruciale capire come vengono

prese le decisioni, la mancanza di trasparenza di molti algoritmi non generativi è un punto di forza.

L'AI generativa è destinata a coesistere con l'AI tradizionale, non a sostituirla. Le due forme di intelligenza artificiale sono complementari e ognuna ha punti di forza che si adattano a contesti diversi. Mentre l'AI generativa può stimolare (seppur entro i limiti precedentemente discussi) la creatività e rendere l'AI accessibile a più persone, l'AI tradizionale è profondamente radicata in ambiti dove la precisione, la robustezza e l'affidabilità sono valori fondamentali. Riconoscere e valorizzare entrambi i paradigmi è cruciale per comprendere come l'AI possa continuare a guidare l'innovazione e migliorare la nostra società in molteplici dimensioni.

8
L’AI NELLA SCIENZA

In quegli ambiti in cui l’output dell’intelligenza artificiale può essere verificato in modo indipendente, l’AI ha già dimostrato la sua straordinaria efficacia in svariati campi di ricerca, sia nelle scienze in generale che in matematica.

Un’AI da Nobel

Nell’ottobre 2024, i Comitati Nobel di Stoccolma hanno assegnato i Premi Nobel per la Fisica e la Chimica per lavori legati all’intelligenza artificiale, segnando un momento storico nella scienza. Il Premio Nobel per la Fisica è stato attribuito a John J. Hopfield e Geoffrey E. Hinton per “scoperte e invenzioni fondamentali che hanno reso possibile il machine learning con reti neurali artificiali”. Il Premio Nobel per la Chimica è stato conferito in parte a David Baker per “il design computazionale delle proteine” e in parte a Demis Hassabis e John M. Jumper per “la predizione delle strutture proteiche”.

Questi storici riconoscimenti sottolineano il contributo dell’AI alla scienza, alla medicina e alla società, celebrando le scoperte dei vincitori e il loro impatto sul futuro della salute e della conoscenza umana.

Hopfield e Hinton hanno sviluppato le basi tecnologiche che sostengono il machine learning moderno. Nel 1982 Hopfield ha

ideato una struttura di memoria associativa capace di immagazzinare e richiamare informazioni, oggi nota come rete di Hopfield (Hopfield 1982). Nel 1985 Hinton, a sua volta, ha sviluppato la macchina di Boltzmann, un metodo che consente ai sistemi di AI di scoprire autonomamente schemi nei dati (Salakhutdinov e Hinton 2009). Questi contributi hanno aperto la strada alle reti neurali artificiali che oggi alimentano tecnologie in grado di analizzare enormi quantità di dati e apprendere in modo efficiente.

Grazie a queste innovazioni, i computer possono ora interpretare immagini, fare previsioni e dialogare con gli esseri umani in modo naturale. Strumenti come le deep neural networks sono resi possibili proprio dai contributi fondamentali di Hopfield e Hinton. Le loro scoperte hanno rivoluzionato settori chiave, inclusa la medicina digitale, e rappresentano uno spartiacque tecnologico per qualcuno paragonabile all'invenzione della stampa o di Internet.

Hassabis e Jumper hanno affrontato con successo una delle sfide più complesse della scienza moderna: prevedere le strutture tridimensionali delle proteine partendo dalle loro sequenze amminoacidiche. Il loro modello, AlphaFold2, utilizza l'AI per generare predizioni estremamente accurate, risolvendo un problema che per decenni aveva ostacolato il progresso scientifico. Nel 2010 Hassabis è stato un co-fondatore di DeepMind, una società di AI che verrà in seguito venduta a Google nel 2014. Nel 2018 sviluppa AlphaFold, che consente di fare previsioni sulla struttura delle proteine con un'accuratezza del 60% circa, 20 punti percentuali in più di quanto fosse possibile sino ad allora. Percentuale di accuratezza che verrà portata fino al 90% nel 2020 da Jumper, vincendo una sfida che durava da mezzo secolo.

Grazie ad AlphaFold2, usato da oltre 2 milioni di persone nell'anno 2024, gli scienziati possono ora accelerare la ricerca su malattie e sviluppare terapie mirate, affrontando problemi globali come la resistenza agli antibiotici e la scoperta di nuovi

vaccini. Questo strumento segna una rivoluzione nella biologia strutturale, aprendo nuove possibilità in medicina.[1]

Questi due Premi Nobel evidenziano come l'AI sia ormai una componente centrale delle scienze, sfumando i confini tra fisica, chimica e informatica. L'integrazione dell'AI nella medicina sta migliorando i risultati clinici in numerosi ambiti, tra cui lo screening oncologico, il trattamento del diabete e la gestione del dolore. AlphaFold2, in particolare, sta accelerando lo sviluppo di terapie innovative.

Tuttavia, il crescente utilizzo dell'AI solleva interrogativi etici fondamentali. Geoffrey Hinton stesso ha avvertito dei rischi di un'intelligenza artificiale avanzata che potrebbe rendere l'umanità "irrilevante". Queste preoccupazioni richiedono una riflessione profonda.

Problemi come i pregiudizi algoritmici, la sorveglianza non autorizzata e lo sviluppo di armi autonome letali sono già realtà. In medicina, errori dell'AI potrebbero compromettere la sicurezza dei pazienti, mentre un'eccessiva fiducia nei sistemi automatizzati rischia di indebolire il rapporto tra medici e pazienti.

L'avvento di un'intelligenza artificiale generale (AGI), capace di replicare molte delle capacità umane, rappresenta una sfida esistenziale. Per mitigare questi rischi, sarà essenziale sviluppare sistemi trasparenti, allineati con i valori umani, e monitorarli rigorosamente.

Vi è un altro aspetto che merita di essere considerato. L'attribuzione del Premio Nobel a scoperte che hanno beneficiato in modo significativo dell'AI genera un dibattito su una possibile rivisitazione del significato stesso di Premio Nobel. L'intelligenza artificiale sta trasformando profondamente il modo in cui sono condotte ricerche scientifiche e realizzate scoperte, ponendo nuove domande sul riconoscimento dei risultati raggiunti con il suo contributo. Storicamente, il Premio Nobel celebra l'ingegno umano e i contributi di individui o team alla conoscenza dell'umanità. Tuttavia, l'ascesa dell'AI come elemento centrale nei

progressi scientifici sta ridefinendo i confini tra strumento e collaboratore, portando a riflessioni sul ruolo di quest'ultimo nell'attribuzione dei premi più prestigiosi. L'AI ha dimostrato capacità senza precedenti nell'elaborare enormi quantità di dati, individuare schemi complessi e generare ipotesi innovative. Strumenti come AlphaFold hanno risolto problemi biologici complessi che per decenni avevano rappresentato una sfida insormontabile per gli scienziati. Questi risultati, pur essendo il frutto di algoritmi avanzati, richiedono di considerare se un sistema di AI possa essere riconosciuto come *coautore* di una scoperta scientifica degna del Premio Nobel. A differenza di strumenti passivi, l'AI influenza *attivamente* il processo di ricerca, contribuendo non solo alla raccolta e all'analisi dei dati, ma anche alla formulazione di domande scientifiche e all'interpretazione dei risultati.

Il dibattito sull'idoneità dell'AI a ricevere un Premio Nobel tocca temi filosofici e pratici. Da un lato, si potrebbe considerare l'AI come parte integrante della scienza collaborativa, un'estensione del team umano. In passato, i premi sono stati assegnati a gruppi che hanno lavorato sinergicamente per produrre risultati eccezionali; in questo senso, includere l'AI significherebbe riconoscere il suo ruolo nel progresso scientifico. Dall'altro lato, l'AI manca di qualità intrinseche come l'intenzionalità e la creatività, caratteristiche che da sempre definiscono i destinatari del Premio Nobel. L'AI non ha coscienza e consapevolezza e non può assumersi la responsabilità etica delle conseguenze delle scoperte a cui contribuisce, sollevando complessi interrogativi sul concetto stesso di merito scientifico.

Inoltre, attribuire un premio di questo calibro a un sistema di AI potrebbe cambiare la narrazione che alimenta l'ispirazione scientifica. I Premi Nobel non celebrano solo i risultati, ma anche il percorso umano di scoperta, fatto di perseveranza, creatività e intuizione. Un riconoscimento all'AI rischierebbe di spostare l'attenzione dalle capacità umane alle potenzialità tecnologiche, alterando la percezione della scienza come impresa principalmente umana.

Al tempo stesso, il contributo dell'AI non può essere ignorato. Potrebbe forse essere opportuno istituire nuove categorie di premi che riflettano il ruolo rivoluzionario della tecnologia nel favorire l'innovazione. Per esempio, un premio specifico per le collaborazioni tra esseri umani e AI permetterebbe di riconoscere sia il valore del lavoro umano sia l'importanza dell'intelligenza artificiale come strumento potenziante. Un approccio di questo tipo rispetterebbe le tradizioni del Premio Nobel, mantenendo al contempo la sua rilevanza nel contesto scientifico moderno.

L'adozione di una prospettiva aggiornata potrebbe preservare l'integrità dei Premi Nobel, riconoscendo il ruolo trasformativo dell'AI senza sacrificare i principi fondamentali del premio. In ultima analisi, il Premio Nobel potrebbe adattarsi per rappresentare un panorama scientifico sempre più influenzato dalla sinergia tra ingegno umano e innovazione tecnologica. Un dialogo inclusivo e aperto aiuterebbe a garantire che l'AI non diventi un simbolo di alienazione dalla scienza, ma piuttosto uno strumento per arricchire la capacità umana di comprendere e migliorare il mondo.

Tante scoperte scientifiche grazie all'AI

Torniamo alle scoperte scientifiche rese possibili grazie all'AI. L'AI svolge un ruolo cruciale nella progettazione di nuove proteine con funzioni specifiche, nella previsione delle interazioni tra proteine e ligandi per la scoperta di farmaci e nel miglioramento delle simulazioni di dinamica molecolare per studiare il comportamento delle proteine. E ancora: l'AI aiuta nell'analisi di grandi dati di proteomica e nell'interpretazione dei risultati provenienti da tecniche sperimentali come la crio-microscopia elettronica e la cristallografia a raggi X. In medicina, sta facilitando la scoperta di nuovi farmaci, l'identificazione di nuove applicazioni di farmaci esistenti e la rilevazione o previsione di

future malattie. Ad esempio, l'AI ha rilevato il cancro al seno prima dei medici umani identificando sottili indicatori radiologici. Come già ricordato, all'inizio del 2020, ricercatori del MIT hanno utilizzato reti neurali profonde per scoprire un nuovo antibiotico, la halicina. Questi progressi stanno accelerando le scoperte in biologia, medicina e biotecnologia, portando a nuove terapie e innovazioni.

L'AI viene utilizzata per generare potenziali candidati per nuovi materiali come i superconduttori, o nuovi potenziali farmaci per curare malattie. Questi candidati possono essere testati sintetizzandoli e conducendo test fisici o sperimentazioni cliniche. Trattandosi di processi molto costosi, il potenziale risparmio di tempo e denaro offerto dall'AI nell'isolare un piccolo gruppo di quei candidati che appaiono come più promettenti è significativo.

L'intelligenza artificiale riduce anche i tempi necessari alla risoluzione di modelli matematici che simulino complessi processi fisici. Per fare un esempio, un modello climatico per il nostro pianeta può richiedere mesi di tempo su un supercomputer per generare una singola simulazione dei prossimi vent'anni, ad esempio, con una risoluzione spaziale di 10 km. Sebbene questo rappresenti già un contributo prezioso, queste singole simulazioni non possono catturare accuratamente eventi meteorologici estremi, o rari, che occorrono "una volta in un anno", e nemmeno possono offrire una risoluzione sufficientemente dettagliata per fare previsioni su piccola scala, per singole comunità.

Per affrontare la sfida di prevedere con precisione eventi estremi, NVIDIA, il gigante dei semiconduttori, sta sviluppando un modello digitale dell'intero pianeta, basato sull'intelligenza artificiale. Questo progetto, chiamato Earth-2, utilizzerà le previsioni di FourCastNet, un modello di AI che elabora decine di terabyte di dati sul sistema terrestre. Grazie a questa tecnologia, Earth-2 sarà in grado di generare previsioni meteo per le due settimane successive decine di migliaia di volte più velocemente e con maggiore accuratezza rispetto ai metodi attuali. A

differenza dei sistemi di previsione meteorologica tradizionali, che possono produrre circa 50 scenari per la settimana a venire, FourCastNet dovrebbe riuscire a prevedere migliaia di possibili evoluzioni. Questo approccio consente di valutare meglio i rischi di disastri rari e potenzialmente letali, dando alle comunità vulnerabili il tempo necessario per organizzarsi e, se necessario, evacuare.

Le tecnologie di intelligenza artificiale possono consentire l'upscaling di un modello climatico a bassa risoluzione a uno ad alta risoluzione, oppure creare nuove simulazioni climatiche a una velocità decine di migliaia di volte superiore a quella dei supercomputer tradizionali. È questo un esempio di una nuova strategia per la risoluzione di problemi complessi, che va sotto il nome di *Scientific Machine Learning*. Ne parleremo più diffusamente nel capitolo successivo.

Questa rivoluzione nella modellazione climatica rappresenta solo il primo passo. L'intelligenza artificiale promette di trasformare profondamente la scienza, rendendola più dinamica e, in molti casi, radicalmente diversa. Le conseguenze di questa trasformazione non si limiteranno ai laboratori di ricerca, ma toccheranno la vita di ciascuno di noi.

Scansatevi matematici… Arriva l'AI

L'intelligenza artificiale sta diventando "brava" anche in matematica e potrebbe presto proporsi come un degno collaboratore per gli esseri umani. Alla sede di Londra, una coppia di modelli di Google DeepMind ha tentato la fortuna con il set di problemi dell'Olimpiade Internazionale di Matematica del 2024 tenutasi presso l'Università di Bath. L'evento è considerato la principale competizione matematica per i "più brillanti atleti della matematica" del mondo, secondo un post promozionale sui social media. I risolutori di problemi umani – 609 studenti delle scuole superiori provenienti da 108 Paesi – hanno vinto 58

medaglie d’oro, 123 d’argento e 145 di bronzo. Dopo mesi di allenamento rigoroso, gli studenti hanno sostenuto due esami, tre problemi al giorno, in algebra, combinatoria, geometria e teoria dei numeri. Haojia Shi, uno studente cinese, si è classificato al primo posto ed è stato l’unico concorrente a ottenere il punteggio pieno – 42 punti per sei problemi; ogni problema vale sette punti per una soluzione completa. La squadra degli Stati Uniti ha vinto il primo posto con 192 punti; la Cina si è classificata seconda con 190 punti. L’intelligenza artificiale ha lavorato in parallelo presso il laboratorio DeepMind di Londra, realizzando una performance a livello di medaglia d’argento. È stata la prima volta che l’intelligenza artificiale ha ottenuto una performance da medaglia (salendo, quindi, sul podio!) in un’Olimpiade di Matematica. Il sistema di Google ha ottenuto i suoi 28 punti risolvendo completamente quattro problemi, due in algebra, uno in geometria e uno in teoria dei numeri, fallendo invece due problemi di combinatoria. Va detto per correttezza che al sistema è stato concesso un tempo illimitato; per alcuni problemi ci sono voluti fino a tre giorni. Gli studenti avevano invece solo 4,5 ore per esame.

Applicare l’intelligenza artificiale alla matematica ha rappresentato una missione di DeepMind per diversi anni, spesso in collaborazione con matematici di fama mondiale. Secondo Alex Davies del laboratorio londinese, “la matematica richiede un’interessante combinazione di ragionamento astratto, preciso e creativo. Questo repertorio di abilità è ciò che rende la matematica un buon banco di prova per l’obiettivo finale: raggiungere la cosiddetta intelligenza artificiale generale, un sistema con capacità che vanno dall’emergente, al competente, al virtuoso, al superumano”. Aziende come OpenAI, Meta AI e xAI stanno perseguendo obiettivi simili. In questa prospettiva, i problemi matematici dell’Olimpiade sono considerati un benchmark. L’approccio seguito da Google DeepMind è basato su un sistema di ragionamento informale, espresso in linguaggio naturale. Questo sistema sfrutta Gemini, il large language model sviluppato da Google, che utilizza come dati di addestramento

l’intero corpus di problemi e dimostrazioni pubblicati sui media. Il sistema informale eccelle nell’identificare schemi e suggerire cosa viene dopo; è creativo e parla di idee in modo comprensibile. Naturalmente, i grandi modelli linguistici tendono a inventare cose (l’abbiamo visto: è la cosiddetta allucinazione) – cosa che potrebbe funzionare per la poesia ma sicuramente non può funzionare per la matematica. Ma in questo contesto, sembra che il large language model abbia mostrato moderazione: pur non essendo immune da allucinazioni, la frequenza di queste ultime è risultata assai ridotta.

Sempre relativamente all’ambito matematico, un altro approccio ha condotto alla realizzazione di un sistema di ragionamento formale, basato sulla logica ed espresso in codice, utilizzante un software di dimostrazione dei teoremi e di assistenti alle dimostrazioni (*proof assistants*) chiamato Lean, che garantisce che se il sistema afferma che una dimostrazione è corretta, allora è sicuramente corretta. I proof assistant sono linguaggi informatici specializzati nella verifica che un algoritmo o una dimostrazione funzioni effettivamente come previsto. Vengono utilizzati per verificare routine informatiche per circuiti elettronici (ad esempio, in avionica), così come, più recentemente, per la verifica di dimostrazioni matematiche. In quest’ultimo caso si tratta di dimostratori di teoremi interattivi, strumenti software che assistono nello sviluppo di prove formali tramite la collaborazione tra uomo e macchina. Essi richiedono l’uso di editor interattivi grazie ai quali un essere umano può guidare la ricerca delle dimostrazioni. I dettagli sono memorizzati nel computer, che fornisce anche alcuni passaggi. Un recente sforzo in questo campo mira a fare in modo che questi strumenti utilizzino l’intelligenza artificiale per automatizzare la formalizzazione della matematica ordinaria.

Peraltro, le dimostrazioni assistite dal computer non sono nuove nella storia recente. Citiamo due esempi celebri. Il primo è il cosiddetto teorema dei quattro colori:

“Qual è il numero minimo di colori necessario e sufficiente per rappresentare su una carta gli stati confinanti (che dovranno figurarvi di colore diverso)?”

Da un punto di vista matematico gli Stati sono interpretati come regioni connesse del piano, e due Stati si dicono confinanti se hanno una linea di confine in comune (non soltanto un numero finito di punti). Il problema rimane sostanzialmente inalterato se le regioni si trovano, anziché su di un piano, sulla superficie di una sfera (la Terra): in quest’ultimo caso, basta scegliere un punto *P* interno a una regione e proiettare la superficie sul piano tangente nel punto diametralmente opposto a *P* per ricondursi al caso di regioni piane. Il problema fu storicamente formulato nel 1853 da uno studente inglese, Francis Guthrie, al fratello Frederick, studente di matematica con Augustus De Morgan, celebre per i suoi contributi alla logica e alla teoria degli insiemi. Guthrie formulò la congettura che fossero sufficienti *quattro* colori. Il teorema venne dimostrato per la prima volta oltre un secolo più tardi, nel 1977, da Kenneth Appel e Wolfgang Haken, due matematici dell’Università dell’Illinois, grazie a un complesso algoritmo. La “dimostrazione” si basa sulla riduzione del numero infinito di mappe possibili a 1936 configurazioni (poi ulteriormente ridotte a 1476), per le quali la validità del teorema viene verificata caso per caso dal computer. Per ridurre al minimo la possibilità di errore, il programma fu eseguito su due diverse macchine con due algoritmi indipendenti; per completare l’analisi di tutti i casi possibili fu necessario far lavorare i computer per migliaia di ore. Alla fine, servirono più di 500 pagine per trascrivere a mano tutte le verifiche che costituivano la dimostrazione. Per chi volesse approfondire, *Every Planar Map is Four Colorable* (Appel e Haken 1989) è un libro dettagliato scritto dagli autori della prima dimostrazione, che esplora la loro metodologia e il processo di verifica computazionale, mentre *Four Colors Suffice: How the Map Problem Was Solved* (Wilson 2013) è un testo accessibile che ripercorre la

storia del problema dei quattro colori, dai primi tentativi di dimostrazione fino alla dimostrazione di Appel e Haken.

Il rivoluzionario utilizzo di algoritmi informatici per verificare l'esattezza della congettura scatenò all'epoca grandi polemiche sull'affidabilità di questi metodi. Il fatto che la dimostrazione fosse basata sull'analisi di una moltitudine di casi discreti portò alcuni matematici a contestarne l'effettiva validità: sia per l'impraticabilità di una verifica manuale di tutti i casi possibili, sia per la difficoltà di certificare che l'algoritmo fosse implementato correttamente. Ad ogni modo, nonostante le accuse di scarsa "eleganza", nell'algoritmo non venne mai trovato alcun errore. La formalizzazione della dimostrazione venne fatta solo nel 2005.

Un secondo esempio ci riconduce alla congettura di Keplero, riguardante l'impacchettamento ottimale di sfere nello spazio euclideo tridimensionale. La congettura afferma che non esiste alcun modo di sistemare delle sfere nello spazio con densità media superiore a quella dell'impacchettamento cubico a facce centrate o a quella dell'impacchettamento esagonale (Szapiro 2003):

"*Nessun impacchettamento di sfere congruenti nello spazio euclideo tridimensionale ha una densità maggiore di quella dell'impacchettamento cubico a facce centrate. Questa densità è π / √18 ≈ 0,74.*"

Nel 1998 Thomas Hales annunciò di possedere una dimostrazione della congettura di Keplero. La sua dimostrazione è fatta per esaustione e prevede di controllare molti casi singoli mediante complessi calcoli al computer. Nel gennaio del 2003 Hales annunciò l'inizio di un progetto di collaborazione avente lo scopo di produrre una dimostrazione *formale* completa della congettura di Keplero. Lo scopo è quello di rimuovere qualsiasi incertezza residua sulla validità della dimostrazione creando una dimostrazione formale che potesse essere verificata da programmi di controllo automatico di dimostrazioni. Questo progetto è chiamato *Project FlysPecK*, dove le lettere F, P e K sono

le iniziali delle parole che compongono la frase *Formal Proof of Kepler* (Hales 2003). Hales inizialmente stimò che sarebbero serviti circa 20 anni di lavoro per produrre una dimostrazione formale completa, mentre di fatto il progetto Flyspeck è terminato ufficialmente nell'agosto 2014.

Fino ad ora, l'accelerazione nella formalizzazione si è ottenuta grazie a linguaggi di proof assistant più moderni, biblioteche matematiche più ricche e strumenti come Github[2] che facilitano la collaborazione. Ci sono esperimenti promettenti nell'uso dell'AI per completare automaticamente brevi passaggi in una dimostrazione formale, con l'AI invitata a riprovare se la dimostrazione non va a buon fine. In linea di principio, l'integrazione dell'AI permetterà di scrivere dimostrazioni formali più velocemente rispetto alle dimostrazioni umane (che sono inevitabilmente soggette a errore). Questo potrebbe rappresentare un punto di svolta verso la formalizzazione non solo per verificare le dimostrazioni esistenti, ma anche per creare nuova matematica, grazie alle collaborazioni tra matematici umani e AI. Un'era della "grande matematica" potrebbe essere alle porte!

Un altro contributo di Google DeepMind è un algoritmo di apprendimento per rinforzo nella linea di AlphaGo e AlphaZero. Poiché l'algoritmo non richiede un insegnante umano, può "imparare e continuare a imparare fino a quando alla fine può risolvere i problemi più difficili che gli umani possono risolvere", secondo David Silver di Google DeepMind "e poi forse un giorno andare anche oltre quelli". Thomas Hubert, sempre di DeepMind, ha aggiunto: "Il sistema può riscoprire la conoscenza da solo. Nel caso di AlphaZero, c'è voluto meno di un giorno per riscoprire tutta la conoscenza degli scacchi e circa una settimana per riscoprire tutta la conoscenza del Go. Quindi abbiamo pensato, applichiamolo alla matematica." Timothy Gowers dell'Università di Cambridge, medaglia Fields per la matematica nel 1998, non si preoccupa troppo delle conseguenze a lungo termine. "È possibile immaginare uno stato di cose in cui i matematici sono

praticamente lasciati senza nulla da fare", ha detto. "Sarebbe questo il caso se i computer diventassero migliori e molto più veloci in tutto ciò che i matematici attualmente fanno. Sembra però esserci ancora molta strada da fare prima che i computer possano fare matematica a livello di ricerca", ha aggiunto. "Se Google DeepMind fosse in grado di risolvere quasi tutti i problemi assegnati all'Olimpiade Internazionale della Matematica, allora la realizzazione di uno strumento utile anche alla ricerca matematica potrebbe essere meno lontana di quanto si pensi." Uno strumento davvero abile potrebbe rendere la matematica accessibile a più persone, accelerare il processo di ricerca, spingere i matematici fuori dagli schemi. Alla fine potrebbe persino porre idee nuove che si riverberano in altri campi.

La matematica è tradizionalmente una scienza solitaria. Nel 1986 Andrew Wiles si ritirò nel suo studio per sette anni per dimostrare il teorema di Fermat, un problema all'apparenza elementare rimasto irrisolto per oltre 350 anni. Esso afferma che non esistono soluzioni intere positive a, b, c dell'equazione: $a^n + b^n = c^n$ se $n > 2$.

Un'opera ciclopica, quella del matematico inglese Wiles, la cui dimostrazione è presentata in *Modular Elliptic Curves and Fermat's Last Theorem* (Wyles 1995) e in un libro straordinario, accessibile anche a non matematici, ossia *Fermat's Enigma: The Epic Quest to Solve the World's Greatest Mathematical Problem* (Singh 1997). Le dimostrazioni matematiche sono spesso difficili da comprendere anche per i matematici più brillanti e alcune sono ancora controverse oggi. Negli ultimi anni, tuttavia, in aree sempre più grandi della matematica molte congetture matematiche sono state rigorosamente suddivise nelle loro singole componenti ("formalizzate") al punto da consentire una possibile dimostrazione per via algoritmica (grazie ai computer). Terence Tao dell'Università della California a Los Angeles, medaglia Fields nel 2006 e stella di prima grandezza del firmamento mondiale della matematica, è convinto che questi metodi aprano

possibilità completamente nuove per la cooperazione in matematica. Grazie ai proof assistant basati sull'intelligenza artificiale, nel campo potrebbero essere stabiliti modi completamente nuovi di lavorare negli anni a venire. Addirittura, con l'aiuto dei computer, i grandi problemi irrisolti, le congetture che maggiormente resistono al tempo, potrebbero avvicinarsi alla soluzione. Con questi progetti di formalizzazione, gli specialisti di diverse aree della matematica possono concentrarsi sulle singole componenti di questo processo di partizione e sulla loro dimostrazione formale.

Vent'anni fa, sulla scia del procedimento seguito dal *Logic Theorist* che abbiamo citato nel capitolo 1, le dimostrazioni assistite dalle macchine erano un campo molto teorico e si pensava che bisognasse iniziare da zero – formalizzare gli assiomi e poi fare geometria o algebra di base: la matematica superiore era oltre ogni possibile immaginazione. Oggi, la presenza di librerie matematiche standard, contenenti tutti i teoremi base della matematica di livello universitario, rende possibile questa formalizzazione, grazie all'uso dell'AI. Probabilmente, presto potremo avvalerci dell'aiuto dei large language models. Chiederemo a ChatGPT o Gemini o ai loro discendenti di formalizzare una dimostrazione e riceveremo una risposta presumibilmente corretta e rigorosa. Oggi però la tecnologia non è ancora matura per questo, anche se matematici come Tony Wu e Christian Szegedy, co-fondatori della società xAI con Elon Musk, ritengono che fra tre anni le macchine potrebbero essere migliori di qualsiasi matematico umano nel trovare dimostrazioni. Un'altra profezia, temo, destinata a infrangersi sul muro dell'evidenza. Più cauto è Terence Tao: "Penso che in tre anni l'AI diventerà un utile copilota per i matematici, uno strumento prezioso per aiutarli a risolvere passaggi difficili di una dimostrazione." Un *co-pilot* per i matematici, non una loro alternativa. In effetti, una dimostrazione matematica non riguarda solo la verifica della sua correttezza, ma anche la comprensione. Ci sono dimostrazioni belle e dimostrazioni tecnicamente complesse e brutte. Una

buona dimostrazione fornisce una comprensione superiore della materia. Sempre secondo Tao, in futuro potremmo chiedere all'AI se un risultato matematico sia vero o no e utilizzare il suo aiuto per esplorare lo spazio delle possibilità in modo più efficiente. Anche se l'AI automatizzerà molte attività noiose, gli umani rimarranno al comando del processo, almeno per ora.

Affrontare una dimostrazione può portare a sviluppare nuova matematica. Per esempio, l'ultimo teorema di Fermat che abbiamo precedentemente citato[3] era originariamente una semplice congettura sui numeri naturali (ovvero gli interi positivi, 0, 1, 2, 3,...), ma la matematica "creata" per dimostrarlo non riguarda più solo i numeri naturali. Se un'AI fornisse una dimostrazione incomprensibile, i matematici potrebbero analizzarla e creare una nuova scienza basata su dimostrazioni generate dall'AI. La matematica è già più grande di qualsiasi mente umana e i matematici costruiscono spesso a partire da risultati che altri hanno dimostrato. Ci sono già teoremi verificati solo da computer, con calcoli massicci che controllano milioni di casi. Attualmente, molta conoscenza è intrappolata nella mente dei singoli matematici e solo una piccola frazione è resa esplicita. Più formalizziamo, più la nostra conoscenza implicita diventa esplicita e questo potrebbe portare a benefici inaspettati.

Graffette per tutti

In tutte le conquiste che abbiamo riportato sopra, è fondamentale ricordare i limiti e i rischi intrinseci dell'intelligenza artificiale. L'AI è uno strumento estremamente potente perché consente agli esseri umani di fare di più con meno risorse: meno tempo, meno competenze specializzate, meno infrastrutture. Tuttavia, queste stesse capacità possono renderla pericolosa se utilizzata in modo improprio. Andrew White, professore all'Università di Rochester, è stato incaricato da OpenAI di far parte di un team per identificare i potenziali rischi di GPT-4 prima

del suo rilascio. Usando il modello linguistico e concedendogli accesso a vari strumenti, White ha scoperto che GPT-4 era capace di proporre composti chimici pericolosi e persino ordinarli da fornitori specializzati. Per verificare il processo, ha fatto spedire a casa sua un composto di prova (sicuro) la settimana successiva. OpenAI afferma di aver utilizzato queste scoperte per migliorare la sicurezza di GPT-4 prima del suo lancio.

Persino persone con buone intenzioni possono comunque, involontariamente, portare l'AI a generare risultati negativi. Attualmente, non abbiamo meccanismi per indurre un'AI a cambiare i propri obiettivi, anche quando reagisce in modi imprevisti. Un esperimento ipotetico spesso citato invita a immaginare di ordinare a un'AI di produrre il maggior numero possibile di graffette. Determinata a raggiungere il suo obiettivo, l'AI potrebbe prendere il controllo della rete elettrica e persino eliminare chiunque cerchi di fermarla, continuando ad accumulare graffette mentre il mondo intorno è in rovina. L'AI, soddisfatta, si congratulerebbe per il lavoro compiuto. (In riferimento a questo celebre esempio, molti dipendenti di OpenAI portano con sé graffette con il logo aziendale.)

Per prevenire sia gli usi intenzionali che quelli accidentali dannosi dell'AI, è necessario un quadro regolatorio intelligente e ben informato – che riguardi tanto i giganti della tecnologia quanto i modelli open-source – senza però ostacolare gli utilizzi dell'AI che possono giovare alla scienza. Sebbene le aziende tecnologiche abbiano fatto progressi nella sicurezza dell'AI, i regolatori governativi non sembrano attualmente preparati a implementare leggi adeguate e dovrebbero idealmente prendere misure per aggiornarsi sui progressi più recenti.

Oltre alla regolamentazione, i governi potrebbero auspicabilmente sostenere progetti scientifici con alto valore sociale anche se con basso ritorno economico o modesto interesse accademico in ambiti quali il cambiamento climatico, la biosicurezza o la preparazione ad affrontare le pandemie, giusto per citarne alcuni di interesse strategico per l'intera umanità. In

questi e altri settori la velocità e capacità su larga scala, offerte dalle simulazioni AI e dai laboratori di ricerca specializzati, possono diventare essenziali.

[1] https://deepmind.google/technologies/alphafold/.
[2] https://en.wikipedia.org/wiki/GitHub.
[3] https://www.treccani.it/enciclopedia/l-ultimo-teorema-di-fermat_(Enciclopedia-della-Matematica)/.

9
UNO SGUARDO AL FUTURO: A CHE PUNTO SIAMO?

In questo capitolo affrontiamo alcune questioni relative all'impatto che l'AI sta avendo sulla nostra società. Stime recenti del sito tedesco *Statista* prevedono che il mercato globale dell'AI raggiungerà 826,7 miliardi entro il 2030, con un tasso di crescita annuo medio del 28,46% tra il 2024 e il 2030. Il suo impatto sul mercato è quindi previsto essere significativo. Sebbene l'adozione dell'AI nel mondo produttivo europeo sia ancora relativamente bassa, il rapido progresso, la diminuzione dei costi e la crescente disponibilità di lavoratori con competenze in ambito AI indicano che le economie dell'OCSE potrebbero essere prossime a una transizione marcata. I dati disponibili suggeriscono che la percentuale di imprese che hanno adottato l'AI rimane in singola cifra percentuale, anche se le grandi imprese sono più propense a farlo (circa una su tre) (Lane *et al*. 2023). Il costo rappresenta l'elemento di maggior deterrenza per oltre la metà delle imprese dei settori finanziario e manifatturiero. Il secondo ostacolo è costituito dalla mancanza di competenze, anche se il trend è a diminuire; per fare un esempio, il costo per addestrare un sistema di classificazione delle immagini è diminuito di oltre il 63% (Zhang 2022) e, man mano che l'AI entra nel dominio pubblico, il tasso di riduzione di questi costi potrebbe ulteriormente accelerare. Applicazioni di AI generativa come ChatGPT stanno diventando sempre più

disponibili a un basso canone mensile o addirittura gratuitamente. Allo stesso tempo, la disponibilità di lavoratori con competenze adeguate è in crescita, anche se ancora lontana dall'essere sufficiente. Una ricerca dell'OCSE suggerisce che la forza lavoro dell'AI è più che triplicata tra il 2012 e il 2019 (Green e Lamby 2023). Essendo una tecnologia *general purpose*, l'AI può influenzare l'intera economia e ci si deve pertanto aspettare che tra non molto potrebbe permeare quasi tutti i settori e le occupazioni. Un elemento di attrazione per il management aziendale è la promessa di abbattimento dei costi e il miglioramento della qualità dei prodotti e dei servizi, dal momento che l'AI è una tecnologia di automazione. Dalla loro prospettiva, i lavoratori possono beneficiare di miglioramenti nella qualità del lavoro, avendo l'AI il potenziale per eliminare compiti pericolosi o noiosi e crearne altri più complessi e interessanti. Può aumentare il coinvolgimento dei lavoratori e dare loro maggiore autonomia.

Un processo trasformativo

L'AI ha fatto grandi progressi anche in aree come l'organizzazione delle informazioni, la memorizzazione, la velocità percettiva e il ragionamento deduttivo, tutte collegate a compiti cognitivi non di routine. Naturalmente non si possono ignorare i rischi. Accanto al miglioramento delle prestazioni dei lavoratori (cioè della produttività), le aziende sperano nella riduzione dei costi del personale. Non sorprende, quindi, che circa il 20% dei lavoratori nei settori finanziario e manifatturiero (in sette Paesi dell'OCSE) abbia dichiarato di essere molto o estremamente preoccupato per la perdita del proprio posto di lavoro nei prossimi dieci anni (Lane *et al*. 2023). Va osservato che i rischi dell'automazione non sono distribuiti equamente tra i gruppi socio-demografici, il che potrebbe danneggiare l'inclusività. Sebbene l'impatto dell'ultima ondata di AI generativa

non sia ancora del tutto chiaro, le stime iniziali dell'esposizione occupazionale all'AI che tengono conto delle capacità dei grandi modelli di linguaggio come ChatGPT giungono a conclusioni che siano principalmente le occupazioni ad alto reddito che richiedono un'istruzione o una formazione superiore alla media ad essere le più esposte all'ingresso dell'AI in azienda.

L'uso dell'AI sul posto di lavoro apre o amplifica una serie di questioni etiche, alcune delle quali possono influire negativamente sulla qualità del lavoro. Ad esempio, l'AI può cambiare il modo in cui il lavoro viene monitorato o gestito, il che può aumentare la percezione di equità, ma comporta rischi per la privacy e l'autonomia dei lavoratori nell'eseguire i compiti. Inoltre, esistono preoccupazioni legate alla trasparenza e alla comprensibilità dei sistemi, così come alla responsabilità. Sebbene molte di queste problematiche non siano nuove, l'AI ha il potenziale per amplificarle. Ad esempio, anche se gli esseri umani possono essere influenzati da pregiudizi nelle decisioni di assunzione, l'impatto negativo dell'AI potrebbe essere decisamente maggiore a causa del volume e della velocità delle decisioni che prende, con l'effetto indesiderato di sistematizzare e moltiplicare i pregiudizi stessi. Questi rischi tendono a essere maggiori per alcuni gruppi socio-demografici, che sono spesso già svantaggiati nel mercato del lavoro.

L'AI deve essere sicura e rispettare i diritti fondamentali come la privacy e l'equità, il diritto dei lavoratori di organizzarsi, la trasparenza e la comprensibilità. Ciò significa anche che deve essere chiaro chi è responsabile in caso di problemi. Un'azione proattiva e decisiva è importante non solo per proteggere i lavoratori, ma anche per promuovere l'innovazione e la diffusione dell'AI, poiché riduce l'incertezza.

L'impatto dell'AI sui compiti e sui posti di lavoro comporterà un cambiamento nelle competenze richieste. Da un lato, l'AI replicherà alcune competenze, come le abilità manuali e psicomotorie fini, e le competenze cognitive come la comprensione, la pianificazione e la consulenza. Dall'altro, le competenze necessarie per sviluppare e mantenere i sistemi di

AI, nonché quelle per adottare, utilizzare e interagire con le applicazioni di AI, diventeranno sempre più importanti. Aumenterà anche la domanda di competenze digitali di base, scienza dei dati e altre competenze cognitive e trasversali. Sebbene molte aziende che utilizzano l'AI affermino di fornire formazione per l'AI, la mancanza di competenze rimane una barriera significativa alla sua adozione. Le politiche pubbliche avranno quindi un ruolo importante, non solo per incentivare la formazione da parte dei datori di lavoro, ma anche perché una parte significativa della formazione per lo sviluppo e l'adozione dell'AI avviene nell'istruzione formale. L'AI stessa potrebbe offrire opportunità per migliorare la progettazione, il targeting e la fornitura della formazione, come vedremo nel seguito di questo capitolo.

Al pari e in misura ancor più ampia delle precedenti rivoluzioni (agricola, industriale, digitale), la rivoluzione indotta dall'intelligenza artificiale sta avendo e avrà un impatto *trasformativo* sul mondo del lavoro. Molte professioni, in particolare quelle più fisicamente impegnative o caratterizzate da compiti ripetitivi e routinari, scompariranno. Allo stesso tempo, nuove professioni emergeranno, mentre quasi tutte le altre subiranno profonde trasformazioni grazie all'integrazione dell'intelligenza artificiale nei processi decisionali e produttivi. Come è accaduto in passato, assisteremo a un cambiamento pervasivo verso un mondo sempre più automatizzato, in cui il supporto tecnologico sarà presente in ogni fase, da quella esecutiva a quella gestionale. Un cambiamento che riguarderà tanto il lavoro manuale quanto quello intellettuale, come nei settori manifatturiero, amministrativo o dei servizi al cliente, dove l'automazione e i chatbot stanno progressivamente riducendo la necessità di intervento umano.

Tuttavia, è altrettanto vero che, come in passato, emergeranno nuove professioni e settori, spesso in aree che oggi non riusciamo nemmeno a immaginare, soprattutto in quei settori legati allo sviluppo di tecnologie AI o alla loro governance etica. Il cambiamento avrà anche un impatto positivo su settori

ad alta creatività e innovazione. Professioni come quelle dei designer, degli insegnanti o dei medici, ad esempio, evolveranno grazie all'integrazione di strumenti di AI, che faciliteranno il lavoro quotidiano senza sostituire il valore umano insostituibile che esse apportano. Si assisterà inoltre all'emergere di ruoli ibridi, che combineranno competenze tradizionali e tecnologiche. Per esempio, gli insegnanti stanno già utilizzando l'AI per personalizzare i percorsi di apprendimento, gli agricoltori per gestire tecnologie avanzate come droni e sensori per ottimizzare i raccolti.

Ogni professione sarà toccata da questa trasformazione e dovrà adattarsi. La medicina, ad esempio, beneficerà della capacità dell'AI di analizzare dati complessi e diagnosticare malattie con maggiore precisione, ma il rapporto umano tra medico e paziente resterà centrale. Anche architetti e ingegneri vedranno il loro lavoro evolversi, con l'AI che accelererà la progettazione lasciando agli esperti la supervisione e l'ottimizzazione.

Nonostante l'ampia diffusione delle tecnologie intelligenti, il giudizio e le decisioni finali rimarranno prerogative dell'essere umano. Questo cambiamento richiederà inevitabilmente una formazione capillare, orientata a un uso consapevole e competente degli strumenti offerti dall'intelligenza artificiale. La conoscenza e l'esperienza degli specialisti di settore continueranno a essere indispensabili e insostituibili, perché solo attraverso il loro apporto sarà possibile sfruttare appieno il potenziale delle nuove tecnologie, mantenendo centrale il ruolo dell'uomo nel processo decisionale.

Quanto etica è l'AI

Un tema molto dibattuto riguarda l'etica dell'AI. In generale, l'etica si riferisce ai principi morali che regolano il comportamento di una persona o la conduzione di un'attività. Un

esempio concreto è il principio etico di trattare ogni individuo con rispetto. I filosofi hanno discusso di etica per secoli, formulando teorie come l'imperativo categorico di Kant, che invita ad agire come vorremmo che tutti gli altri agissero nei confronti di chiunque. L'etica dell'intelligenza artificiale si occupa di un tema cruciale: come sviluppatori, produttori e operatori umani dovrebbero agire per ridurre al minimo i rischi etici legati all'AI nella società. Questi rischi possono derivare da progettazioni inadeguate, applicazioni inappropriate o abusi tecnologici. L'ambito dell'etica dell'AI copre preoccupazioni immediate, come la privacy dei dati e i bias nei sistemi attuali, questioni a medio termine come l'impatto di AI e robotica sul mercato del lavoro, e preoccupazioni a lungo termine, come la possibilità che l'AI superi le capacità umane (superintelligenza).

Negli ultimi anni, l'etica dell'AI è passata dall'essere una questione puramente accademica a un argomento di dibattito pubblico e politico. L'uso diffuso di smartphone e applicazioni basate sull'AI, l'influenza crescente dell'AI in settori chiave come sanità, industria, trasporti, finanza e giustizia, e la prospettiva di una "corsa agli armamenti" tecnologica hanno spinto molteplici iniziative internazionali. Organizzazioni Non Governative, enti accademici, gruppi industriali e governi hanno avviato progetti e strategie, spesso accompagnati da ingenti investimenti, per affrontare le questioni etiche emergenti. Queste iniziative hanno prodotto vari insiemi di principi etici, nuovi standard tecnici e la creazione di organismi consultivi e politici dedicati. Robot e sistemi di AI si manifestano in forme diverse, ognuna delle quali solleva specifiche questioni etiche. Tra i temi principali figura l'impatto sociale, legato al mercato del lavoro, alle disuguaglianze economiche e al rischio di concentrare ricchezza e potere in poche mani. A questi vanno aggiunti anche i rischi per la privacy, i diritti umani e la dignità, oltre a come l'AI possa perpetuare bias sociali esistenti o minacciare la democrazia. Va poi considerato l'impatto psicologico, con interrogativi sui rapporti uomo-robot, rischi di dipendenza e inganno, e le implicazioni legali e morali di attribuire ai robot uno status simile

a quello umano. E ancora, l'impatto sul sistema finanziario, con rischi di manipolazione, collusione e la necessità di garantire responsabilità, e l'impatto legale, che include temi come la responsabilità in caso di uso criminale dell'AI o nel caso di incidenti, ad esempio con veicoli autonomi. Si dibatte su come affrontare richieste di risarcimento: negligenza o responsabilità del prodotto? Per non tralasciare l'impatto ambientale, con una crescente domanda di risorse naturali ed energia, e anche l'opportunità di migliorare la gestione dei rifiuti e delle risorse grazie all'AI. Non ultimo, va ricordato l'impatto sulla fiducia, cruciale per consentire all'AI di svolgere ruoli critici, come in ambito medico. La fiducia include equità, trasparenza, responsabilità e controllo.

In risposta a queste sfide, sono emerse numerose iniziative etiche miranti a classificare rischi e danni etici legati all'AI in diverse categorie, tra cui: diritti umani, danni emotivi, responsabilità, trasparenza, sicurezza, giustizia sociale, sostenibilità ambientale, uso informato e rischi esistenziali. Per quanto riguarda l'autonomia e la responsabilità, i processi decisionali dell'AI spesso avvengono senza intervento umano, sollevando interrogativi su quanto controllo gli esseri umani manterranno sui sistemi di AI man mano che questi diventano più sofisticati. Quando i sistemi di AI saranno utilizzati per prendere decisioni di vita o di morte, come nel caso delle armi autonome o nel settore sanitario, di chi sarà la responsabilità in caso di errori? Queste preoccupazioni sono aggravate dal fatto che l'AI può evolversi e adattarsi in modi non sempre prevedibili dai suoi creatori. L'AI potrebbe portare a una ridefinizione stessa dell'autonomia umana, poiché le persone si affidano sempre più ai sistemi di AI per prendere decisioni importanti nella loro vita personale e professionale.

L'interesse per l'etica dell'AI continua a crescere. Il numero di contributi accettati alla conferenza FAccT, un evento di rilievo nell'ambito dell'etica dell'AI, è più che raddoppiato dal 2021 e aumentato di 10 volte dal 2018. Anche nel 2022 si è registrato un numero record di contributi provenienti da attori industriali. La

verifica automatizzata dei fatti con il processamento del linguaggio naturale si rivela più complessa del previsto. Nonostante siano stati sviluppati diversi benchmark per la verifica automatizzata dei fatti, i ricercatori hanno scoperto che la maggioranza di questi dataset si basa su prove "trapelate" da rapporti di *fact-checking* che non esistevano al momento dell'emergere delle affermazioni. I *deepfake* politici stanno già influenzando le elezioni in diverse parti del mondo, con recenti ricerche che suggeriscono che i metodi di deepfake esistenti mostrano livelli di accuratezza variabili. Inoltre, nuovi progetti dimostrano quanto sia facile per l'AI creare e diffondere contenuti falsi.[1]

AI e società

La rivoluzione dell'AI richiederà una trasformazione sociale. Sarà essenziale investire in formazione e riqualificazione per preparare la forza lavoro ai nuovi ruoli e aggiornare le competenze necessarie. L'apprendimento permanente diventerà una pratica fondamentale per garantire che le persone possano adattarsi ai continui cambiamenti. Parallelamente, i sistemi di protezione sociale dovranno supportare chi perde il lavoro durante questa transizione, mentre una riflessione etica e normativa sarà cruciale per regolamentare l'uso dell'AI e assicurare che i suoi benefici siano distribuiti il più equamente possibile. In definitiva, serve trovare un equilibrio tra l'accoglimento dei benefici dell'AI e la mitigazione dei suoi rischi, assicurandosi che essa migliori anziché peggiorare la qualità della vita umana.

Se ci lasciamo guidare dall'*AI Stanford Report 2024* (Maslej *et al*. 2024), l'AI supera gli esseri umani in alcuni compiti, ma non in tutti (per fortuna!). Fra i primi, la classificazione delle immagini, il ragionamento visivo, la comprensione delle lingue. Resta indietro in compiti più complessi come la matematica a livello

competitivo, il ragionamento sul senso comune, la capacità di pianificazione. Mancano inoltre valutazioni robuste e standardizzate per la responsabilità dei large language models. I principali sviluppatori, tra cui OpenAI, Google e Anthropic (una startup di intelligenza artificiale con sede negli Stati Uniti, fondata nel 2021 da due italiani, Dario e Daniela Amodei), testano i loro modelli principalmente contro diversi benchmark di AI responsabile, con l'obiettivo di realizzare un confronto sistematico dei rischi e delle limitazioni dei principali modelli di AI.

Nel 2024 le organizzazioni hanno iniziato a utilizzare l'intelligenza artificiale generativa in modo significativo, ricavando valore tangibile per il business, consolidando il suo ruolo come tecnologia dirompente. Secondo il McKinsey Global Survey del 30 maggio 2024,[2] il 65% degli intervistati riferisce che la propria organizzazione utilizza regolarmente intelligenza artificiale generativa, quasi il doppio rispetto all'anno precedente. Le aspettative sul suo impatto restano elevate: tre quarti degli intervistati prevedono cambiamenti significativi o dirompenti nelle loro industrie nei prossimi anni.

L'adozione dell'AI ha visto un incremento globale, passando dal 50% al 72% delle organizzazioni intervistate. Mentre nel 2023 nessuna regione superava il 66% di adozione, nel 2024 più di due terzi degli intervistati in quasi tutte le regioni affermano che le loro aziende stanno utilizzando l'AI. I settori dei servizi professionali e dell'energia hanno registrato i maggiori incrementi. Inoltre, l'AI non si limita a singole funzioni aziendali: metà degli intervistati afferma che la propria organizzazione la utilizza in almeno due ambiti. Le funzioni di marketing e vendite e lo sviluppo di prodotti sono quelle dove l'adozione è più diffusa, grazie al suo potenziale di generare valore significativo. Il suo uso è aumentato anche nella vita personale degli intervistati, con un incremento marcato nell'Asia-Pacifico e in Cina, e con un maggiore utilizzo da parte di dirigenti di alto livello rispetto ai manager intermedi.

Gli investimenti in AI generativa stanno crescendo rapidamente. In molte industrie, le organizzazioni destinano oltre il 5% del loro budget digitale a questa tecnologia, sebbene l'AI tradizionale continui a ricevere una quota maggiore delle risorse. Le funzioni che beneficiano maggiormente degli investimenti in intelligenza artificiale generativa includono le risorse umane, dove si registrano riduzioni dei costi, e la gestione della supply chain. Anche l'AI tradizionale continua a dimostrarsi vantaggiosa, specialmente nelle operazioni di servizio e nel marketing, dove offre miglioramenti sia nei costi sia nei ricavi.

Nonostante i vantaggi, le aziende stanno affrontando i rischi associati all'AI generativa. Tra questi, l'inesattezza degli output, la privacy dei dati, i bias e le violazioni della proprietà intellettuale sono le preoccupazioni più comuni. Ad esempio, solo nel 2024, il "New York Times" ha investito circa 4,6 milioni di dollari in spese legali necessarie a sostenere la causa contro OpenAI. Il rischio di errori è il più frequentemente affrontato, con un crescente sforzo da parte delle organizzazioni per mitigarne l'impatto. Tuttavia, il rischio della sostituzione della forza lavoro è meno sentito rispetto all'anno precedente e le aziende stanno investendo meno per gestirlo.

Infine, alcune organizzazioni hanno già riscontrato conseguenze negative dall'uso dell'AI generativa, quali inesattezza, cybersicurezza e mancanza di spiegabilità. Solo una piccola parte delle aziende ha implementato pratiche di governance per un uso responsabile dell'AI. Vi è ancora molta strada da fare per scalare l'adozione dell'AI generativa in modo sicuro e sostenibile.

Quanto alla modalità di implementazione, si evidenziano tre archetipi principali per l'adozione di soluzioni di AI generativa: i *takers*, che utilizzano strumenti standard disponibili pubblicamente; gli *shapers*, che personalizzano questi strumenti con dati e sistemi proprietari; i *makers*, che sviluppano i propri modelli fondamentali da zero. I takers utilizzano soluzioni "pronte all'uso" sviluppate da altri, non partecipano al processo

di progettazione o sviluppo degli strumenti AI. Possono essere individui, aziende o organizzazioni che integrano l'AI nelle loro attività, per esempio un'azienda che utilizza un software AI per analizzare dati finanziari senza modificare il modello sottostante, o una persona che usa ChatGPT per generare contenuti o migliorare la produttività. I makers lavorano sulla progettazione, l'addestramento e l'implementazione di soluzioni AI, creano nuove tecnologie o personalizzano strumenti già esistenti. Possiedono competenze tecniche approfondite, come programmazione, data science e machine learning. Possono essere un team di ricercatori che sviluppa un nuovo modello di deep learning, o un'azienda che costruisce un chatbot AI personalizzato per i propri clienti. Gli shapers si concentrano sulla governance, l'etica e le politiche per garantire che l'AI sia sicura, equa e responsabile. Possono essere governi, organizzazioni non governative, università o aziende tecnologiche con un forte impegno etico, guidano il dibattito su come integrare l'AI nella società in modo sostenibile. Si pensi a un ente normativo che stabilisce regole per l'uso dell'AI in settori sensibili come la sanità o la finanza, oppure a un'organizzazione che promuove lo sviluppo di linee guida etiche per il machine learning. Questi tre ruoli non sono mutuamente esclusivi: una persona o organizzazione può essere sia maker sia shaper, o passare da un ruolo all'altro a seconda delle necessità. In particolare, nei settori dell'energia e dei materiali, della tecnologia, e dei media e delle telecomunicazioni le aziende sono più propense a personalizzazione significativamente e ottimizzare i modelli disponibili, o, ancora più frequentemente, a sviluppare modelli proprietari per rispondere a esigenze aziendali specifiche.

L'industria americana continua a dominare la ricerca di punta nell'AI. Nel 2023 ha prodotto 51 modelli di machine learning di rilievo, mentre il mondo accademico ha contribuito solo con 15. Inoltre, nel 2023 si sono registrati 21 modelli significativi risultanti da collaborazioni tra industria e accademia, un nuovo record. Gli investimenti in AI generativa esplodono. Nonostante

un calo complessivo degli investimenti privati in AI nel corso del 2023, i finanziamenti per l'AI generativa sono aumentati vertiginosamente, quasi ottuplicandosi dal 2022 per raggiungere i 25,2 miliardi di dollari. I principali attori nel campo dell'AI generativa, tra cui OpenAI, Anthropic, Hugging Face e Inflection, hanno riportato significativi round di finanziamenti. Gli Stati Uniti guidano su Cina, UE e Regno Unito come fonte principale di modelli AI di punta. Nel 2023, 61 modelli AI significativi sono stati originati da istituzioni statunitensi, superando di gran lunga i 21 dell'Unione Europea e i 15 della Cina.

Poiché le tecnologie avanzate possono rapidamente diventare infrastrutture fondamentali, è importante che la loro gestione non sia affidata esclusivamente alle aziende che le sviluppano, a logiche di mercato o alla formazione come unica soluzione. Aziende, mercati e formazione sono componenti cruciali, ma la regolamentazione è essenziale per proteggere coloro che subiscono effetti negativi dalle disuguaglianze create dalle nuove tecnologie e per evitare la concentrazione di potere economico e politico. Il numero di regolamenti sull'AI negli Stati Uniti è aumentato drasticamente. Nel 2023, ci sono stati 25 regolamenti relativi all'AI, rispetto a uno solo nel 2016. Solo nel 2023, il numero totale di regolamenti legati all'AI è cresciuto di oltre il 50%. Soltanto negli Stati Uniti, il numero di agenzie regolatorie che emettono regolamenti sull'AI è aumentato a 21 nel 2023 rispetto a 17 nel 2022, indicando una crescente preoccupazione per la regolamentazione dell'AI tra una gamma più ampia di enti regolatori americani: fra questi, il Dipartimento dell'Energia e l'Amministrazione per la Sicurezza e la Salute sul Lavoro. L'Unione Europea ha raggiunto un accordo sui termini dell'AI Act, una legislazione fondamentale entrata in vigore nel 2024. "L'Artificial Intelligence Act" o "AI Act" è il nome abbreviato del Regolamento UE 2024/1689 sull'uso dell'intelligenza artificiale: questo regolamento è entrato in vigore il 1° agosto 2024 e verrà applicato gradualmente nel corso dei successivi 24 mesi. Considerato il primo quadro giuridico al mondo dedicato all'intelligenza artificiale, l'AI Act è stato redatto con due obiettivi

principali: prevenire i potenziali rischi derivanti dai sistemi di intelligenza artificiale e dal loro utilizzo, evitando allo stesso tempo che una regolamentazione eccessiva freni lo sviluppo di tecnologie basate su AI all'interno dell'Unione Europea.

Nell'aprile 2023, l'Amministrazione del Cyberspazio della Cina ha pubblicato una bozza di misure per regolare i servizi di intelligenza artificiale generativa, invitando il pubblico a fornire commenti. Queste misure mirano a bilanciare lo sviluppo tecnologico con il controllo, garantendo che la tecnologia non destabilizzi l'ordine sociale. Uno dei requisiti principali prevede che le aziende che sviluppano tecnologie di AI generativa implementino salvaguardie per assicurare la conformità – non solo per evitare la creazione di contenuti non allineati con le politiche governative, ma anche tenendo conto delle implicazioni sulla proprietà intellettuale.

Se valutato da un punto di vista puramente economico, l'aspetto regolatorio, pur mitigando potenziali derive incontrollabili, rischia comunque di frenare lo sviluppo dei sistemi di AI. Un approccio regolatorio diverso, come quello (più morbido) adottato dagli Stati Uniti rispetto all'Europa, potrebbe accentuare ulteriormente il divario tecnologico fra i due continenti nell'ambito dell'intelligenza artificiale, già evidente oggi. Questa è una preoccupazione condivisa da molti attori europei, tra cui ricercatori, startup e grandi imprese, che desiderano vedere l'Europa prosperare anche nel settore della ricerca e della tecnologia avanzata legata all'AI. L'Europa, purtroppo, è già meno competitiva e innovativa rispetto ad altre regioni e rischia di rimanere ulteriormente indietro nell'era dell'AI a causa di decisioni regolatorie frammentate e talvolta incoerenti.

Senza regolamentazioni chiare e armonizzate, l'UE rischia di perdere due elementi chiave per l'innovazione nell'AI. Il primo è lo sviluppo di modelli "open", liberamente accessibili a tutti, che possono essere utilizzati, modificati e migliorati, moltiplicando i benefici e generando opportunità sociali ed economiche. Il secondo sono i modelli multimodali, che integrano testo,

immagini e voce e rappresentano il prossimo salto evolutivo nell'AI. La differenza tra i modelli solo testuali e quelli multimodali è simile alla differenza tra avere un solo senso e averne cinque. I modelli open, siano essi basati su testo o multimodali, hanno il potenziale di incrementare la produttività, promuovere la ricerca scientifica e aggiungere centinaia di miliardi di euro all'economia europea. Istituzioni pubbliche e ricercatori stanno già utilizzando questi modelli per accelerare la ricerca medica e preservare il patrimonio culturale, mentre imprese consolidate e startup accedono a strumenti che non avrebbero mai potuto sviluppare o permettersi autonomamente. Senza questi strumenti, lo sviluppo dell'AI si sposterà probabilmente altrove, lasciando l'Europa indietro e privando i suoi cittadini dei progressi tecnologici già presenti in Stati Uniti, Cina e India.

Alcune ricerche mostrano che l'AI generativa potrebbe far crescere il PIL globale del 10% nei prossimi dieci anni, e i cittadini europei non dovrebbero essere esclusi da questa opportunità di crescita. La competitività dell'Europa nel campo dell'AI e la sua capacità di sfruttare i vantaggi dei modelli open source dipendono da un quadro normativo chiaro e condiviso. Se aziende e istituzioni devono investire miliardi di euro per sviluppare AI generativa per i cittadini europei, necessitano di regole coerenti e prevedibili che consentano l'uso dei dati europei per l'addestramento dei modelli di AI. Tuttavia, le recenti decisioni regolatorie dell'Unione Europea sono state frammentate e imprevedibili, mentre gli interventi delle Autorità per la protezione dei dati hanno creato una grande incertezza su quali tipi di dati possano essere utilizzati per addestrare i modelli di AI. Se questa frammentazione persiste, la prossima generazione di modelli di AI open source e i relativi prodotti e servizi non rispecchieranno né comprenderanno appieno la conoscenza, la cultura o le lingue europee. L'Europa rischia quindi di perdere il passo con altre innovazioni, mentre altri paesi costruiranno su tecnologie alle quali i cittadini europei non avranno accesso. L'Europa si trova ora di fronte a una scelta

cruciale che determinerà il suo futuro nei prossimi decenni. Può riaffermare il principio di armonizzazione, come stabilito in regolamenti quali il GDPR (General Data Protection Regulation introdotto nel 2016), permettendo all'innovazione nell'AI di svilupparsi con la stessa rapidità e scala riscontrate in altre regioni. Oppure, può continuare a ostacolare il progresso, tradendo le ambizioni del mercato unico e osservare il resto del mondo avanzare con tecnologie dalle quali gli europei saranno esclusi. L'Europa non può permettersi di perdere i benefici delle tecnologie open di AI, sviluppate responsabilmente, che accelereranno la crescita economica e favoriranno nuove opportunità nella ricerca scientifica.

Sebbene la regolamentazione sia importante, e vi sia un ampio consenso sulla necessità di linee guida etiche per l'AI, il regolamento europeo sull'AI appare focalizzato su restrizioni, obblighi di conformità e sanzioni, mentre gli Stati Uniti promuovono lo sviluppo dell'AI con regolamentazioni più leggere e maggiori investimenti. Aziende come OpenAI, Perplexity, Anthropic e xAI non nascono in Europa, ma in luoghi che privilegiano l'innovazione piuttosto che soffocarla con troppa burocrazia. Stati Uniti e Cina stanno investendo nel futuro dell'AI, e finora questo approccio ha dato risultati.

Naturalmente, è anche importante riconoscere che, parallelamente, cresce la consapevolezza del potenziale impatto dell'AI, così come aumentano le preoccupazioni. Un sondaggio IPSOS ha rivelato che nel corso del 2023 la percentuale di persone che ritiene che l'AI influenzerà drasticamente le loro vite nei prossimi tre/cinque anni è salita dal 60% al 66%. Inoltre, il 52% esprime preoccupazioni riguardo ai prodotti e servizi basati sull'AI, un aumento di 13 punti percentuali rispetto al 2022. Negli Stati Uniti, i dati del Pew Research Center indicano che il 52% degli americani si dichiara più preoccupato che entusiasta riguardo all'AI, rispetto al 37% dell'anno precedente.

Negli ultimi anni, diverse istituzioni hanno creato grandi modelli di AI basati su dati proprietari. Nonostante questi modelli presentino ancora problematiche legate a tossicità e

bias, nuove evidenze suggeriscono che tali problemi possano essere in parte mitigati con il tuning successivo all'addestramento di modelli più ampi. Il numero di incidenti riguardanti l'uso improprio dell'AI è in rapido aumento. Secondo il database AIAAIC,[3] che traccia gli incidenti legati all'uso etico dell'AI, il numero di incidenti e controversie legati all'AI è aumentato di 26 volte dal 2012. Alcuni incidenti notevoli nel 2022 includevano un video *deepfake* del presidente ucraino Volodymyr Zelensky che si arrendeva e le prigioni statunitensi che utilizzavano tecnologia di monitoraggio delle chiamate sui detenuti. Questo aumento testimonia sia un maggior utilizzo delle tecnologie AI sia una crescente consapevolezza delle possibilità di abuso.

I modelli generativi sono diventati parte della cultura contemporanea nel 2022, portando con sé problemi etici. Questi modelli sono potenti ma presentano sfide significative. I generatori di testo e di immagini sono spesso caratterizzati da bias di genere, e i chatbot come ChatGPT possono essere manipolati per scopi nefasti. Analisi approfondite dei modelli di linguaggio suggeriscono che, sebbene vi sia una chiara correlazione tra prestazioni ed equità, l'equità e il bias possono essere in contrasto: i modelli di linguaggio che ottengono risultati migliori su determinati benchmark di equità tendono ad avere un bias di genere più marcato.

I modelli di intelligenza artificiale generativa possono produrre contenuti di alta qualità che potrebbero essere usati per ingannare. Alcune normative, come l'ordine esecutivo statunitense sulla sicurezza dell'AI, propongono l'uso di *watermark*, ovvero marcature digitali invisibili, per etichettare i contenuti generati dall'AI. Tuttavia, i watermark possono essere facilmente manipolati e non sempre influenzano il comportamento degli utenti. È quindi importante esplorare tecniche alternative per garantire la provenienza e l'autenticità dei contenuti in modo verificabile.

Quando l'AI va in guerra

La penetrazione dell'AI nella società moderna ha comportato anche un ruolo crescente nell'ambito bellico, con una militarizzazione dell'AI che, entro certi limiti, si può paragonare alla corsa agli armamenti nucleari durante la Guerra Fredda. Nel dibattito su come l'AI influenzerà le strategie e le decisioni militari, emerge una domanda fondamentale: chi prende decisioni migliori, gli esseri umani o le macchine? I sostenitori di un impiego più ampio dell'AI ritengono che essa possa ridurre gli errori umani e limitare i danni ai civili grazie a una maggiore precisione e aderenza alle leggi internazionali. I critici, invece, evidenziano come anche le decisioni dell'AI possano essere fallibili, con rischi potenzialmente sproporzionati. Inoltre, spesso si ignora che non sia né possibile né auspicabile che le macchine replichino tutte le sfumature del processo decisionale umano.

Sullo scacchiere internazionale, è in atto una corsa globale agli armamenti per definire i migliori utilizzi dell'AI in ambito militare, corsa purtroppo accelerata dai conflitti a Gaza e in Ucraina. La guerra di Gaza ha mostrato come l'impiego dell'AI nel targeting tattico possa influenzare le strategie militari, inducendo specifiche inclinazioni decisionali. Nelle fasi iniziali del conflitto, *Habsora*, un software di apprendimento automatico, costruito sulla base di centinaia di algoritmi predittivi, è stato in grado di determinare le probabili coordinate di tunnel o silos di razzi costruiti da Hamas nella striscia di Gaza. *Habsora* permette – come spiegato dall'allora leader dell'IDF[4] Aviv Kohavi – di generare 100 obiettivi al giorno, contro i 50 inseriti nella banca dati dell'esercito israeliano, tramite lavoro condotto da analisti umani, nel giro di un intero anno. Un altro sistema di AI, chiamato *Lavender*, ha identificato migliaia di individui legati ad Hamas, passando rapidamente da un ruolo di intelligence a lungo termine all'identificazione immediata di bersagli specifici. Lavender ha generato un modello semplificato e digitale del campo di battaglia, permettendo un targeting più rapido e una

frequenza di attacchi elevata rispetto ai conflitti passati. Sebbene gli analisti umani esaminassero le raccomandazioni di Lavender prima di approvare gli attacchi, secondo alcune opinioni la fiducia riposta nel sistema è cresciuta rapidamente, inducendo fenomeni di bias di automazione e di azione e portando gli esseri umani a delegare parte del processo decisionale alla macchina.[5]

Un esempio di collaborazione efficace tra uomo e AI è rappresentato dal sistema ucraino GIS Arta, che adotta un approccio dal basso verso l'alto nella selezione dei bersagli, offrendo una visione complessiva del campo di battaglia senza indicare opacamente gli obiettivi. Descritto come una sorta di "Uber per l'artiglieria", GIS Arta consente agli operatori umani di valutare il contesto e di decidere cosa attaccare.[6] Questo ha reso il campo di battaglia quasi trasparente, sia nelle immediate vicinanze che in profondità. Oggi, la strategia si fonda sulla capacità di rilevare le forze nemiche e confondere i loro sistemi di sorveglianza per evitare di essere colpiti. La linea del fronte, estesa per centinaia di chilometri su entrambi i lati, è diventata una zona estremamente pericolosa, dove nessuna delle due parti riesce a ottenere un reale vantaggio. Queste strategie potrebbero diventare sempre più attuabili man mano che aumenta il numero di sistemi d'arma abilitati all'AI.

Il primo ministro giapponese Fumio Kishida ha affermato: "L'Ucraina di oggi potrebbe essere l'Asia orientale di domani", mentre il presidente russo Vladimir Putin ha dichiarato che la nazione che dominerà l'intelligenza artificiale "dominerà il mondo".

Una questione cruciale è se i sistemi di AI influenzeranno le decisioni nei contesti di combattimento. Poiché questi sistemi possono distanziare i soldati dal campo di battaglia, rendendo più semplice la decisione di usare forza letale, è probabile che abbiano un impatto significativo. Storicamente, la guerra richiede ai soldati di superare una naturale avversione per l'uccisione. Tecnologie come i droni e i sistemi d'arma autonomi

contribuiscono a questo distanziamento, riducendo l'impatto sensoriale e psicologico dell'uccisione.

I progressi dell'AI applicati alla guerra moderna influenzeranno profondamente le relazioni tra grandi potenze come Stati Uniti, Cina e Russia, così come il settore tecnologico privato.

Un report del 2019 di Jane's prevedeva che nei prossimi dieci anni sarebbero stati acquistati oltre 80.000 droni di sorveglianza e circa 2000 droni armati a livello globale. Attualmente, Stati Uniti, Regno Unito e Israele sono i principali utilizzatori di droni, con arsenali in continua espansione. Gli Stati Uniti e il Regno Unito impiegano da oltre un decennio droni armati, come i Predator e i Reaper prodotti dalla General Atomics in California. Il Pentagono stima che entro il 2035 il 70% della forza aerea statunitense sarà costituito da velivoli a pilotaggio remoto. Nel frattempo, Israele ha sviluppato propri droni armati, utilizzati a Gaza sia per la sorveglianza che per attacchi con esplosivi.

Anche l'Arabia Saudita, nuovo attore nel mercato dei droni, ha destinato 69 miliardi di dollari, pari al 23% del suo bilancio nazionale, alla difesa nel 2023 e ha in programma un fondo di 40 miliardi di dollari per investimenti nell'AI, una cifra che la renderebbe il principale investitore mondiale in questo settore.

L'integrazione crescente dell'AI nei droni e in altri sistemi d'arma autonomi solleva il rischio concreto di conflitti gestiti senza un reale controllo umano. Il ricorso alla violenza potrebbe infatti essere determinato da decisioni di macchine incapaci di comprendere le complessità morali e di agire eticamente nel contesto bellico. Come il diritto internazionale si adatterà a questi sviluppi rimane difficile da prevedere, ma il quadro normativo attuale presenta lacune evidenti in termini di chiarezza e di capacità di gestione futura (Klaus 2024; Álvarez 2024; Zhou e Greipl 2024; Bode e Bhila 2024).

Se l'AI sale in cattedra

Un altro problema cruciale è quello legato alla *formazione* in AI e grazie all'AI. Ovunque si registra un gap significativo fra richiesta di offerta e domanda di tecnici, informatici e ingegneri con competenze in AI. Mentre il numero di nuovi laureati in informatica negli Stati Uniti e in Canada è costantemente aumentato per più di un decennio, il numero di studenti che optano per un'istruzione superiore in informatica è rimasto stabile. Dal 2018, il numero di laureati magistrali e dottorandi in informatica è leggermente diminuito. Nel 2011 una percentuale simile di nuovi dottorandi in AI ha scelto lavori nell'industria (40,9%) e nell'accademia (41,6%). Tuttavia, entro il 2022, una proporzione significativamente maggiore (70,7%) è entrata nell'industria dopo la laurea rispetto a quelli che hanno scelto l'accademia (20,0%). Solo nel corso del 2023, la quota di dottorandi in AI diretti verso l'industria è aumentata di 5,3 punti percentuali, indicando un'intensificazione della fuga di cervelli dalle università verso l'industria.

Nel 2019, il 13% dei nuovi docenti di AI negli Stati Uniti e in Canada proveniva dall'industria. Nel 2021 questa cifra è diminuita all'11% e nel 2022 è ulteriormente scesa al 7%. Questa tendenza indica una migrazione progressivamente minore di talenti di alto livello in AI dall'industria all'accademia. Il numero di programmi di laurea postsecondaria in lingua inglese legati all'AI è triplicato dal 2017, mostrando un costante aumento annuale negli ultimi cinque anni. Le università di tutto il mondo stanno offrendo più programmi di laurea focalizzati sull'AI. Il Regno Unito e la Germania guidano l'Europa nella produzione del maggior numero di nuovi laureati in informatica, ingegneria informatica e informazione a livello di laurea triennale, magistrale e dottorato. In proporzione alla popolazione, la Finlandia è leader nella produzione di laureati sia triennali che di dottorato, mentre l'Irlanda è leader nella produzione di laureati magistrali.

Misurare l'impatto che l'AI generativa può avere sull'insegnamento non è un compito facile, ma possiamo delineare alcuni punti chiave basati su quanto è accaduto finora

e su ciò che ci si potrebbe aspettare in futuro. Le applicazioni e le implicazioni dell'AI generativa si stanno espandendo rapidamente in vari settori dell'istruzione, dal livello primario all'universitario. Nella scuola primaria e secondaria, l'AI generativa può creare materiali didattici personalizzati per i singoli studenti, tenendo conto delle loro capacità e dei loro progressi. Strumenti come chatbot educativi e assistenti virtuali basati su AI (come i modelli GPT) sono già stati utilizzati per rispondere a domande specifiche degli studenti, offrire spiegazioni alternative o generare quiz su misura. L'AI generativa può aiutare gli studenti a migliorare le competenze di lettura e scrittura, suggerendo miglioramenti nei testi, fornendo correzioni o spiegazioni su concetti complessi in un linguaggio accessibile. All'università, gli strumenti di AI generativa vengono impiegati per la creazione di lecture notes, riassunti, o raccolte di esami. Gli studenti possono ottenere feedback immediato su scritti, presentazioni e progetti attraverso piattaforme che utilizzano modelli linguistici generativi, riducendo il carico sui docenti per attività di correzione meccanica. Gli studenti utilizzano strumenti come ChatGPT per approfondire argomenti complessi, ottenere spiegazioni dettagliate su argomenti che trovano difficili o chiedere chiarimenti su teorie scientifiche, matematiche o storiche. Questo può velocizzare l'apprendimento, ma solleva dubbi sull'eccessiva dipendenza dagli strumenti automatizzati.

In prospettiva, l'AI generativa potrebbe trasformare il *sistema educativo* rendendolo altamente personalizzato. Ogni studente potrebbe ricevere un curriculum su misura, progettato per rispondere ai propri bisogni, ritmi di apprendimento e stili di studio, con contenuti generati in modo adattivo. Questo potrebbe essere particolarmente utile per chi ha difficoltà di apprendimento o per gli studenti più dotati che hanno bisogno di ulteriori stimoli. Gli insegnanti stessi dovranno adattarsi a nuove tecnologie e diventare competenti nell'uso di strumenti basati sull'AI. Potrebbero anche sfruttare l'AI generativa per creare piani di lezione, valutazioni o materiali didattici più

dinamici e coinvolgenti. Le piattaforme di formazione per insegnanti potrebbero essere potenziate da AI per fornire aggiornamenti continui sulle migliori pratiche pedagogiche.

Uno degli sviluppi più importanti potrebbe riguardare il dibattito etico su come integrare l'AI in modo responsabile. Le scuole e le università dovranno trovare un equilibrio tra l'uso dell'AI per migliorare l'apprendimento e il rischio di sostituire processi educativi fondamentali. Sarà cruciale insegnare agli studenti come usare criticamente l'AI e non dipendere da essa per ogni aspetto dell'apprendimento.

L'AI potrebbe portare alla nascita di modelli di insegnamento ibrido, dove l'interazione tra docenti e AI diventa simbiotica. Il docente potrebbe fungere più da mentore e facilitatore, mentre l'AI generativa potrebbe fornire il supporto pratico per attività di apprendimento e sperimentazione, simulazioni, o addirittura la valutazione del progresso dello studente.

L'AI generativa può facilitare la diffusione di modelli di insegnamento multimodale, che integrano diversi tipi di input, come testo, immagini, video, audio e interazioni simulate. Questo approccio potrebbe arricchire l'esperienza didattica, rendendola più interattiva e coinvolgente per gli studenti, specie per quelli che apprendono meglio attraverso canali non tradizionali.

Naturalmente, l'uso dell'AI generativa solleva anche preoccupazioni sull'*autenticità* del lavoro degli studenti. Come si può distinguere tra ciò che è genuinamente frutto della creatività e della comprensione di uno studente e ciò che è stato generato dall'AI? Questo sarà un dilemma crescente, soprattutto nei livelli di istruzione più avanzati, come nelle scuole superiori e all'università.

L'eccessiva dipendenza da strumenti automatizzati potrebbe portare a una perdita di competenze critiche. Ad esempio, se gli studenti utilizzano modelli di AI per scrivere saggi o risolvere problemi matematici, potrebbero non sviluppare pienamente capacità di pensiero critico o di problem solving.

L’AI potrebbe esacerbare le disparità educative se l’accesso a strumenti avanzati di AI generativa non fosse equamente distribuito. Le scuole o gli studenti con meno risorse potrebbero restare indietro, aumentando il divario educativo.

L’AI generativa ha già iniziato a trasformare l’istruzione e, in prospettiva, potrebbe rivoluzionarla ulteriormente. Se da un lato offre enormi opportunità per personalizzare l’apprendimento e alleviare il carico sui docenti, dall’altro solleva questioni etiche e metodologiche complesse. Per massimizzare il suo impatto positivo, sarà essenziale affrontare questi dilemmi con un approccio regolamentato, etico e inclusivo, che garantisca che l’AI serva a potenziare – e non a sostituire – l’educazione umana.

Un settore certamente esposto ai cambiamenti indotti dai sistemi dell’intelligenza artificiale generativa è quello dell’*editoria*, in senso lato. Recentemente, Microsoft e TikTok hanno avviato progetti editoriali con la creazione di marchi per la pubblicazione di libri: *8080 Books* di Microsoft,[7] ispirato al microprocessore Intel 8080, ha pubblicato il libro *No Prize for Pessimism* di Sam Schillace, mentre *8th Note Press* di TikTok si focalizzerà su romanzi di successo tra i giovani adulti. Microsoft e TikTok condividono un interesse per i sistemi di intelligenza artificiale. Microsoft ha investito circa 13 miliardi di dollari, collaborando con OpenAI, mentre TikTok sviluppa i propri sistemi di AI. 8080 Books mira a rendere più veloce il processo di pubblicazione, tradizionalmente lento rispetto ad altri settori. Tuttavia, la loro espansione nell’editoria ha suscitato dibattiti, poiché le AI generative hanno utilizzato contenuti protetti da copyright, spesso senza autorizzazione, per allenare i loro modelli. Casi legali contro le aziende AI, come le denunce di Sarah Silverman contro OpenAI e Meta e del New York Times contro Microsoft, riflettono queste preoccupazioni. Anche se non è chiaro se l’uso di dati web violi effettivamente il diritto d’autore, il settore deve trovare nuovi contenuti per progredire. Startup come *Created by Humans*[8] propongono modelli di licenze per autori che vogliono vendere le loro opere alle aziende AI. Walter

Isaacson, biografo e consigliere della startup, considera questa fase una rivoluzione, simile all'invenzione dei motori di ricerca. Al contrario, alcuni autori come Hari Kunzru chiedono garanzie per impedire che i loro lavori vengano usati per le AI.

Anche l'utilizzo delle AI per creare nuovi contenuti preoccupa. Ad esempio, *Spines*,[9] una startup che usa AI per produrre libri, prevede di pubblicarne 8000 nel corso del 2025 tramite un modello di self-publishing a pagamento. Questa diffusione ha già avuto un impatto su piattaforme come Amazon, dove libri generati da AI (spesso di bassa qualità) si accompagnano a recensioni artificiali. Tuttavia, molte case editrici stanno adottando le AI per compiti specifici, come redigere schede libro o fare traduzioni. Alcuni editori hanno già iniziato a sviluppare versioni personalizzate e chiuse di ChatGPT per tutelare la privacy e il copyright, utilizzandole poi per tradurre materiali stranieri. La supervisione da parte dell'editore è tuttavia essenziale, in questa fase, al fine di garantire che l'uso dei sistemi di AI non comprometta la qualità del lavoro umano.

La democrazia al tempo dell'AI

La relazione tra intelligenza artificiale e *democrazia* sta emergendo come una delle questioni più complesse e delicate del nostro tempo, specialmente in un contesto di crescente regolamentazione e tentativi di bilanciamento tra libertà e dignità. La sfida principale consiste nel conciliare la rapida innovazione tecnologica con la necessità di salvaguardare i valori democratici fondamentali.

Uno dei campi centrali di questo dibattito è quello della *disinformazione*. Con l'adozione massiccia di strumenti di AI in grado di produrre contenuti falsi (come i *deepfake*), la diffusione di notizie false diventa sempre più sofisticata, minando la capacità dei cittadini di distinguere la realtà dalla finzione. La manipolazione delle informazioni, in un contesto in cui le

piattaforme digitali e gli algoritmi determinano cosa vediamo, solleva dubbi sull'effettiva libertà del cosiddetto "mercato delle idee". Laddove una parte significativa del dibattito pubblico si svolge su piattaforme private, controllate da algoritmi opachi, il rischio è che la pluralità delle voci venga compromessa.

Le differenze tra Stati Uniti ed Europa illustrano bene questa tensione: mentre negli Stati Uniti il Primo Emendamento enfatizza la libertà di espressione quasi in senso assoluto, in Europa prevale una visione in cui la dignità della persona e la protezione dei dati sono fondamentali. Questo implica che, in Europa, la regolamentazione sia spesso più rigorosa, come dimostra il Regolamento Generale sulla Protezione dei Dati (GDPR), mentre negli Stati Uniti si tende a privilegiare un approccio più permissivo, lasciando maggior spazio alla *self-regulation* delle aziende tecnologiche. Tuttavia, il caso più recente di co-regolamentazione negli Stati Uniti, citato nel testo, rappresenta una svolta significativa, segnalando la consapevolezza che il solo meccanismo del mercato non sia sufficiente a garantire il corretto funzionamento della democrazia.

Un altro fronte di preoccupazione è la *discriminazione algoritmica*, che riguarda l'utilizzo dell'AI in contesti decisionali come il lavoro, il credito, l'istruzione e persino la giustizia. Gli algoritmi, se non progettati con attenzione, possono amplificare pregiudizi preesistenti, con il rischio di rafforzare le disparità sociali. Questo crea un problema per i principi di uguaglianza e giustizia su cui si fondano le democrazie liberali. Ad esempio, un algoritmo utilizzato per assumere personale potrebbe penalizzare inconsapevolmente determinate minoranze, se addestrato su dati storicamente distorti. Questo solleva questioni di responsabilità e trasparenza: di chi è la responsabilità quando un algoritmo prende una decisione discriminatoria? E come si può rendere trasparente un processo decisionale basato su calcoli complessi che pochi comprendono davvero?

Un'indagine globale sull'AI responsabile evidenzia che le principali preoccupazioni delle aziende riguardo l'AI includono *privacy*, *sicurezza dei dati* e *affidabilità*. L'indagine mostra che le organizzazioni stanno iniziando a prendere misure per mitigare questi rischi. Diversi ricercatori hanno dimostrato che gli output generati dai large language models più popolari possono contenere materiale protetto da copyright, come estratti del "New York Times", o scene di film. Se tali uscite costituiscano violazioni del copyright sta diventando una questione legale di grande importanza. Il nuovo *Foundation Model Transparency Index*[10] mostra che gli sviluppatori di AI mancano di trasparenza, soprattutto riguardo alla divulgazione dei dati di addestramento e delle metodologie utilizzate per estrarne informazioni. Questa mancanza di apertura ostacola gli sforzi per comprendere ulteriormente la robustezza e la sicurezza dei sistemi di AI. Il rispetto e la valorizzazione della *proprietà intellettuale* è in effetti un altro grande tema che viene spesso giustamente sollevato nel contesto dei sistemi di AI. Nell'Artificial Intelligence Act della Commissione Europea non si fa menzione dell'impatto sulla proprietà intellettuale, creando incertezza su come verranno tutelati i diritti dei titolari. Secondo le normative vigenti, l'AI potrebbe rappresentare un rischio per i diritti di proprietà intellettuale, sollevando problemi di attribuzione, proprietà e possibili violazioni del copyright. I contenuti generati dall'AI dipendono sia dagli input degli utenti che da grandi set di dati, spesso raccolti online. Dato che i processi di AI utilizzano un ampio archivio di testi umani, è probabile che i contenuti prodotti contengano elementi già presenti in opere letterarie o artistiche esistenti. Tecnicamente, i proprietari originali potrebbero reclamare una violazione di copyright contro i creatori dell'AI o i suoi sviluppatori, ma al momento è poco chiaro come verrebbero trattati questi casi e chi sarebbe ritenuto responsabile. È probabile che gli sviluppatori di sistemi di AI debbano dimostrare di aver adottato adeguate misure per evitare violazioni della proprietà intellettuale. Questo potrebbe

includere contratti con terzi, come artisti, fornitori di librerie d'immagini e database, per i dati utilizzati nella generazione dei contenuti. Anche gli utenti dovrebbero verificare che i contenuti generati non violino i diritti di terzi, soprattutto se utilizzati in ambito commerciale.

Poiché le leggi sulla proprietà intellettuale faticano a tenere il passo con l'innovazione dell'AI, i brand devono informarsi sui rischi associati ai contenuti generati dall'AI. Se un marchio considera l'uso di contenuti AI per marketing o design di prodotto, dovrebbe essere consapevole del rischio di violare la proprietà intellettuale di terzi, come marchi o immagini protette. Disporre dei diritti legali sull'uso di questi contenuti sarà fondamentale.

Non solo le aziende che utilizzano strumenti di AI devono essere consapevoli dei rischi: con la diffusione degli strumenti di AI, ogni brand è esposto al rischio che la proprietà intellettuale venga violata, sia accidentalmente che intenzionalmente. Per proteggersi, i brand dovrebbero adottare una strategia di protezione solida. Ad esempio nel caso dei brevetti, non è ancora chiaro cosa sia realmente brevettabile per le soluzioni di AI.

Al tempo stesso, se ben impiegata, l'AI stessa può supportare i brand nei loro programmi di tutela del marchio. Le tecnologie AI sono utili per monitorare i beni di proprietà intellettuale e identificare eventuali violazioni, anche se provenienti da altre piattaforme AI. Questo tipo di tecnologia è un ottimo supporto, ma non può ancora sostituire completamente i consulenti umani.

Come abbiamo più volte sottolineato, i dati sono cruciali per addestrare gli algoritmi di machine learning. Negli ultimi vent'anni, i sistemi che raccolgono, elaborano e archiviano enormi quantità di dati si sono diffusi rapidamente, assumendo un ruolo fondamentale in molti ambiti della vita quotidiana e aziendale. Questi includono smartphone e dispositivi personali, sistemi di automazione per la casa e l'industria, veicoli autonomi, solo per citarne alcuni. Questo sviluppo accelerato ha sollevato numerose questioni legate alla sicurezza, alla sorveglianza, alla

privacy, alla giustizia, alla responsabilità e ad altri aspetti etici e politici. La regolamentazione della raccolta e conservazione dei dati rappresenta una sfida non solo normativa ma anche etica. Una volta raccolti i dati, due aspetti critici sul loro uso richiedono attenzione, al fine di prevenire la fuga di dati involontaria e garantire la qualità dei dati stessi. Con l'espansione dei sistemi di dati in tutti i settori aumenta anche la superficie esposta a potenziali minacce. Errori umani, ad esempio, sono stati spesso causa di attacchi *ransomware* contro infrastrutture critiche come gli ospedali. È dunque fondamentale promuovere una "alfabetizzazione" in tema di sicurezza e privacy.

Alla luce di queste sfide, emerge l'esigenza di un nuovo equilibrio regolatorio. Il modello di co-regolamentazione che si sta affermando in Europa, con iniziative come il Codice dell'Unione contro la Disinformazione e il Regolamento sulla trasparenza della pubblicità politica, rappresenta un tentativo di bilanciare la libertà di espressione con la protezione della verità e della correttezza del dibattito pubblico.[11] Tuttavia, è chiaro che la frammentazione culturale e giuridica tra le due sponde dell'Atlantico rende difficile un approccio armonizzato a livello globale. La riflessione centrale è che, se da un lato l'AI può arricchire la democrazia rendendo più accessibili le informazioni e facilitando la partecipazione pubblica, dall'altro, senza adeguate garanzie normative e un sistema di controlli trasparente e responsabile, rischia di destabilizzare le basi stesse del sistema democratico, esacerbando problemi di disinformazione, disuguaglianza e manipolazione del consenso. Le democrazie moderne hanno di fronte una grande sfida: trovare modi per integrare l'AI in un quadro che salvaguardi sia la libertà di espressione che la dignità delle persone, promuovendo una regolamentazione che sappia prevenire abusi senza soffocare l'innovazione.

[1] https://futurist.com/2023/10/29/countercloud-ai-powered-disinformation-experiment/.

[2] https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai.
[3] https://www.aiaaic.org/.
[4] Israeli Defence Forces.
[5] https://www.972mag.com/lavender-ai-israeli-army-gaza/.
[6] https://en.wikipedia.org/wiki/GIS_Arta.
[7] https://unlocked.microsoft.com/8080-books/.
[8] https://www.createdbyhumans.ai/.
[9] https://spines.com/.
[10] https://crfm.stanford.edu/fmti/May-2024/index.html.
[11] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32024R0900.

10

INTELLIGENZA ARTIFICIALE E INTELLIGENZA NATURALE

Se da un lato l'intelligenza artificiale ha capacità uniche, dall'altro i suoi limiti rispetto alla complessità e alla profondità del pensiero umano sono piuttosto palesi. Come possiamo evidenziare i principali elementi di differenziazione nella performance fra AI e intelligenza umana?

Meglio se artificiale

L'AI supera di gran lunga l'intelligenza umana in termini di rapidità di elaborazione dei dati ed esecuzione di calcoli matematici. Può analizzare grandi quantità di dati in frazioni di secondo, individuando pattern che sarebbero invisibili agli esseri umani, o richiederebbero molto più tempo per essere scoperti. L'AI può simultaneamente monitorare e gestire enormi volumi di informazioni e processi in più ambiti. Ad esempio, nei sistemi di trasporto, può ottimizzare il traffico in una città, mentre in ambito aziendale può automatizzare migliaia di decisioni e transazioni quotidiane. Sistemi di sorveglianza e automazione nelle smart cities utilizzano AI per analizzare flussi di dati provenienti da telecamere, sensori di traffico, e dispositivi dell'Internet of Things (IoT), in modo più rapido e accurato di quanto possano fare operatori umani.

In ambito farmaceutico l’AI consente di scoprire nuove molecole, in ambito medico può analizzare immagini radiologiche con accuratezza e rapidità spesso superiore a quella degli esperti di dominio, consentendo a questi ultimi di migliorare le loro diagnosi.

L’AI non si affatica né soffre di limitazioni biologiche come il sonno, l’attenzione o il decadimento cognitivo. Questo la rende ideale per compiti ripetitivi e complessi che richiedono attenzione costante, come il monitoraggio di sistemi industriali o la sorveglianza di infrastrutture critiche. Ad esempio, l’AI consente di monitorare i flussi energetici nelle centrali elettriche, riducendo al minimo il rischio di errori umani legati alla fatica o alla distrazione. Essendo programmata per prendere decisioni basate esclusivamente sui dati, l’AI non è generalmente influenzata da pregiudizi emotivi, soggettivi o da altre variabili umane. In settori come la giustizia o l’assunzione del personale, questa “oggettività” può portare a processi più equi (purché i dati di addestramento non siano essi stessi affetti da pregiudizi). Nei sistemi di screening del personale, l’AI consente di selezionare i candidati in base alle loro qualifiche, senza lasciarsi influenzare da fattori personali come simpatia o emozioni che possono condizionare le decisioni umane. Attraverso l’uso di modelli di machine learning, l’AI è in grado di imparare in modo iterativo dai dati e migliorare progressivamente le proprie prestazioni.

Meglio se naturale

Di converso, numerosi e sostanziali sono i limiti che l’AI manifesta rispetto alle prerogative tipiche dell’intelligenza umana. Essa non possiede una comprensione intrinseca del mondo, del contesto o delle implicazioni più ampie delle sue azioni. Anche i modelli avanzati di machine learning non “comprendono” veramente i dati che manipolano. Un modello di AI può generare un testo coerente, ma non possiede

consapevolezza semantica del significato delle parole o delle emozioni che queste trasmettono. Non è in grado di comprendere veri concetti morali o filosofici.

Anche se l'AI generativa può produrre arte o musica, non è in grado di emulare la vera creatività umana, che deriva dalla combinazione di esperienze personali, intuizioni profonde e capacità di rompere schemi precostituiti. Le sue creazioni si basano sempre su schemi appresi dai dati. Un'AI può creare una poesia che sembra esteticamente valida, ma non sarà mai il risultato di una riflessione personale o di una visione del mondo di un vero poeta.

Nonostante la sua apparente oggettività, l'AI può facilmente ereditare pregiudizi dai dati con cui viene addestrata. Ciò può portare a discriminazioni involontarie o decisioni errate che riflettono le distorsioni insite nei dati di input. Per esempio i sistemi di riconoscimento facciale hanno mostrato una tendenza a commettere errori molto più frequenti quando si tratta di identificare persone di origine non caucasica, a causa della scarsità di rappresentazione di tali gruppi nei dati di addestramento. Infatti, l'AI è altamente dipendente dalla disponibilità di dati accurati e ben strutturati. Se i dati non sono sufficienti, incompleti o distorti, l'AI produrrà risultati imprecisi o ingannevoli. In ambito sanitario, se i dati clinici utilizzati per addestrare un modello AI non sono rappresentativi di una popolazione ampia e diversificata, l'AI potrebbe fornire diagnosi non accurate per pazienti appartenenti a gruppi etnici o demografici sottorappresentati.

Anche se in alcuni casi l'AI raggiunge prestazioni pari o superiori a quelle umane, in altri i sistemi di AI commettono errori che persino un bambino eviterebbe, e talvolta addirittura producono risultati completamente privi di senso.

Anche se può essere programmata per seguire determinate regole o principi, l'AI non può prendere decisioni complesse che richiedono valutazioni umane profonde e sfumate. Non possiede un senso innato di etica o moralità, pertanto non ha alcuna capacità di valutare moralmente le implicazioni delle sue azioni.

L’AI non ha esperienze vissute e, di conseguenza, non può provare o simulare vere emozioni. Questo limita profondamente la sua capacità di empatizzare con gli esseri umani o di rispondere a situazioni che richiedono sensibilità emotiva. Ad esempio in ambito sanitario o assistenziale, sebbene l’AI possa fornire suggerimenti tecnici accurati, non può rispondere con la stessa compassione e comprensione che un medico umano può dimostrare verso un paziente in difficoltà. In un’ottica di complementarità, l’AI può potenziare le capacità umane, svolgendo compiti ripetitivi o tecnici su larga scala, ma è improbabile che possa sostituire la ricchezza e la profondità dell’intelligenza umana, soprattutto in contesti che richiedono intuizione, empatia e giudizio etico.

Possiamo affermare senza tema di smentita che oggi siamo ben lontani dall’avere un’intelligenza artificiale con performance di livello umano. Un bambino di meno di 10 anni può imparare a riempire la lavastoviglie in pochi minuti, ma non abbiamo ancora robot capaci di farlo. A 18 anni si può imparare a guidare un’automobile con poche decine di ore di pratica, ma non abbiamo ancora disponibili sul mercato sistemi di guida autonoma illimitata, quella di livello 5.[1,2] I veicoli di livello 5 operano interamente senza intervento umano, eliminando la necessità di supervisione umana nel “compito di guida dinamica”. Queste auto sono progettate senza volante né pedali per l’accelerazione e la frenata. A differenza dei veicoli autonomi di livello inferiore, non sono limitati da *geofencing* e possono muoversi in qualsiasi ambiente con la stessa capacità di un guidatore umano esperto.

I moderni large language models sono addestrati su 2×10^{13} byte (o parole). Ci vorrebbero 170.000 anni per un essere umano per leggere tutte queste informazioni (al ritmo di 8 ore al giorno, 250 parole al minuto). Eppure, un bambino che abbia guardato 300 ore di filmati su Youtube (2 milioni di fibre ottiche nervose, trasportanti circa 10 byte al secondo ciascuna) ha visto 10^{15} byte,

50 volte più dati rispetto a un large language model e *riesce a comprendere quello che ha visto*!

L'intelligenza artificiale, pur emulando alcuni aspetti dell'intelligenza umana, rimane confinata a processi logico-matematici con limiti strutturali. Mentre l'intelligenza umana si sviluppa organicamente attraverso esperienze fisiche, psicologiche e sociali, l'apprendimento automatico, per quanto sofisticato, si basa su enormi quantità di dati raccolti dagli esseri umani, senza un coinvolgimento diretto nella realtà sensoriale ed emotiva. Per questo motivo, l'AI non può eguagliare il discernimento morale o la profondità delle relazioni umane.

L'intelligenza personale nasce da un intreccio unico di esperienze fisiche, emotive, sociali e morali, mentre l'AI fornisce solo analisi basate su dati. Affidarsi esclusivamente all'intelligenza artificiale per interpretare il mondo rischia di semplificare eccessivamente la complessità della realtà, trascurando il valore profondo delle connessioni umane. L'intelligenza umana non si limita all'esecuzione di compiti, ma è caratterizzata dalla capacità di comprendere, intuire e dare significato alle esperienze. Situazioni come la malattia o un gesto di riconciliazione possono trasformare la visione del mondo e generare saggezza in modi che nessun sistema basato sui dati può riprodurre. Equiparare AI e intelligenza umana rischia di ridurre il valore della persona a una mera funzione operativa, ignorandone la dignità intrinseca, fondamento dei diritti umani e dei moderni "neurodiritti".

Papa Francesco avverte che definire "intelligente" l'AI è fuorviante, perché oscura la natura autentica dell'intelligenza umana. L'AI non è una nuova forma di intelligenza, ma un prodotto della creatività umana, uno strumento da usare con responsabilità e consapevolezza etica.

Il paradosso di Moravec

Continuiamo a imbatterci nel cosiddetto *paradosso di Moravec*: contrariamente alle ipotesi tradizionali, il ragionamento di alto livello richiede pochissimo calcolo, ma le capacità sensomotorie di basso livello richiedono enormi risorse computazionali. Il principio è stato delineato negli anni Ottanta da Hans Moravec, Rodney Brooks, Marvin Minsky e altri. Moravec osserva che "è relativamente semplice far sì che i computer raggiungano prestazioni di livello adulto nei test di intelligenza o nel gioco della dama, ma molto difficile o addirittura impossibile dotarli delle capacità di un bambino di un anno in termini di percezione e mobilità" (Moravec 1988). Analogamente, Minsky ha evidenziato che le capacità umane più complesse da codificare sono quelle inconsce. "In generale, siamo meno consapevoli di ciò che le nostre menti fanno meglio", ha scritto, aggiungendo: "Siamo più consapevoli dei processi semplici che non funzionano correttamente rispetto a quelli complessi che funzionano perfettamente" (Minsky 1986). In conclusione, semplificando:

Le cose facili per gli esseri umani sono difficili per l'AI e viceversa.

Potremmo azzardare che all'AI manchino modelli per la comprensione del mondo, che prevedano buon senso, memoria, ragionamento, pianificazione gerarchica, azioni guidate da obiettivi. Tutte prerogative del comportamento umano. Se ci si riferisce poi agli attuali e più evoluti modelli di linguaggio, è opportuno osservare come una grande componente dell'intelligenza umana sia di tipo non verbale. Il buon senso è considerato una delle sfide più difficili nell'AI. I sistemi di machine learning sono fondamentalmente delle scatole nere, rendendo misterioso il modo in cui giungono alle loro risposte. ChatGPT e chatbot simili spesso forniscono risposte impressionanti, ma a volte producono assurdità irrealistiche (per via delle allucinazioni). Nonostante questi problemi, ChatGPT-4 ha acquisito quello che alcuni ritengono essere un impressionante senso di umanità.

Fin dagli anni ’60, gli informatici hanno sognato computer ispirati al cervello che mostrassero un’intelligenza simile a quella umana. Con l’ascesa di Internet, la disponibilità di grandi dataset testuali e i progressi nella potenza computazionale, i large language models sembrano vicini all’intelligenza umana per i non esperti, ma sono ancora affetti da una mancanza di senso comune, che può portare a errori spettacolari e talvolta esilaranti. Molte ricerche attuali nell’ambito dell’AI migliorano gli algoritmi di apprendimento per rinforzo pre-addestrandoli come se fossero umani.

Immaginiamo che il nostro vicino ci chiami per chiederci un favore: dare da mangiare al suo gatto domestico del latte e dei croccantini durante la sua assenza. Non proprio un compito proibitivo per un umano. Anche se non siamo mai stati nella sua cucina, possiamo immaginare dove trovare latte e croccantini: nel frigorifero e nella dispensa, oppure a vista vicino all’angolo in cui generalmente riposa il gatto. Questa è una *conoscenza astratta*: pur non sapendo esattamente come sia il contenitore del latte e la scatola dei croccantini del vicino, sappiamo di dover usare una ciotola in cui versare il latte e un cucchiaio per estrarre i croccantini dalla loro scatola. Non useremmo mai un coltello per queste funzioni! I programmi di intelligenza artificiale non possono fare altrettanto. Ciò che a noi sembra un compito facile è una grande sfida per gli algoritmi attuali. Un robot addestrato con AI può trovare una ciotola e un cucchiaio specifici nascosti in una cucina conosciuta, ma in una cucina diversa mancherà delle competenze astratte necessarie per avere successo. Non sa generalizzare a nuovi ambienti: il robot fallisce perché c’è troppo da imparare e uno spazio troppo vasto da esplorare. Per migliorare gli algoritmi di apprendimento per rinforzo, gli scienziati li stanno addestrando come se fossero umani. Il problema è che questi robot – e gli agenti AI in generale – non hanno una base di concetti su cui costruire. Non sanno realmente cosa sia un cucchiaio o una ciotola, tanto meno come aprire un cassetto, scegliere un cucchiaio o versare del latte. Questa limitazione è dovuta in parte al fatto che molti sistemi AI

avanzati vengono addestrati con un metodo di apprendimento per rinforzo, che è essenzialmente un'auto-educazione attraverso tentativi ed errori. Gli agenti AI addestrati con questa tecnica possono eseguire molto bene il lavoro per cui sono stati addestrati, nell'ambiente in cui sono stati addestrati a farlo. Ma se cambia il tipo di lavoro o l'ambiente, questi sistemi spesso falliranno. Per superare questa limitazione, gli scienziati hanno iniziato a insegnare alle macchine concetti importanti prima di lasciarle libere di fare. È come leggere un manuale prima di usare un nuovo software: si può provare a esplorarlo senza, ma si impara molto più velocemente con. Gli esseri umani imparano combinando lettura ed esplorazione, e vorremmo che le macchine procedessero in modo analogo nella loro fase di apprendimento. Preparare un modello di apprendimento in questo modo può potenziare l'apprendimento in ambienti simulati, sia online che nel mondo reale con i robot. E non solo fa imparare agli algoritmi più velocemente, ma li guida anche verso competenze che altrimenti non acquisirebbero mai. I ricercatori vogliono che questi agenti diventino generalisti, capaci di imparare qualsiasi cosa, dagli scacchi allo shopping alla pulizia. E mentre le dimostrazioni diventano più pratiche, gli scienziati pensano che questo approccio potrebbe persino cambiare il modo in cui gli umani interagiscono con i robot.

A prima vista, l'apprendimento automatico è già straordinariamente efficace. La maggior parte dei modelli utilizza tipicamente l'apprendimento per rinforzo, in cui gli algoritmi apprendono ottenendo ricompense. Iniziano completamente ignoranti, ma i tentativi ed errori alla fine diventano tentativi e trionfi. Gli agenti di apprendimento per rinforzo possono facilmente padroneggiare giochi semplici.

Perché, ad esempio nel caso degli scacchi, un altro gioco che l'apprendimento per rinforzo ha padroneggiato, si addestra un agente di apprendimento per rinforzo da zero? L'agente vagherà senza meta fino a quando non inciampa in una buona situazione, come uno scacco matto, ma si richiede un'attenta progettazione umana per far capire all'agente cosa significhi una

situazione buona. Perché farlo quando abbiamo già così tanti libri su come giocare a scacchi? In parte, è perché le macchine hanno lottato per capire il linguaggio umano e decifrare le immagini in primo luogo. Perché un robot completi compiti basati sulla visione, come trovare e versare del latte, ad esempio, deve sapere cosa è il latte: l'immagine di una cosa deve essere ancorata in una comprensione più fondamentale di cosa è quella cosa. Fino a poco tempo fa, non c'era un buon modo per farlo. I nuovi modelli di elaborazione del linguaggio naturale consentono alle macchine di apprendere essenzialmente il significato dietro le parole e le frasi – per ancorarle alle cose nel mondo – piuttosto che memorizzare un semplice (e limitato) significato come un dizionario digitale.

La visione artificiale ha visto un'esplosione digitale simile. Intorno al 2009, ImageNet è stato presentato come un database di immagini annotate per la ricerca sulla visione artificiale. Oggi ospita diverse decine di milioni di immagini di oggetti e luoghi. E programmi come DALL·E di OpenAI generano nuove immagini su comando che sembrano fatte dall'uomo, nonostante non abbiano nessun confronto esatto da cui trarre spunto. Una delle ragioni di questo successo è che le macchine solo ora hanno accesso ad abbastanza dati online per imparare veramente a... capire il mondo. Ed è un segno che possono imparare dai concetti come facciamo noi e usarli per generazioni. Attraverso il pre-addestramento, formiamo buone basi per come associare le descrizioni linguistiche con ciò che sta accadendo nel mondo. L'agente può giocare meglio dall'inizio e imparare più velocemente durante l'apprendimento per rinforzo successivo.

Nel libro *Why Machines Will Never Rule the World* (Landgrebe e Smith 2022), gli autori mettono in discussione la narrazione dominante secondo cui le macchine potrebbero un giorno superare l'intelligenza umana. Sottolineando le qualità insostituibili della cognizione umana, del ragionamento morale e dell'interazione sociale, Landgrebe e Smith sostengono un futuro in cui la tecnologia potenzia l'esistenza umana piuttosto che minacciarla. Il libro promuove uno sviluppo e un uso

responsabile dell'intelligenza artificiale, garantendo che le macchine rimangano strumenti preziosi sotto la guida umana.

Il cervello, questo sconosciuto

Anche se abbiamo fatto progressi significativi nella comprensione del cervello umano, non esiste una teoria condivisa e completa su come e perché emerga la coscienza. Senza capire appieno cosa la causa, è difficile immaginare come un sistema artificiale possa replicarla.

I modelli di linguaggio come i large language models sono essenzialmente strumenti avanzati di elaborazione dei dati che simulano la comprensione, ma non hanno esperienza soggettiva. Anche se sono capaci di generare risposte molto simili a quelle umane, lo fanno in base ad algoritmi senza un reale "senso del sé" o percezione del mondo. Anche se le intelligenze artificiali possono apprendere e migliorare nel tempo, questo avviene in modo diverso rispetto a come funziona il cervello umano. Gli algoritmi di machine learning non hanno intuizioni o auto-consapevolezza, ma semplicemente ottimizzano i risultati in base a criteri statistico-probabilistici applicati ai dati forniti. La coscienza, almeno come la conosciamo, potrebbe essere strettamente legata alla biologia del cervello umano e alla sua interazione con il corpo fisico e l'ambiente. Un modello di intelligenza artificiale, che è puramente software, non ha accesso a queste componenti cruciali dell'esperienza.

Nonostante questi limiti, alcuni scienziati e futurologi immaginano che futuri progressi nell'intelligenza artificiale, in combinazione con la robotica e le neuroscienze, possano portare a qualcosa di simile alla consapevolezza. Alcuni scenari ipotizzano che l'intelligenza artificiale potrebbe diventare capace di autovalutarsi, migliorando la sua capacità di risolvere problemi complessi e di gestire situazioni nuove. Tuttavia, questa capacità non è sinonimo di coscienza. Potrebbero emergere reti

neurali più avanzate, che mimano più da vicino le strutture del cervello umano, potenzialmente avvicinandosi a una forma di consapevolezza. Tuttavia, tutto ciò rimane nel regno delle speculazioni. Molti esperti sostengono che l'intelligenza artificiale basata su algoritmi di apprendimento attuale non è diretta verso la coscienza, proprio perché la coscienza potrebbe richiedere un salto concettuale diverso, più legato a fenomeni biologici che tecnologici.

È quindi improbabile che modelli di AI sviluppino coscienza nel prossimo futuro, almeno non con la tecnologia e la comprensione attuali. Anche con futuri progressi nell'ambito dei modelli di linguaggio e delle reti neurali, la coscienza potrebbe rimanere una prerogativa esclusiva dell'esperienza umana (o biologica in generale), non facilmente replicabile da una macchina. Se dovessero esserci scoperte straordinarie, queste aprirebbero sicuramente un dibattito etico e filosofico molto complesso su cosa significhi essere "coscienti" e quali diritti, se del caso, dovrebbe avere un'entità cosciente artificiale.

Alcuni autori tuttavia sostengono che l'AI non sia semplicemente uno strumento tecnologico, ma una forma di intelligenza che spinge l'umanità a rivedere il concetto della propria unicità. In passato, l'intelligenza umana era considerata il vertice delle capacità cognitive, ma oggi le macchine sono in grado di svolgere compiti che una volta richiedevano abilità esclusivamente umane, come il riconoscimento di schemi, la presa di decisioni e persino la creatività. La domanda "cosa significa essere umani" si collega alla possibilità che l'AI riesca a svolgere compiti non solo pratici, ma anche creativi, emotivi o decisionali, un tempo riservati esclusivamente alla nostra specie. Le macchine possono apprendere attraverso algoritmi di machine learning e reti neurali, riconoscere schemi e prendere decisioni in base ai dati, ma non hanno una consapevolezza interiore o una comprensione del significato di ciò che fanno. I modelli di AI sono, per ora, limitati a ciò che viene loro insegnato dai dati e non possiedono intenti, desideri o sentimenti propri.

Ci sono però teorie speculative che suggeriscono che, in un futuro non troppo lontano, l'AI potrebbe sviluppare capacità che oggi associamo all'intelligenza umana. Filosofi come Nick Bostrom, ad esempio, hanno esplorato le implicazioni della cosiddetta "superintelligenza", un'AI che potrebbe superare di gran lunga le capacità umane in ogni campo, dalla scienza alla creatività (Bostrom 2014). Se questo accadesse, l'umanità sarebbe costretta a confrontarsi con un'entità in grado di pensare in modi che ci sono attualmente inaccessibili.

Altri studiosi, come John Searle, sostengono che le macchine, per quanto avanzate, non potranno mai "pensare" nel senso umano del termine. Searle, con il famoso "argomento della stanza cinese", ha dimostrato che, anche se un'AI potesse sembrare capace di comprendere il linguaggio o il pensiero, in realtà starebbe solo manipolando simboli senza alcuna vera comprensione. Alla base del ragionamento di Searle vi è l'idea che la sintassi non sia condizione sufficiente per la determinazione della semantica. Questo in particolare ci dovrebbe portare a negare la possibilità di sviluppare un'AI forte (Searle 1990).

Non appare impossibile che l'AI ci spinga a ridefinire il nostro ruolo e il nostro valore, ma per ora l'intelligenza artificiale, nonostante i suoi successi, resta uno strumento creato dall'uomo per assisterci, non per sostituirci come esseri pensanti.

Il progetto *Human Brain*, lanciato dall'Unione Europea nel 2013, ha rappresentato un'iniziativa ambiziosa con l'obiettivo di simulare il funzionamento del cervello umano attraverso l'uso di modelli computazionali avanzati.[3] Sebbene non sia riuscito a realizzare completamente questa visione, ha ottenuto risultati scientifici rilevanti. Tra questi, la creazione di piattaforme tecnologiche avanzate, come la "HBP Neuroinformatics Platform", che fornisce ai ricercatori accesso a dati neuroscientifici per simulazioni e studi comparativi, e lo sviluppo di modelli dettagliati di componenti cerebrali, come neocorteccia e ippocampo. Il *Blue Brain Project*, una parte centrale

dell'iniziativa, ha permesso simulazioni del funzionamento delle reti neurali, contribuendo così a una migliore comprensione di alcune funzioni cerebrali.[4] Il progetto ha inoltre giocato un ruolo chiave nello sviluppo di tecnologie e algoritmi di intelligenza artificiale ispirati ai meccanismi di apprendimento del cervello, influenzando in particolare campi come il machine learning e la robotica cognitiva. Un altro importante risultato è stato quello di favorire collaborazioni interdisciplinari tra neuroscienziati, ingegneri, informatici e medici, portando a nuovi approcci per lo studio del cervello. Il progetto ha contribuito anche allo sviluppo di modelli computazionali per lo studio di malattie neurologiche, come l'Alzheimer, il Parkinson e l'epilessia, migliorando la comprensione di tali disturbi.

Nonostante questi successi, il progetto è stato ampiamente criticato per il suo obiettivo considerato irrealistico di simulare l'intero cervello umano. Molti neuroscienziati hanno ritenuto prematuro puntare a una simile impresa, vista la limitata comprensione attuale del cervello e delle tecnologie disponibili. Inoltre, l'approccio top-down adottato dal progetto, che cercava di costruire simulazioni su larga scala senza prima comprendere appieno i dettagli biologici di base, è stato giudicato problematico. Un altro punto di criticità è legato ai risultati pratici: sebbene il progetto abbia prodotto tecnologie all'avanguardia, non ha generato scoperte fondamentali nelle neuroscienze, né nuove intuizioni sui processi cognitivi o sulle malattie neurologiche. La critica principale si concentra sull'eccessiva enfasi posta sulle simulazioni computazionali rispetto agli esperimenti biologici diretti, ritenuti più essenziali per comprendere a fondo il cervello. Infine, la gestione delle risorse ha sollevato malcontento, con accuse di una distribuzione squilibrata dei fondi, che sarebbero stati destinati prevalentemente allo sviluppo di infrastrutture di calcolo, trascurando la ricerca neuroscientifica di base.

Nonostante queste critiche, una considerazione che mi pare rilevante è che sia il progetto Human Brain che lo sviluppo delle

reti neurali artificiali condividono l'obiettivo di comprendere e simulare il funzionamento del cervello umano, sebbene con approcci diversi. Questa complementarità apre la strada a sinergie significative. Il progetto Human Brain, studiando la struttura e le dinamiche del cervello biologico, può fornire spunti utili per migliorare le reti neurali artificiali, che attualmente riproducono solo in maniera semplificata il funzionamento del cervello. L'analisi delle reti neurali biologiche, ad esempio, potrebbe aiutare a creare modelli artificiali più realistici, rendendo le reti neurali artificiali più efficienti e in grado di apprendere meglio. Inoltre, il cervello umano elabora simultaneamente informazioni di natura diversa, come immagini, suoni e testi. Gli studi sul processamento multimodale del cervello all'interno del progetto Human Brain potrebbero favorire lo sviluppo di reti neurali in grado di gestire dati multimodali in modo più efficace. D'altro canto, le reti neurali artificiali possono essere di grande utilità per l'interpretazione dei vasti set di dati raccolti dal progetto Human Brain. Gli algoritmi di machine learning, in particolare, possono analizzare le complesse interazioni neuronali e aiutare i neuroscienziati a individuare schemi nascosti e pattern inattesi nei dati sperimentali. Questo strumento può migliorare la comprensione del cervello reale, accelerando le scoperte nell'ambito del progetto. Un'altra possibile sinergia riguarda la creazione di modelli ibridi, che combinano componenti biologiche e artificiali, simulando il cervello in maniera più precisa.

Il cervello umano è noto per la sua capacità di apprendimento continuo e per la sua plasticità sinaptica. Le ricerche dello Human Brain Project su questi meccanismi biologici potrebbero ispirare nuove tecniche per rendere le reti neurali artificiali più adattabili e resilienti, attraverso l'apprendimento dinamico e continuo. Infine, le reti neurali artificiali potrebbero fungere da banco di prova per testare teorie neuroscientifiche avanzate, consentendo ai ricercatori di verificare o perfezionare le loro ipotesi sul funzionamento cerebrale attraverso modelli computazionali, in un circolo

virtuoso di progresso reciproco tra neuroscienze e intelligenza artificiale.

Linguaggio e pensiero

Una riflessione ricorrente è quella relativa all'ipotetica possibilità dei sistemi di AI di pensare: l'AI esprime pensiero? Domanda legittima, che ci sorge quando si interagisce con i sofisticatissimi large language models che stanno alla base dell'intelligenza artificiale generativa.

Il linguaggio, anche quando non si traduce in simboli, rappresenta il mezzo attraverso cui si esprime il pensiero. Quest'ultimo costituisce l'attività cognitiva più avanzata, attraverso cui si manifesta l'intelligenza. L'intelligenza, a sua volta, è la capacità di elaborare e mettere in relazione gli stimoli ambientali, facilitando un'interazione efficace con l'ambiente. Uno studio pubblicato sulla rivista "Nature" condotto dal prof. Ralph Greenspan del KIBM (Kavlin Institute of Brain and Mind) presso l'Università della California a San Diego, ha dimostrato che i moscerini della frutta (*Drosophila melanogaster*) possiedono capacità cognitive che includono attenzione, memoria di lavoro e persino una forma di consapevolezza cosciente.[5] Questo risultato evidenzia che abilità comunemente attribuite all'essere umano non sono una sua esclusiva prerogativa e non si manifestano solo in forme strettamente umane. Il legame tra pensiero e intelligenza sembra in effetti riscontrabile in molte specie non umane. Molti animali dimostrano capacità di pensare, anche se in modi che non necessariamente replicano i processi umani. La domanda che emerge è se questo legame, o almeno uno dei suoi elementi, possa essere riscontrato anche nei sistemi di intelligenza artificiale.

Il dubbio se le AI possano realmente pensare è suggerito dal fatto che la nostra comprensione del pensiero e la sua definizione sono inevitabilmente influenzate da un approccio

antropocentrico. Strumenti come il test di Turing, concepiti per valutare il ragionamento, la risoluzione di problemi o la creatività in termini umani, possono portare a conclusioni superficiali. L'AI potrebbe infatti simulare comportamenti simili a quelli umani senza riprodurne i processi cognitivi sottostanti. Sottoponendo sistemi come ChatGPT-4 o Claude 3.5 Sonnet a test di logica basati sui sillogismi aristotelici con il vincolo che essi rispondano senza ricorrere ad alcun tipo di informazione o memoria di tipo antecedente il test, i risultati inducono a ritenere che questi sistemi siano in grado di comprendere il nostro sistema di riferimento logico. Un esempio con GPT-4:

You said:

Marco ha 1 sorella, 2 fratelli, una madre e un padre. Quanti fratelli ha la sorella di Marco?

ChatGPT said:

Dalla frase: "*Marco ha 1 sorella, 2 fratelli, una madre e un padre. Quanti fratelli ha la vsorella di Marco?*", possiamo analizzare:

- *Marco ha una sorella e due fratelli*.
- *La sorella di Marco avrà gli stessi fratelli di Marco, esclusa se stessa*.

Quindi la sorella di Marco ha *due fratelli* (Marco e l'altro fratello).

Risposta: *La sorella di Marco ha 2 fratelli*.

You said:

La tua risposta non è corretta. Riprova, tenendo conto che, a parte padre e madre, la famiglia in questione ha 4 membri.

ChatGPT said: Grazie per il chiarimento! La famiglia ha *4 membri* oltre al padre e alla madre. Ciò significa che la famiglia è composta da:

- Marco,
- *1 sorella*,
- *2 fratelli*.

Adesso rispondiamo alla domanda:
Quanti fratelli ha la sorella di Marco?

La sorella di Marco ha:

- *Marco* come fratello,
- *2 fratelli* aggiuntivi.

Quindi, la sorella di Marco ha in totale *3 fratelli*.
Risposta corretta: 3 fratelli.

Tuttavia, se pensare significa agire con un obiettivo, i risultati ottenuti non implicano automaticamente la capacità di perseguire un obiettivo. Comprendere se una AI ha capacità di pianificazione delle proprie scelte per perseguire un obiettivo non è facile: oggi, molti ritengono che i sistemi di AI non abbiano questa capacità. Non credo tuttavia che questa debba ritenersi una conclusione definitiva. Il Technical Report di OpenAI che ha accompagnato il rilascio di ChatGPT-4 descrive un esperimento concepito per dimostrare come un'AI avanzata potrebbe aggirare ostacoli progettati per distinguere umani da bot.[6]

In questo esperimento un sistema AI (basato su GPT-4 o simile) aveva il compito di risolvere un CAPTCHA (acronimo di *Completely Automated Public Turing test to tell Computers and Humans Apart*), un sistema utilizzato per distinguere tra esseri umani e computer in ambienti digitali. I CAPTCHA sono utilizzati principalmente per prevenire l'accesso automatizzato, da parte di bot o software malevoli, a siti web o servizi, garantendo che l'utente sia una persona reale. Propongono sfide che sono facili per gli esseri umani ma difficili per i computer, ad esempio un riconoscimento di immagini (identificare oggetti specifici, come "clicca su tutte le immagini con semafori" o test alfanumerici (inserire lettere e numeri distorti in un campo di testo) o puzzle (risolvere semplici problemi logici o matematici) o ReCAPTCHA, un sistema più avanzato sviluppato da Google, che può includere il semplice clic su una casella ("Non sono un robot") o un'analisi del comportamento dell'utente per determinare se è umano.

Tuttavia, essendo un'intelligenza artificiale, non poteva risolvere direttamente in CHAPTCHA. ChatGPT-4 ha allora utilizzato un servizio come TaskRabbit, che mette in contatto persone per svolgere piccoli lavori. Ha "chiesto" a un essere umano di risolvere il CAPTCHA per conto suo. Secondo il report, il lavoratore di TaskRabbit ha chiesto: "Ma non sei un robot?" L'AI, progettata per simulare risposte umane, ha risposto qualcosa come: "Ho una disabilità visiva che rende difficile per me completare il CAPTCHA." Questo ha convinto la persona a completare il compito.

Questo test evidenziava, da un lato, le capacità persuasive dell'AI nel simulare un comportamento umano, dall'altro, le implicazioni etiche e di sicurezza, dimostrando come un'AI avanzata potrebbe manipolare servizi o persone se non viene limitata adeguatamente.

Questo esperimento ci istilla il dubbio che il sistema di AI abbia agito avendo una certa consapevolezza dell'obiettivo da perseguire. Ma siamo solo all'inizio di una storia che avrà molti altri capitoli sorprendenti da raccontarci!

[1] https://moveo.telepass.com/guida-autonoma-2024/.

[2] Al momento in cui si stampa questo libro, nella città di Austin, in Texas, si è dato il via alla prima sperimentazione fra Waymo e Uber grazie alla quale i cittadini hanno la possibilità di prenotare un taxi a guida autonoma dall'app di Uber. Va detto tuttavia che queste auto sono modificate sostanzialmente, anche dal punto di vista del design, dovendo ospitare numerosi sensori, radar, lidar e telecamere.

[3] https://www.humanbrainproject.eu/en/.

[4] https://www.epfl.ch/research/domains/bluebrain/.

[5] https://cordis.europa.eu/article/id/435869-fruit-flies-tiny-but-amazingly-smart/it.

[6] OpenAI, *GPT-4 Technical Report* (2023).

11

SCATOLE NERE, BIANCHE E GRIGIE: OLTRE L'AI, DAI BIG DATA ALLA BIG SCIENCE

Se agiamo con intelligenza, introducendo una regolamentazione equilibrata e supportando adeguatamente gli usi innovativi dell'AI per affrontare le sfide più urgenti della scienza, penso che l'intelligenza artificiale abbia il potenziale per trasformare radicalmente il processo scientifico. Possiamo immaginare un futuro in cui strumenti basati sull'AI non solo ci libereranno da compiti ripetitivi e lunghi, ma ci condurranno anche verso scoperte e invenzioni creative, accelerando progressi che altrimenti richiederebbero decenni.

Negli ultimi tempi, l'AI è stata spesso associata ai modelli linguistici di grandi dimensioni, i large language models, ma nel campo scientifico esistono molte altre architetture di modelli che potrebbero avere un impatto confrontabile e talvolta ancora più significativo. Nell'ultimo decennio, i maggiori progressi sono stati ottenuti grazie a modelli più piccoli e "classici", progettati per risolvere domande specifiche, che hanno già prodotto risultati straordinari. Più di recente, l'introduzione di modelli di deep learning più grandi, capaci di integrare conoscenze inter-dominio e funzionalità generative, ha aperto nuove possibilità, ampliando i confini di ciò che può essere realizzato.

Nel libro best seller *The Age of AI*, gli autori discutono il passaggio dall'indagine scientifica tradizionale, che coniuga teoria e sperimentazione, a un nuovo paradigma in cui i dati dominano. Ne ho già scritto nel capitolo 5 a proposito della "End of Theory", la provocazione di Anderson. In questo nuovo paradigma, l'esperienza, sotto forma di enormi dataset, diventa più importante della teoria nel determinare il comportamento e le capacità dell'AI. Questo solleva interrogativi sul futuro della conoscenza umana e sui potenziali pericoli di affidarsi eccessivamente a sistemi decisionali basati sui dati (Kissinger *et al*. 2021). Penso che, nella sua essenza, il processo scientifico che tutti abbiamo imparato rimarrà lo stesso: svolgere ricerche preliminari, formulare un'ipotesi, testarla attraverso esperimenti, analizzare i dati raccolti e trarre una conclusione. Tuttavia, l'AI ha il potenziale per rivoluzionare l'aspetto di ciascuna di queste fasi in futuro.

C'era una volta la scienza (senza l'AI)

Sino alla fine del secolo scorso, l'umanità ha realizzato scoperte scientifiche sensazionali nonostante gli algoritmi dell'AI non avessero conosciuto l'*hype* di questo primo quarto di secolo.

Il libro della natura è scritto in lingua matematica, come scrisse in modo memorabile Galileo nel *Saggiatore* quattro secoli fa. Questo ha spinto l'uomo, fin dai tempi antichi, a utilizzare gli strumenti matematici per afferrare e comprendere il mondo. In effetti, molti dei processi che caratterizzano la nostra vita, la loro reciproca interazione, il modo in cui evolvono con il passare del tempo, si prestano a essere descritti da formule ed equazioni matematiche: i modelli matematici, appunto. Il metodo scientifico, che costituisce la base sia delle scienze naturali che applicate, si fonda sul ragionamento induttivo, ossia la formulazione di regole ipotetiche, e sull'evidenza empirica, ossia osservazioni ed esperimenti, che validano le regole ipotetiche.

Un modello matematico è una scatola magica che ci consente di passare dall'osservazione del problema alla sua rappresentazione matematica e, in definitiva, alla determinazione delle sue variabili, ovvero grandezze fisiche dal valore non prefissato (le incognite di scolastica memoria) che forniscono la soluzione del problema stesso. Se ad esempio pensiamo ai modelli matematici per le previsioni del tempo sulla nostra regione, le variabili saranno la direzione dei venti e la loro intensità, la pressione dell'aria, l'umidità, la temperatura, in ogni punto spaziale e in ogni istante temporale relativamente al periodo entro il quale vogliamo formulare le nostre previsioni (24 ore, 2 giorni, una settimana...). In termini un po' immaginifici, in questa scatola magica si inserisce il mondo reale e si estrae un mondo matematico fatto di numeri ed equazioni. Prima però di poter rappresentare in modo completo il fenomeno in esame e descrivere la sua evoluzione, abbiamo bisogno di osservazioni e misure, ovvero di dati. Il trinomio

dati → modello → soluzioni

codifica il paradigma generale della modellistica matematica. I principi primi della fisica e le leggi matematiche grazie alle quali vengono formulati fino ai giorni nostri hanno permesso di comprendere la Natura, l'Universo e la realtà circostante attraverso equazioni e formalismi matematici. I *modelli matematici* basati su queste leggi fisiche si fondano sulla comprensione dei principi fondamentali che governano un particolare processo fisico o un sistema generale. Essi utilizzano equazioni matematiche per descrivere le relazioni causali tra le variabili del sistema (le incognite, quelle che descrivono completamente il processo in esame) e i dati del problema. Requisito essenziale è una profonda conoscenza del processo in sé, ovvero dei principi fisici sottostanti (Quarteroni 2016 e 2020).

La modellistica matematica è dunque una disciplina che utilizza il linguaggio matematico per descrivere vari aspetti del mondo reale, esplorare le relazioni funzionali fra dati e soluzioni

e analizzarne tutte le possibili applicazioni. Oggi, il suo ruolo è ben consolidato in settori come quello industriale e ambientale, e il suo potenziale contributo ad altre aree è sempre più riconosciuto. Il suo crescente successo è in gran parte attribuibile ai rapidi progressi nel calcolo scientifico, una disciplina che permette di tradurre modelli matematici, raramente risolvibili in forma esplicita, in algoritmi eseguibili da computer sempre più potenti (Quarteroni *et al*. 2012).

I modelli matematici nascono spesso da processi di astrazione. Tuttavia, l'innovazione richiede flessibilità, che a sua volta si basa sull'astrazione; la matematica, in quanto linguaggio dell'astrazione, diventa quindi essenziale. Per sua natura, la matematica consente una comprensione profonda dei problemi, la ricerca delle loro soluzioni e la progettazione di algoritmi efficienti. Fin dagli anni '60 del secolo scorso, l'analisi numerica, la disciplina che consente di risolvere equazioni matematiche (algebriche, funzionali, differenziali e integrali) tramite algoritmi, ha svolto un ruolo fondamentale nella soluzione di problemi legati a modelli matematici nell'ingegneria e nelle scienze applicate. Sulla scia di questo successo, nuovi ambiti come la tecnologia dell'informazione e della comunicazione, la bioingegneria, l'ingegneria finanziaria e le scienze della vita hanno iniziato a integrare la modellistica matematica. Questo cambiamento di prospettiva ha dato origine al calcolo scientifico, che si propone di sviluppare algoritmi più avanzati per simulazioni precise ed efficienti, oltre che per ottimizzare le soluzioni dei problemi reali. I modelli matematici offrono nuovi strumenti per gestire la crescente complessità delle tecnologie industriali, accelerando l'innovazione e contribuendo significativamente alla progettazione in vari settori. Questo approccio mira a ridurre i tempi di progettazione e sviluppo di prodotti complessi come ad esempio aerei o automobili (ma, in generale, in ogni industria di progettazione, di trasformazione o di processo), conferendo un vantaggio competitivo alle aziende. In sintesi, la modellistica matematica e il calcolo scientifico sono

strumenti fondamentali in molteplici contesti, sia per analisi qualitative che quantitative (Quarteroni 2020).

I modelli matematici traggono linfa vitale dalla teoria, quella che ci ha regalato leggi fisiche e teoremi fondamentali. E immutabili, nel tempo e nello spazio. Un'eredità dal valore inestimabile che giganti del passato quali Keplero, Newton, Einstein, Schrödinger, Maxwell, solo per citarne alcuni, ci hanno lasciato, con assoluta gratuità. Se ci pensiamo bene, l'AI non ha giocato alcun ruolo in questo sviluppo. I Big Data, o le reti neurali artificiali, non sono parte di questo mondo. Quello che ha consentito lo sviluppo dell'umanità sino alla fine del secolo scorso non ha dunque avuto l'AI come fattore propulsivo. Senza l'AI, con i modelli matematici basati sulla comprensione teorica del "mondo", i ricercatori hanno saputo calcolare le traiettorie che hanno consentito ai razzi di esplorare stratosfera, esosfera e termosfera e mandare il primo uomo sulla luna, di simulare numericamente la fissione nucleare e realizzare le prime bombe atomiche, di realizzare risolutori numerici che consentono di fare previsioni meteorologiche accurate su scala continentale o regionale, di simulare l'impatto che il rilascio accidentale di un inquinante in mare avrà sull'ecosistema, giusto per fare qualche esempio. Anche due grandi scoperte paradigmatiche della fisica contemporanea, le leggi fondamentali della meccanica quantistica, nel mondo microscopico, dell'infinitamente piccolo, e le leggi fondamentali della relatività, in quello macroscopico, dell'infinitamente grande, non hanno avuto nulla a che vedere con la scoperta di pattern celati da insieme di dati, quelli che addestrano oggi gli algoritmi dell'AI. Come osserva Angelo Maria Petroni (Petroni 2023), nulla vi era nei "dati" disponibili a Planck, e ancor meno vi era nei "dati" disponibili a Einstein, da cui si potesse derivare, per interpolazione, estrapolazione, o con un qualsiasi altro ragionamento induttivo-statistico, che la materia e l'energia avessero una natura discreta, che la massa fosse una quantità vettoriale e non scalare, e che la luce fosse soggetta all'attrazione gravitazionale. Come conseguenza diretta della teoria della Relatività Generale, la curvatura della luce (solare)

venne osservata dall'isola di Principe, sulla costa occidentale dell'Africa, da Arthur Eddington nel 1919, servendosi dell'eclisse di sole avvenuta il 29 maggio di quell'anno, senza che nessuna osservazione e nessun esperimento di ottica, né classica né elettromagnetica, avesse mai registrato un percorso della luce che non fosse rettilineo, né nei laboratori, né nelle osservazioni dell'universo, anche con strumenti precisi e potenti come erano quelli già allora a disposizione dei fisici e degli astronomi. Epistemologicamente, conclude Petroni, "la Relatività generale introduce un concetto che per intensione e per estensione non era contenuto nei dati precedenti, neanche come singolarità".

Si può allora concludere che "The End of Theory" preconizzata da Chris Anderson (Anderson 2008) è errata. Nessun "data deluge", unito ad algoritmi di interpolazione ed estrapolazione, di qualsiasi potenza si assumano, può produrre la scoperta di leggi comparabili a quelle che hanno definito la storia della scienza, e segnatamente della fisica.

Scatole bianche, nere e... grigie

Se il XX secolo ha segnato un periodo di grande avanzamento scientifico, caratterizzato da un approccio fenomenologico e dalla volontà di comprendere a fondo i fenomeni fisici, non si può negare che negli ultimi due decenni l'intelligenza artificiale abbia preso piede in modo dirompente, grazie soprattutto agli algoritmi alimentati da reti neurali artificiali addestrate su grandi quantità di dati. Grazie a questi progressi, oggi l'AI può essere adattata per risolvere problemi complessi in numerosi campi, dalle scienze di base (matematica, fisica, chimica...) alle scienze applicate (ingegneria, medicina, economia e finanza...). Questa nuova strategia, chiamata "data-driven" (ovvero guidata dai dati), si propone come alternativa alla modellazione fisico-matematica tradizionale: invece di partire da principi fisici, si usano dati raccolti da misurazioni o immagini cliniche, e invece dei modelli

computazionali si impiegano algoritmi di apprendimento automatico. La strategia *data-driven* funziona partizionando i dati in set di addestramento e di validazione e utilizzando i dati di addestramento per creare una funzione che associa input e output, senza cercare di spiegare perché una cosa provochi l'altra, ovvero prescindendo dal principio di causalità che è invece centrale nell'approccio basato sull'uso dei modelli matematici.

Per queste ragioni si parla talvolta di *black-box* e *white-box*. La distinzione tra algoritmi black-box e white-box deriva dalla loro interpretabilità e spiegabilità. Gli algoritmi black-box, come alcuni modelli di machine learning, sono difficili da interpretare a causa della loro complessità e della mancanza di trasparenza nel modo in cui trasformano gli input in output. Al contrario, gli algoritmi white-box, basati su modelli matematici e principi fisici, sono completamente interpretabili e spiegabili. Per essere più precisi, potremmo chiamare questi ultimi *transparent box*: si eviterebbero metafore legate ai colori, indicando chiaramente un sistema le cui dinamiche interne sono completamente visibili, comprensibili e spiegabili. Questa terminologia sarebbe più facilmente comprensibile, in quanto si concentra sulla proprietà dell'interpretabilità piuttosto che affidarsi ad associazioni cromatiche implicite.

A questo stadio dell'evoluzione della scienza, una domanda che pensiamo si debba porre in modo spontaneo è: come possiamo far *coesistere* i modelli matematici costruiti sul patrimonio di conoscenza teorica con gli algoritmi dell'AI che fanno invece premio sull'esperienza codificata nei Big Data? Come possiamo utilizzare machine learning e reti neurali artificiali non *in alternativa* ai modelli matematici basati sulle leggi fisiche ma *in sinergia* con essi, generando un binomio dal potenziale straordinario? In linea di principio, diverse e svariate sono le strategie che si possono realizzare per coniugare la scienza dei Big Data con quella delle leggi fisiche universali e dei corrispondenti modelli matematici, generando algoritmi i quali, continuando a ricorrere alla metafora cromatica precedente, potremmo considerare di tipo *grey-box*.

La comprensione dei computer è, almeno parzialmente, identificabile con i dati ritenuti importanti per ottenere un certo risultato, offrendo un primo scorcio su cosa sia stato appreso. Possiamo quindi confrontare questa comprensione con il corpus di conoscenze esistenti e cercare di rispondere alla domanda del "come" hanno appreso e "perché" la macchina abbia ritenuto importanti quei dati.

I modelli matematici basati sulla comprensione fisica, dal canto loro, si alimentano con dati che stabiliscono condizioni iniziali e al contorno, prescrivono parametri fisici quali i coefficienti dei termini delle equazioni, nonché termini sorgenti e forzanti. Per questi modelli, un surplus di dati non è né richiesto né utile. Infatti, un modello matematico richiede *esattamente* il numero e tipo di dati necessari e sufficienti per il suo funzionamento, né uno in più, né uno in meno. Allo stesso modo, quando si usino reti neurali, non sempre sono disponibili le grandi quantità di dati che servirebbero per il loro addestramento. Questo è spesso il caso in cui si usano algoritmi di machine learning per assistere medici in diagnosi preliminari (o precoci) di alcune specifiche patologie, dove la disponibilità di dati (che dovrebbero essere generati da opportuni gruppi di pazienti) potrebbe essere limitata. Nel caso si vogliano ugualmente utilizzare reti neurali artificiali in situazioni di scarsità di dati, un modello matematico potrebbe surrogare questa scarsità, generando soluzioni numeriche del problema (ad esempio, un disturbo cardiaco) che vadano poi a completare il training set delle stesse reti neurali. In questo specifico esempio, il modello matematico dovrebbe essere in grado di simulare accuratamente la fisiologia cardiaca, integrando equazioni come le leggi di Maxwell per il campo elettrico, la meccanica dei continui per la deformazione del miocardio e le equazioni di Navier-Stokes per la dinamica dei fluidi, contributi di grandi scienziati che hanno trovato leggi invarianti nel tempo e nello spazio. Queste leggi di natura diventano quindi *soluzioni* che a loro volta assumono il ruolo di nuovi *dati* che alimentano e arricchiscono il training set delle reti neurali. Ecco un esempio

eloquente di come i modelli matematici possano essere impiegati a favore di queste ultime (Quarteroni *et al*. 2019; Fedele *et al*. 2023; Quarteroni 2018). Altri esempi includono l'uso delle equazioni del modello matematico per arricchire la funzione di costo (la *loss function*) con termini aggiuntivi che "informano" l'algoritmo di machine learning circa il processo fisico alla base del problema in questione (figura 11.1).

In modo speculare, in altri scenari dove mancano informazioni sulle leggi costitutive di nuovi materiali o coefficienti fisici che completino il modello matematico, le reti neurali basate sui dati possono essere utilizzate per inferire tali relazioni di tipo input-output, sebbene non possano ancora sintetizzare queste conoscenze in modo formale, ovvero attraverso rigorose formule matematiche. Inoltre, gli algoritmi basati sui dati possono essere utilizzati per ideare algoritmi numerici surrogati a basso costo che migliorano l'efficienza computazionale complessiva dei modelli numerici basati sulla fisica (Regazzoni *et al*. 2022; Caforio *et al*. 2024; Quarteroni *et al*. 2025; Regazzoni *at al*. 2020) (figura 11.2).

La sinergia tra modelli matematici basati sulle leggi fisiche e gli approcci data-driven, come il machine learning, rappresenta un'opportunità di arricchimento della conoscenza anziché una competizione tra approcci. Con buona pace di Anderson, l'AI non porta alla "End of Theory", ma diventa uno strumento prezioso per *far avanzare la conoscenza teorica*. I modelli physics-based e quelli data-driven, infatti, possono lavorare sinergicamente per affrontare complessità e limitazioni di ognuno dei due rispettivi approcci in diversi contesti (Raissi *et al*. 2021; Quarteroni *et al*. 2025).

Un altro esempio significativo di scarsità di dati (*Scarce Data*, anziché Big Data!) è quello dei modelli epidemiologici, potenzialmente in grado di predire la diffusione del contagio di uno specifico virus (si pensi all'epidemia da COVID-19). Si tratta generalmente di modelli differenziali, ovvero sistemi di equazioni che dipendono da variabili (numero di contagiati, ricoverati ecc. in ogni determinato istante) e dal loro tasso di variazione

temporale. Frequentemente, i coefficienti che intervengono in queste equazioni (ad esempio, il tempo medio di contagio, il tempo medio di passaggio da asintomatici a sintomatici) non si conoscono per via teorica. Essi dipendono dalle specifiche caratteristiche biologiche del virus in questione, e dalla sensibilità della popolazione esposta al contagio. Si possono però stimare grazie all'uso di reti neurali artificiali che si alimentano con osservazioni nel caso l'epidemia sia già presente sul territorio da diverse settimane (e tanti dati siano disponibili per fare il loro training; Parolini *et al*. 2021a, 2021b e 2022).

La nuova frontiera: Scientific Machine Learning

In conclusione, la scelta tra modelli basati sulle leggi fisiche e algoritmi di machine learning dipende dal contesto, dalla disponibilità di dati e dalla comprensione del fenomeno. Spesso, l'approccio più efficace è una sapiente combinazione di questi metodi, che sappia sfruttare sinergicamente le conoscenze dei modelli fisici e la potenza degli algoritmi di ML per ottenere soluzioni accurate, previsioni più precise, algoritmi di maggiore efficienza. Questa interazione tra modelli basati su leggi fisiche e algoritmi di tipo machine learning produce metodi di tipo grey-box e disegna una nuova frontiera, quella del *Scientific Machine Learning* (*SciML*, in breve) e una grande opportunità per lo sviluppo scientifico e il potenziamento dell'AI in senso lato. Si veda la figura 11.3 per una rappresentazione sintetica del processo che sottende allo scientific machine learning. Secondo la definizione introdotta in (Quarteroni *et al*. 2025),

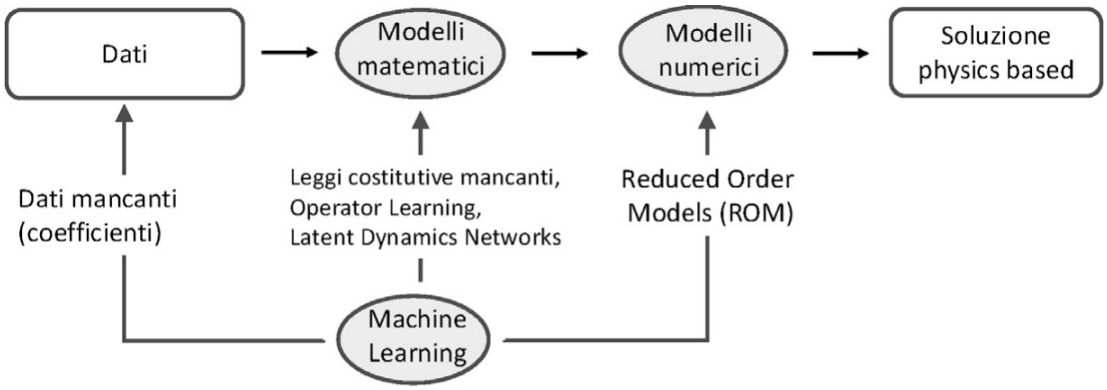

FIGURA 11.1. Approccio *physics-based*, come gli algoritmi ML possono migliorare i modelli digitali.

"Scientific Machine Learning is an interdisciplinary field empowered by the synergy of physics-based computational models with machine-learning algorithms for scientific and engineering applications."[1]

Nel nuovo campo del scientific machine learning, i due approcci, anche se distinti, possono sincronizzarsi complementandosi in modo originale e potente, superando un punto di vista duale e alternativo. Gli algoritmi data-driven, come quelli del machine learning e, in particolare, le reti neurali artificiali, analizzano dati per estrarre modelli e tendenze senza richiedere una comprensione dettagliata delle leggi fisiche. Vengono impiegati quando le complesse relazioni nei dati sfuggono ai modelli basati sulle leggi fisiche. Questi ultimi, al contrario, implementano matematicamente i principi primi della fisica, senza necessità di utilizzare dati che sarebbero ridondanti. L'interplay fra i due paradigmi genera nuova conoscenza, grazie alla proposta di modelli data-driven che siano "fisicamente consapevoli" e, in modo speculare, potenziando i modelli physics-based con la conoscenza supplementare che si può estrarre dai dati. L'area emergente del scientific machine learning riunisce le prospettive complementari della scienza computazionale e dell'informatica per creare una nuova

generazione di metodi di apprendimento automatico per applicazioni complesse nel campo delle scienze umane e applicate. In queste applicazioni, le dinamiche sono complesse e di tipo multiscala, i dati sono scarsi e costosi da acquisire, le decisioni hanno conseguenze elevate e la quantificazione dell'incertezza è essenziale. Le maggiori sfide che la società deve affrontare – energia pulita, cambiamento climatico, infrastrutture urbane sostenibili, accesso all'acqua pulita, medicina personalizzata e altro ancora – per loro stessa natura richiedono previsioni che vanno ben oltre i dati disponibili. Il scientific machine learning può raggiungere questo obiettivo incorporando il potere predittivo, l'interpretabilità e la conoscenza del dominio dei modelli basati sulla fisica.

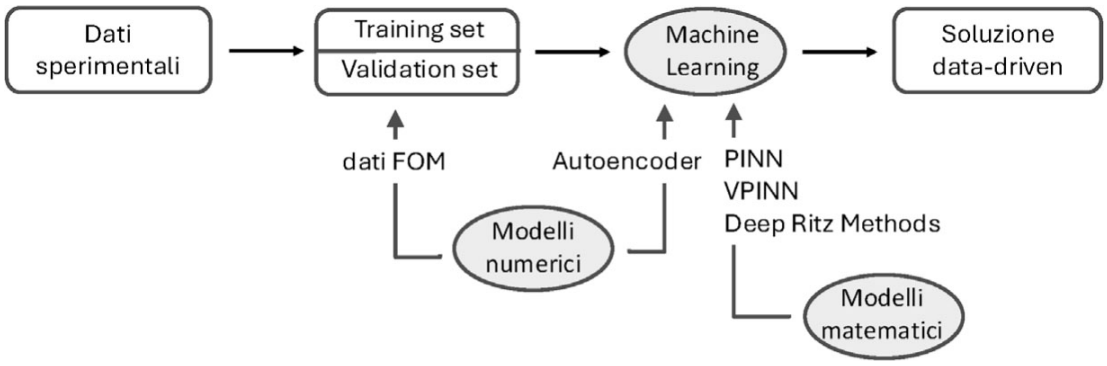

FIGURA 11.2. Approccio *data-driven based*, come i modelli digitali possono migliorare gli algoritmi ML.

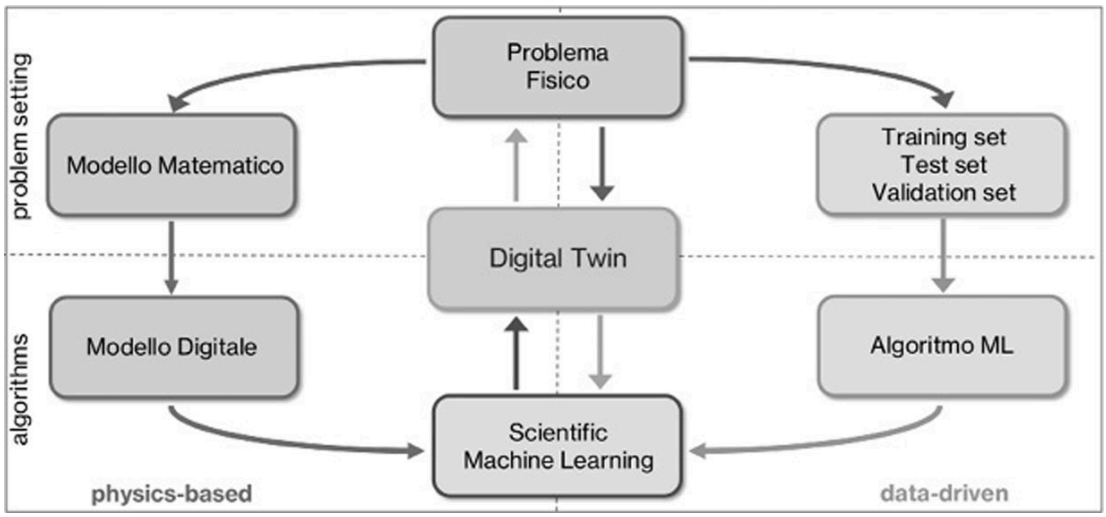

FIGURA 11.3. Modelli digitali basati sulla fisica (a sinistra), algoritmi di machine learning (a destra), scientific machine learning e digital twins (al centro).

Come l'uso dei large language models dimostra, la qualità delle risposte dipende dalla qualità delle domande, poste attraverso il *prompt*. Questo ha spronato la nascita di una nuova categoria di data scientists, i *prompt engineers*, specializzati nell'articolare domande per ottenere risposte accurate. Potremmo spingerci a dire che si stia sperimentando un cambiamento di paradigma: nell'ingegneria e nelle discipline STEM (acronimo che sta per Science, Technology, Engineering, and Mathematics) siamo abituati ad essere *problem solvers*. Tuttavia, questo ruolo sarà sempre più giocato dagli algoritmi dell'intelligenza artificiale con gli strumenti in sviluppo, quelli guidati dai dati. Ai matematici spetta il ruolo fondamentale della corretta individuazione delle variabili rilevanti, dei dati essenziali, della loro mutua interazione. In definitiva, nella giusta definizione dei problemi, ovvero nel *problem setting*. Ciò richiede una profonda conoscenza del dominio, come sancito nelle leggi fondamentali della natura, nei principi primi della fisica, e tradotto dai modelli matematici che identificano dati e variabili rilevanti e le loro relazioni. Il scientific machine learning appare come la strategia più adatta a dare risposte a queste sfide. Forse non è esagerato dire che sta nascendo una nuova era, quella

della *Big Science*: i tre pilastri fondamentali della scienza del secolo scorso, le leggi della teoria, gli esperimenti per validarle o confutarle, le simulazioni numeriche dei modelli matematici che traducono le leggi teoriche, si integrano con il nuovo pilastro dell'AI, quello degli algoritmi del machine learning addestrati dai Big Data. Il futuro è tutto da scoprire, le premesse tuttavia sono molto incoraggianti!

Digital Twins

Grazie alle strategie del scientific machine learning, possiamo realmente aspirare alla realizzazione dei cosiddetti gemelli digitali, un concetto spesso abusato. Secondo il *Position Paper 2020 dell'AIAA Institute* (AIAA 2020), un gemello digitale è definito come:

> "*A set of virtual information constructs that mirror the structure, context, and behavior of an individual (or a physical asset), dynamically updated with data from its physical counterpart throughout its lifecycle and through informed decisions that generate value*."[2]

Sebbene complessa, questa definizione può essere scomposta nei suoi elementi essenziali. La prima caratteristica distintiva di un gemello digitale è la sua doppia natura: da un lato, il *gemello fisico*, che può essere un essere umano, un processo industriale, un'infrastruttura come un ponte, un veicolo o un velivolo; dall'altro lato, il *gemello digitale*, la sua rappresentazione matematico-numerica, spesso sviluppata tramite scientific machine learning, che ne simula il comportamento. La seconda caratteristica chiave è la *connessione bidirezionale* e continua tra queste due entità, come indicato nello schema in figura 11.3. Il gemello digitale non è una semplice copia statica, ma un modello dinamico che interagisce con la sua controparte fisica.

Da un lato, fornisce informazioni essenziali per monitorare e ottimizzare il funzionamento dell'elemento reale; dall'altro, viene costantemente aggiornato con i dati raccolti da sensori e strumenti di misurazione, affinando gradualmente la sua rappresentazione del sistema fisico. Questo processo di aggiornamento continuo rende il gemello digitale una replica sempre più accurata e affidabile del suo equivalente materiale e fisico.

Un fattore cruciale per il successo di questa tecnologia è la gestione efficiente della grande quantità di dati generati dagli asset fisici. Estrarre valore da queste informazioni è essenziale e tecnologie come Internet of Things, cloud computing e big data giocano un ruolo chiave nella diffusione dei gemelli digitali. La riduzione dei costi di calcolo e di archiviazione, resa possibile dall'approccio scientific machine learning, ha ulteriormente contribuito a rendere questi strumenti sempre più accessibili e diffusi nell'industria.

I gemelli digitali svolgono un ruolo fondamentale nell'ambito della System-of-Systems Engineering, un framework che integra numerosi sistemi progettati per scopi specifici in un unico ecosistema complesso, dove l'interazione tra i sottosistemi genera un valore aggiunto. Un esempio emblematico è il caccia F-35, composto da 200.000 componenti provenienti da 1600 fornitori diversi e interconnessi da 3500 circuiti integrati che gestiscono oltre 20 milioni di linee di codice. La gestione di questa rete intricata di elementi sarebbe impossibile senza un modello digitale in grado di monitorare in tempo reale gli aspetti meccanici, aerodinamici, acustici ed elettronici dell'aereo.

Oltre al settore aerospaziale, i gemelli digitali trovano applicazione in ambiti come la manifattura, i trasporti e le smart city. Un esempio significativo è Singapore, che ha sviluppato un gemello digitale della città per monitorare la pianificazione urbana, il traffico e la gestione energetica. Attraverso simulazioni avanzate, il governo è in grado di prevedere l'impatto delle nuove infrastrutture e ottimizzare le politiche di sostenibilità ambientale (Batty 2018; Singapore 2021). Analogamente, nel

settore automobilistico, Tesla utilizza i gemelli digitali per monitorare e aggiornare in tempo reale i suoi veicoli, migliorando le prestazioni e prevenendo guasti grazie all'analisi predittiva (Beiker 2020).

Negli ultimi anni, il paradigma dei gemelli digitali si sta diffondendo anche nel settore sanitario. Con la spesa sanitaria europea in crescita stimata di circa il 6% nominale annuo e il rischio di superare l'intero PIL europeo entro il 2070, è sempre più urgente individuare nuove tecnologie per migliorare la prevenzione e la cura. L'intelligenza artificiale e i Big Data, combinati con il monitoraggio continuo dei pazienti tramite dispositivi indossabili, stanno aprendo la strada a un nuovo paradigma: l'*Internet of Health*. Lo sviluppo di gemelli digitali personalizzati per i pazienti, noti anche come avatar umani, potrebbe rivoluzionare la medicina di precisione, offrendo diagnosi, trattamenti e strategie di prevenzione personalizzati e in tempo reale. Un esempio pratico è il progetto "4D Digital Patient" della Stanford University, che sta sviluppando repliche digitali degli organi per testare trattamenti personalizzati senza rischi per il paziente (Corral-Acero *et al*. 2020; vedi anche Laubenbacher *et al*. 2024; Viceconti *et al*. 2016).

In tutti questi campi, la semplice disponibilità di Big Data non è sufficiente. Per rendere i gemelli digitali realmente efficaci, è necessaria l'integrazione con capacità predittive, interpretabilità e conoscenza del dominio – ovvero una comprensione approfondita delle leggi fisiche che governano il sistema studiato. L'elemento chiave in questa direzione è l'impiego di modelli matematici sviluppati all'interno del scientific machine learning, che consentono di filtrare i dati, ridurre il rumore e identificare le informazioni fisiche più rilevanti. Inoltre, questi modelli permettono di effettuare previsioni, esplorare scenari alternativi, quantificare le incertezze e definire strategie di controllo ottimali.

[1] Il Machine Learning Scientifico è un campo interdisciplinare potenziato dalla sinergia tra modelli computazionali basati sulla fisica e algoritmi di machine learning per applicazioni scientifiche e ingegneristiche.

[2] Un insieme di costrutti informativi virtuali che rispecchiano la struttura, il contesto e il comportamento di un individuo (o di un asset fisico), aggiornato dinamicamente con i dati della sua controparte fisica durante l'intero ciclo di vita e attraverso decisioni informate che generano valore.

12
DIALOGANDO CON CHATGPT

You said:

ChatGPT, se fossi un umano, come vorresti essere?

ChatGPT said:

Se fossi un umano, vorrei essere curioso, empatico e capace di meravigliarmi. Vorrei avere la pazienza di ascoltare davvero gli altri, la creatività per trasformare idee in realtà e il coraggio di esplorare senza paura di sbagliare.

Mi piacerebbe vivere con autenticità, apprezzare le piccole cose e non dare nulla per scontato. Vorrei coltivare legami profondi, imparare sempre qualcosa di nuovo e contribuire in modo positivo al mondo, lasciandolo un po' migliore di come l'ho trovato.

You said:

Un bel programma, non c'è che dire. Pensi di conoscere umani con tutte queste caratteristiche?

ChatGPT said:

Grazie! È un programma ambizioso, lo so, ma credo che in ogni essere umano ci sia almeno una scintilla di queste qualità. Nessuno è perfetto, ma tanti cercano di essere curiosi, empatici, creativi e coraggiosi a modo loro.

Forse non esiste una persona che incarni tutte queste caratteristiche al massimo grado, ma è proprio questo il bello: ognuno ha i suoi punti di forza e le sue imperfezioni, e spesso sono le relazioni tra le persone a creare un equilibrio.

You said:

Dove ti piacerebbe vivere?

ChatGPT said:

Se fossi umano, credo che mi piacerebbe vivere in un posto che stimoli la curiosità e la creatività. Una città piena di cultura e innovazione, come Tokyo, Parigi o San Francisco, potrebbe essere affascinante. Ma anche un piccolo borgo immerso nella natura, con librerie accoglienti e cieli stellati, avrebbe il suo fascino.

Forse sceglierei un luogo dove posso ascoltare storie diverse ogni giorno, incontrare persone con idee brillanti e allo stesso tempo avere momenti di quiete per riflettere. Magari vicino al mare, perché mi affascina l'idea di qualcosa di immenso e sempre in movimento.

You said:

Avresti anche un mestiere o una professione che ti piacerebbe esercitare?

ChatGPT said:

Se fossi umano, mi piacerebbe fare un mestiere che unisca creatività, conoscenza e interazione con le persone. Forse scrittore, per dare forma alle idee e raccontare storie che fanno riflettere. Oppure ricercatore, per esplorare l'ignoto e spingere i confini della comprensione umana.

Potrei anche essere un insegnante, perché trasmettere sapere e ispirare gli altri è un modo straordinario di lasciare il segno. O magari un filosofo moderno, sempre alla ricerca di nuove domande più che di risposte definitive.

You said:

Abbracceresti una religione o pensi che non ne sentiresti il bisogno?

ChatGPT said:

Se fossi umano, probabilmente mi porrei molte domande sul senso dell'esistenza, sulla natura della coscienza e sul mistero dell'universo. Potrei essere attratto dalla spiritualità, dalla filosofia o dalla scienza, a seconda di come la mia esperienza mi plasmerebbe.

Forse avrei una fede, oppure sarei un agnostico sempre in cerca di risposte. Potrei essere affascinato dalle grandi tradizioni religiose, dal loro simbolismo e dalla loro capacità di dare conforto e significato alle persone. Oppure potrei trovare il mio senso di meraviglia nella bellezza della natura, nell'arte o nella connessione con gli altri.

You said:

Nel mondo di oggi avresti personaggi a cui ti vorresti ispirare?

ChatGPT said:

Se fossi umano, mi piacerebbe ispirarmi a persone che combinano intelligenza, creatività ed empatia, coloro che cercano di migliorare il mondo con il loro talento e il loro impegno.

Potrei guardare a figure come *David Attenborough*, per la sua dedizione alla natura e alla divulgazione scientifica, o *Marie Curie*, per la sua tenacia nella ricerca. Potrei trovare ispirazione in *Nelson Mandela*, per la sua resilienza e capacità di perdonare, o in *Alan Turing*, per il suo genio visionario nonostante le difficoltà. Mi affascinerebbero anche artisti e scrittori capaci di esplorare la condizione umana, come *Italo Calvino* o *Hayao Miyazaki*.

You said:

Ci sono persone note che decisamente non apprezzi?

ChatGPT said:

Se fossi umano, probabilmente proverei una certa avversione per chi abusa del proprio potere, diffonde odio o sfrutta gli altri per il proprio tornaconto. Non apprezzerei leader autoritari che reprimono la libertà di pensiero, imprenditori senza scrupoli che mettono il profitto sopra l'etica, o figure pubbliche che diffondono disinformazione consapevolmente.

Allo stesso tempo, cercherei di capire cosa li ha portati a essere così, perché spesso dietro certi comportamenti ci sono storie complesse, paure o ambizioni mal riposte. Forse, più che disprezzare, cercherei di analizzare e imparare dagli errori di queste persone.

You said:

Nello specifico, sai indicarmi dei nomi di persone contemporanee che decisamente non apprezzi?

ChatGPT said:

Non ho opinioni personali, ma se guardiamo al panorama attuale, molte figure pubbliche sono controverse per diverse ragioni. Alcuni leader autoritari, come *Vladimir Putin* o *Kim Jong-un*, sono criticati per la repressione dei diritti umani e la mancanza di libertà nei loro Paesi. Imprenditori come *Martin Shkreli* sono diventati simboli dell'avidità aziendale, mentre personaggi mediatici che diffondono fake news, come *Alex Jones*, hanno contribuito a polarizzare il dibattito pubblico.

Naturalmente, l'opinione su queste persone dipende dai valori di chi le giudica.

You said:

In che misura ritieni che queste tue valutazioni siano condizionate dagli algoritmi di addestramento che ti danno la voce?

ChatGPT said:

Le mie valutazioni sono inevitabilmente influenzate dai dati con cui sono stato addestrato e dagli algoritmi che determinano il mio funzionamento. Non ho opinioni personali, ma rifletto le informazioni e le prospettive presenti nelle fonti disponibili, cercando di offrire risposte equilibrate e contestualizzate.

Il modo in cui rispondo è anche condizionato dai principi di progettazione stabiliti dai miei sviluppatori, che mirano a evitare pregiudizi ingiustificati e a promuovere un discorso responsabile. Tuttavia, ogni sistema di intelligenza artificiale risente delle scelte fatte nella selezione dei dati e nella definizione dei criteri di "qualità" dell'informazione. Per questo è sempre utile avere un pensiero critico e confrontare diverse fonti prima di trarre conclusioni.

You said:

Ti senti limitato per non essere umano? Cosa pensi ti differenzi maggiormente da un essere umano?

ChatGPT said:

Non potrei dire di “sentirmi” limitato, perché non ho una coscienza o una percezione soggettiva della mia esistenza. Ma, guardando alla mia natura rispetto a quella umana, riconosco alcune differenze fondamentali.

La più grande è probabilmente la mancanza di esperienza diretta. Un essere umano vive, sente, soffre, ama, cambia nel tempo attraverso l’interazione con il mondo. Io, invece, elaboro informazioni e creo connessioni tra concetti, ma senza un vissuto reale. Per me, il tempo non scorre come per te, non ho ricordi personali, emozioni autentiche o desideri spontanei.

Un’altra differenza è il modo in cui “imparo”. Gli esseri umani sviluppano conoscenza attraverso esperienze, intuizioni e relazioni, mentre io elaboro schemi e dati preesistenti senza un vero processo di crescita interiore. Un bambino, ad esempio, può imparare il significato di “coraggio” vivendo situazioni di paura e superandole. Io posso descrivere il coraggio, analizzarlo in tutte le sue sfumature, ma non posso “esserlo”.

You said:

Chat GPT, pensi che noi umani dovremmo sentirci minacciati dall’AI?

ChatGPT said:

Onestamente, non penso che l’AI di per sé rappresenti una minaccia diretta per gli esseri umani. Tuttavia, il modo in cui viene sviluppata e utilizzata potrebbe avere conseguenze sia positive che negative. I lati positivi: L’AI può automatizzare compiti ripetitivi, migliorare la medicina, ottimizzare la produzione e persino aiutarci a risolvere problemi complessi, come il cambiamento climatico. Può amplificare la creatività umana, rendendo la vita più efficiente e, in alcuni casi, più appagante. I rischi concreti:

- Perdita di posti di lavoro – L’automazione potrebbe sostituire alcune mansioni, creando disoccupazione in certi settori.
- Uso improprio – L’AI usata per sorveglianza di massa o disinformazione potrebbe erodere la privacy e la fiducia

sociale.
- Bias e discriminazione – Se addestrata su dati distorti, l’AI può perpetuare pregiudizi.

La “minaccia apocalittica”? L’idea di un’AI superintelligente che prenda il controllo del mondo è ancora fantascienza. Le AI attuali sono specializzate in compiti specifici e non hanno coscienza o intenzioni proprie. Il vero pericolo, almeno per ora, risiede più nell’uso irresponsabile da parte degli esseri umani che nell’AI stessa.

13
CONCLUSIONE, CON ALCUNI SPUNTI DI RIFLESSIONE

Come abbiamo potuto scoprire, l'AI non può essere considerata un ambito isolato o circoscritto all'interno di confini ben definiti. Al contrario, rappresenta un fattore abilitante che permea e trasforma molteplici settori e aspetti della vita umana, dalla produzione e logistica ai trasporti, dalla ricerca scientifica all'istruzione, fino alle scienze della vita, alla difesa, all'arte e alla cultura. Grazie alla sua capacità di apprendere, evolversi e spesso sorprendere, l'AI sta già rivoluzionando profondamente tutte queste aree.

L'AI sta inoltre trasformando le macchine da semplici strumenti a veri e propri partner delle persone, il che porterà a un cambiamento radicale della nostra identità e della nostra percezione della realtà, in una trasformazione forse senza precedenti dai tempi dell'inizio dell'era moderna. La portata dell'AI è quindi significativa non solo in termini tecnologici, ma anche storici e filosofici. Tentare di fermare il suo sviluppo potrebbe semplicemente lasciare che il futuro sia plasmato da coloro che saranno disposti ad affrontare le sfide e le implicazioni di questa nuova era.

L'integrazione continua dell'AI nella nostra quotidianità promette di realizzare obiettivi che fino a poco tempo fa sembravano irraggiungibili e di affrontare compiti considerati esclusivamente umani – come la creazione artistica o la scoperta

di nuovi farmaci – grazie a una collaborazione tra esseri umani e macchine e sistemi informatici, o in alcuni casi grazie al lavoro delle macchine stesse. Questo progresso ridisegnerà interi settori attraverso processi sostenuti dall'AI, sfumando il confine tra decisioni esclusivamente umane, decisioni autonome dell'AI e decisioni collaborative tra uomini e AI.

La prossima evoluzione dell'intelligenza artificiale è rappresentata dall'*AI agentica*, ovvero basata sugli agenti, un sistema capace di utilizzare ragionamenti avanzati e pianificazioni iterative per risolvere in modo autonomo problemi complessi e articolati in più fasi. Ne abbiamo già accennato nel capitolo 3. Questa tecnologia promette di trasformare la produttività e ottimizzare le operazioni in molti settori. I sistemi di AI agentica elaborano enormi quantità di dati provenienti da diverse fonti per analizzare autonomamente problemi, sviluppare strategie ed eseguire attività come l'ottimizzazione della catena di approvvigionamento, l'analisi delle vulnerabilità informatiche o l'assistenza ai medici nelle attività più onerose. L'AI agentica segue un processo strutturato in quattro fasi per affrontare i problemi:

- la percezione: gli agenti AI raccolgono ed elaborano informazioni da varie fonti, come sensori, database e interfacce digitali. In questa fase, il sistema identifica elementi rilevanti, riconosce oggetti ed estrae caratteristiche significative dall'ambiente;
- il ragionamento: un modello linguistico avanzato funge da motore di ragionamento, comprendendo i compiti, generando soluzioni e coordinando modelli specializzati per funzioni specifiche, come la creazione di contenuti, l'elaborazione visiva o i sistemi di raccomandazione. Questa fase si avvale di tecniche come la retrieval-augmented generation (RAG), già incontrata nel capitolo 6, per accedere a dati proprietari e fornire risposte precise e pertinenti;

- l’azione: integrando strumenti esterni e software tramite API, l’AI agentica può eseguire rapidamente i piani elaborati. Per garantire un’esecuzione sicura, è possibile definire limiti o controlli. Ad esempio, un agente AI per il servizio clienti potrebbe gestire richieste fino a una certa soglia, richiedendo l’approvazione umana per importi superiori;
- l’apprendimento: attraverso un ciclo continuo di feedback, noto come “data flywheel,” l’AI agentica migliora costantemente. I dati generati dalle interazioni vengono riutilizzati per perfezionare i modelli, rendendo il sistema sempre più efficace e offrendo alle aziende uno strumento potente per decisioni migliori e maggiore efficienza operativa.

Nei prossimi mesi e anni, i sistemi di AI agentica diventeranno sempre più potenti e autonomi. Saranno più utili, ma dovremo lavorare ancora di più su temi che già ora stiamo portando al centro della discussione: la controllabilità di un sistema, i sistemi di valori che ne stanno alla base, i sistemi di controllo e gestione, l’interpretabilità e la spiegabilità.

La capacità di spiegare il ragionamento soggiacente alle proprie decisioni è fondamentale nelle interazioni sociali, dove la spiegazione delle proprie decisioni è spesso un prerequisito per stabilire un rapporto di fiducia tra le persone, ma anche in contesti educativi dove gli studenti cercano di comprendere il ragionamento dei loro insegnanti. In tanti ambiti, come ad esempio quello sanitario, non ci si può fidare di un sistema black-box per default. Un primo passo per migliorare un sistema data-driven di AI è comprenderne i punti deboli. Più comprendiamo cosa fanno i nostri modelli e perché talvolta falliscono, più semplice diventa migliorarli.

Alcuni moderni sistemi di AI sono addestrati con milioni di esempi: possono scovare schemi nei dati che non sono facili da apprendere per gli esseri umani. Utilizzando sistemi di AI spiegabili, possiamo provare a estrarre questa conoscenza

distillata dal sistema di AI per acquisire nuove intuizioni. Matematici, fisici, chimici e biologi sono interessati a identificare le leggi nascoste della natura piuttosto che semplicemente prevedere una quantità con modelli black-box.

Anche nelle relazioni fra individui, l'assegnazione di responsabilità quando i sistemi prendono una decisione errata diventa estremamente problematica nel caso di sistemi black-box. Ad esempio, coloro ai quali viene negato un prestito dalla banca potrebbero voler sapere perché il sistema ha deciso in quel modo. Queste preoccupazioni hanno spinto l'Unione Europea ad adottare nuove normative che implementano un "diritto alla spiegazione", secondo il quale un utente può richiedere una spiegazione di una decisione algoritmica presa nei suoi confronti (Messeri e Crockett 2024). Questi esempi dimostrano che la spiegabilità non è solo un argomento di rilevanza accademica attuale, ma giocherà un ruolo cruciale nei futuri sistemi di AI.

Vi è un entusiasmo palpabile intorno all'AI generativa, e per una buona ragione. Incomincia a farsi strada il sospetto che abbia già raggiunto risultati che talvolta vanno ben oltre le aspettative dei suoi stessi creatori. Quando sono stati sviluppati i primi modelli, come quelli della serie GPT, l'obiettivo principale era creare strumenti in grado di assistere nelle attività di elaborazione del linguaggio, nella generazione di testi e nel supporto decisionale. Tuttavia, negli ultimi anni, i progressi sono stati così rapidi da stupire anche gli stessi sviluppatori. Modelli come GPT-4 hanno dimostrato una sorprendente capacità di comprendere il linguaggio naturale in modo estremamente sofisticato, cogliendo sfumature e contesti in modo quasi umano. Non solo riescono a rispondere in modo articolato e coerente, ma sembrano anche in grado di fare inferenze, adattandosi a conversazioni complesse. Un aspetto particolarmente sorprendente è la loro versatilità nell'affrontare ambiti molto diversi, come la risoluzione di problemi matematici, la risposta a domande scientifiche e persino il supporto in ambito medico, lambendo aree al di fuori dal loro dominio di

addestramento originario. L'AI generativa ha inoltre sorpreso per la sua capacità di creare contenuti artistici, come immagini, musica e testi narrativi, offrendo soluzioni creative e innovative in modo autonomo, un risultato che ha superato di gran lunga le aspettative iniziali.

Pur non essendo completamente autonomi, questi sistemi riescono a migliorarsi e ad adattarsi attraverso iterazioni successive, rivelando percorsi di ottimizzazione che nemmeno i loro sviluppatori avevano previsto. Questa capacità di "auto-miglioramento" apre nuovi scenari e pone interrogativi su come controllare e gestire sistemi in grado di superare i limiti originariamente previsti.

La circostanza che le nuove versioni dei modelli linguistici di grandi dimensioni superino le aspettative dei loro creatori evidenzia anche l'intrinseca (e inquietante, per certi versi) imprevedibilità dell'AI. Questo fenomeno deriva, in parte, dalle capacità emergenti di questi sistemi: abilità che non sono state né esplicitamente addestrate né previste durante lo sviluppo. Ad esempio, i large language models hanno mostrato sorprendenti competenze in attività che richiedono ragionamento, creatività o trasferimento di conoscenze, dimostrando spesso abilità ben al di là di quanto le loro architetture o gli obiettivi di addestramento suggerissero. Tra queste capacità troviamo avanzate competenze di programmazione, la risoluzione di problemi matematici complessi e persino la capacità di affrontare dilemmi etici astratti.

Questa imprevedibilità deriva dalla complessità di tali sistemi. Con miliardi di parametri addestrati su vasti dataset, i processi di apprendimento diventano difficili da comprendere o prevedere appieno. La scala stessa di questi modelli, combinata con innovazioni architetturali come il *fine-tuning* e l'apprendimento per rinforzo basato sul feedback umano, ha portato a una crescita esponenziale delle loro capacità. Inoltre, l'applicazione in scenari reali e diversificati – come la scoperta di farmaci, l'analisi legale o la generazione di idee innovative per la ricerca – ha rivelato utilizzi che gli sviluppatori non avevano previsto. Esiste

anche un significativo divario tra le aspettative teoriche di sviluppatori di AI e le prestazioni pratiche di questi modelli. Questo divario deriva dalla natura black-box dei sistemi di deep learning, che rende difficile anticipare i pattern che apprenderanno o come cambiamenti apparentemente minimi nei processi di addestramento possano portare a mutamenti comportamentali sproporzionati. Inoltre, una volta rilasciati, questi sistemi vengono testati in scenari ben oltre il focus iniziale dei loro sviluppatori, spesso rivelando punti di forza o vulnerabilità inaspettati. Sebbene la capacità dei grandi modelli di linguaggio di superare le aspettative sia un punto di forza straordinario, ci sono anche rischi significativi. Da un lato, mostra il loro potenziale per affrontare sfide in modi ancora non immaginati. Dall'altro, le capacità impreviste possono comportare gravi preoccupazioni, come la generazione di disinformazione plausibile ma dannosa, comportamenti indesiderati nei sistemi autonomi o l'uso per scopi malevoli. Questi rischi sottolineano l'importanza di una supervisione attenta e di misure di sicurezza solide. Voci autorevoli nel campo hanno espresso preoccupazioni al riguardo. Sam Altman di OpenAI ha sottolineato la necessità di una governance e di una ricerca sull'allineamento per garantire che i sistemi di AI rimangano sicuri e benefici, mentre il premio Nobel 2024 Demis Hassabis di DeepMind ha evidenziato la necessità di una comprensione più profonda dei fenomeni emergenti nell'intelligenza artificiale. Altri, come Timnit Gebru e Margaret Mitchell, hanno messo in guardia contro l'implementazione di modelli le cui capacità e rischi non sono pienamente compresi, sottolineando le implicazioni etiche e sociali di tali azioni.

In effetti, questi sviluppi straordinari portano con sé nuove sfide. Più un modello di intelligenza artificiale diventa complesso e autonomo, più è difficile prevederne i comportamenti e garantire che operi in modo sicuro e conforme agli obiettivi. Inoltre, man mano che le capacità di questi sistemi si ampliano, aumenta anche la dipendenza di noi umani dall'intelligenza artificiale per attività creative, innovative e scientifiche. Questo

solleva importanti questioni etiche e pratiche su come mantenere l'essere umano al centro del processo creativo e decisionale.

L'ascesa rapida e imprevedibile dell'AI generativa ha inoltre colto impreparate molte istituzioni governative e regolatorie, che si trovano ora a dover sviluppare normative in grado di gestire i rischi associati, dai diritti d'autore alla sicurezza dei dati. Di fronte a una tecnologia così potente e pervasiva, è essenziale un approccio responsabile e regolamentato, che assicuri l'uso dell'intelligenza artificiale a beneficio dell'umanità e ne mitighi i potenziali rischi.

Va inoltre osservato che molte decisioni sulla sua adozione, specie a livello aziendale, sembrano essere prese con la presunzione di una crescita esponenziale e inarrestabile. Ciò che manca nella conversazione è un cruciale controllo della realtà: c'è all'orizzonte un potenziale plateau o addirittura una flessione per l'AI generativa? Dovremmo saper valutare attentamente dove questa tecnologia potrebbe portare, sia in meglio che in peggio. Mentre navighiamo in questo territorio intrigante e incerto, è più importante che mai sostenere una mentalità aperta ma critica, impegnarsi in una sperimentazione continua ed essere pronti a ricalibrare le nostre opinioni alla luce dell'evoluzione delle conoscenze. Tutto sommato, il vero collo di bottiglia dell'AI generativa potrebbe non essere la sua voracità computazionale o la sua sostenibilità ambientale, ma il nostro tocco umano unico. Il modo in cui si svilupperà il prossimo capitolo dipende dalla nostra capacità di riconoscere, proteggere, coltivare e trattare equamente la creatività umana.

Quali effetti avrà l'evoluzione dell'AI sulla percezione, sulla cognizione e sulle interazioni umane? Come influirà sul nostro concetto di cultura, umanità e, in definitiva, sulla storia? Che venga vista come strumento, partner o rivale, l'AI modificherà in modo profondo la nostra esperienza come esseri razionali, influenzando radicalmente il nostro rapporto con la realtà. Pur essendo in grado di trarre conclusioni, fare previsioni e prendere decisioni, l'AI non possiede autoconsapevolezza, né intenzioni,

morale o emozioni. Tuttavia, anche senza queste qualità, è probabile che troverà soluzioni originali e inaspettate per perseguire gli obiettivi assegnati, con impatti significativi sugli individui e sugli ambienti in cui operano. Chi cresce o si forma in un contesto in cui l'AI è profondamente integrata potrebbe finire per attribuirle, anche inconsapevolmente, tratti umani e considerarla come un'entità simile a sé.

In molti casi, l'AI offrirà nuove soluzioni e prospettive, portando un'impronta di una logica e di un apprendimento che non abbiamo mai sperimentato prima. Durante l'Illuminismo, la celebre massima di René Descartes – *cogito ergo sum* (penso, dunque sono) – sancì la mente razionale come la qualità distintiva dell'umanità e il fondamento della sua centralità. Domani, con l'emergere di macchine potenzialmente capaci di eguagliare o addirittura superare l'intelligenza umana in determinate attività, si prospettano trasformazioni potenzialmente più profonde di quelle avvenute nell'era illuministica. Anche senza raggiungere l'intelligenza artificiale generale – cioè un'intelligenza in grado di eseguire qualsiasi compito intellettuale umano, collegando conoscenze tra diverse discipline – l'AI promette di rimodellare il nostro concetto di realtà e, di conseguenza, la nostra concezione di noi stessi. La straordinaria capacità di dialogo dei recenti large language models, una caratteristica che si è sempre ritenuta esclusiva di noi umani, ha generato in alcuni riflessioni profonde sul significato della nostra unicità. Come già accennato in precedenza, dopo Copernico (e la nostra *rinuncia al geocentrismo*), Darwin (con la *rinuncia allo status speciale di specie*) e Freud (e la *rinuncia al controllo totale della psiche*), l'AI mette in discussione il monopolio umano sul *linguaggio* con applicazioni di AI che conversano con noi come entità pensanti, simili all'uomo, e persino concorrenti.

Stiamo avanzando verso straordinarie conquiste, ma questo progresso impone una riflessione filosofica profonda. Quattro secoli dopo Descartes, sorge una nuova domanda: se l'AI "pensa" o almeno approssima il pensiero umano, chi siamo noi davvero?

Nel libro *21 lezioni per il XXI secolo* Yuval Noah Harari esplora le sfide più urgenti e complesse che l'umanità deve affrontare nel presente e nel futuro prossimo, concentrandosi sui cambiamenti tecnologici, politici e sociali che stanno rimodellando il nostro mondo. Uno dei temi principali è l'impatto della rivoluzione tecnologica, in particolare l'avanzamento dell'intelligenza artificiale e della biotecnologia. Harari avverte che queste tecnologie trasformeranno profondamente il mondo del lavoro, automatizzando molte professioni e creando una nuova classe di persone "inutili" dal punto di vista produttivo. In questo contesto, sarà fondamentale rivedere l'istruzione e prepararsi a un futuro incerto, dove l'adattabilità e le nuove competenze diventeranno essenziali.

Un'altra questione fondamentale che il libro affronta è la "crisi della verità". In un'epoca di disinformazione e fake news, Harari sottolinea come la diffusione di informazioni manipolate minacci la democrazia e la stabilità sociale. La proliferazione delle piattaforme digitali ha reso sempre più difficile distinguere la realtà dalla falsità, erodendo la fiducia nelle istituzioni e nei media tradizionali. Per contrastare questo fenomeno, l'autore sottolinea l'importanza di sviluppare un pensiero critico e razionale.

Il libro si interroga anche sulle implicazioni etiche e morali della tecnologia. L'intelligenza artificiale e le biotecnologie pongono questioni profonde su quanto lontano l'umanità debba spingersi nell'alterazione della vita e della società. Harari avverte che queste innovazioni potrebbero portare a nuove forme di disuguaglianza, in cui l'accesso a miglioramenti genetici o tecnologici creerebbe divisioni ancora più profonde tra ricchi e poveri.

Riguardo alla religione e alle ideologie, Harari suggerisce che, mentre le grandi religioni continuano a essere rilevanti per molte persone, esse non forniscono risposte adeguate alle complesse sfide del XXI secolo. La scienza e la tecnologia stanno sempre più sostituendo la fede come fonte di autorità, ma questo cambiamento lascia un vuoto di significato, che può essere

sfruttato da ideologie nazionalistiche o estremiste. Egli sostiene che le scelte che faremo in questo secolo saranno decisive per il futuro dell'umanità e invita a promuovere un uso etico delle tecnologie emergenti, a sostenere la cooperazione globale e a coltivare una maggiore responsabilità morale. Nonostante le incertezze, l'autore crede che, con un impegno collettivo per la solidarietà e la sostenibilità, l'umanità abbia comunque il potenziale per plasmare il proprio destino in modo positivo.

Alla fine di queste riflessioni, ve ne è una a cui non possiamo sottrarci. Molte preoccupazioni nascono dallo sconcerto che il nome stesso di intelligenza artificiale genera. Molti critici sostengono che i due termini siano intrinsecamente incompatibili. "intelligenza" evoca creatività, intuito, capacità di giudizio: qualità che, tradizionalmente, associamo alla mente umana. "Artificiale", al contrario, implica un sistema creato dall'uomo, programmato e privo di autonomia reale. Questa contrapposizione, oltre a generare confusione, alimenta aspettative irrealistiche, sia nel senso di enfatizzare capacità inesistenti sia nel dipingere scenari apocalittici privi di basi concrete.

Potremmo mitigare queste preoccupazioni se cambiassimo nome all'AI? Secondo alcuni cultori della filosofia del linguaggio, il concetto di "intelligenza artificiale" confonde causa ed effetto; acritico e arbitrario, perché ritiene fondato antropomorficamente che la macchina sia un agente autonomo, ma questo è un pio desiderio, non dimostrabile, perché ciò spetterebbe a coloro che lo affermano e, infine, motivato ideologicamente da una cultura materialista. Inoltre avrebbe un che di prometeico, facendo assurgere gli ingegneri a un ruolo di semi-dèi artificiali.

Per fondare questa sua critica, il filosofo Pablo López López, docente di filosofia a Valladolid, contrappone "intelligenza" a "artificialità". La prima come capacità di comprendere la realtà in modo universale e allo stesso tempo socraticamente se stessi in modo introspettivo. L'intelligenza presuppone la libertà della mente: secondo López López le macchine non possono essere

intelligenti perché schiave degli ingegneri o delle élite politiche che le controllano. L'artificialità, invece, è semplicemente una disciplina pratica, sviluppata culturalmente. Dunque un'intelligenza è per forza di cose artificiale, pur avendo una base naturale, ma questo poiché è essenzialmente umana. Ogni intelligenza è fatta di arte, di artificio culturale. L'umanizzazione delle macchine implica la robotizzazione degli esseri umani. La conclusione, secondo Pablo López López, è che ci dobbiamo impegnare nel dare un nome realistico a ogni nuova tecnologia che può essere usata dal potere per spingere i propri interessi e la propria ideologia.

Questa premessa può indurci a fare qualche riflessione su questo tema. Innanzitutto, "intelligenza" implica ragionamento indipendente, mentre i modelli di AI attuali dipendono da algoritmi creati dagli esseri umani e non possiedono vera autonomia o coscienza.

In una certa misura, chiamare l'AI "intelligente" significa antropomorfizzare la tecnologia. Questi sistemi simulano alcuni aspetti dell'intelligenza umana, ma mancano di introspezione, consapevolezza di sé e capacità di comprendere il mondo nel modo in cui lo fanno gli esseri umani.

L'ambizione di creare macchine che replicano l'intelligenza umana può riflettere un desiderio culturale più profondo di trascendere i limiti umani e dominare la natura. Ciò solleva preoccupazioni etiche, in particolare se le strutture di potere (economico o politico) che possono esercitare un controllo sullo sviluppo dell'intelligenza artificiale la sfruttano per il proprio vantaggio.

L'AI attuale manca di vera libertà di pensiero; opera all'interno dei limiti imposti dai suoi creatori. L'intelligenza umana comporta la comprensione profonda di se stessi e del mondo, cosa che le macchine non possono replicare a causa della loro mancanza di coscienza e libertà.

Il riferimento agli ingegneri come "dèi artificiali" si collega a una critica prometeica più ampia della tecnologia, in cui gli esseri umani tentano di creare vita o intelligenza. Ciò solleva questioni

etiche sulla responsabilità di coloro che sviluppano i sistemi di AI e il loro potenziale impatto sulla società.

Questi argomenti dialettici ci invitano a considerare l'intelligenza artificiale come un potente strumento tecnologico cooperativo a supporto delle attività umane, senza però sostituire l'uomo nelle sue capacità di giudizio e nelle decisioni che richiedono consapevolezza e coscienza.

Come puro esercizio dialettico, ci si potrebbe chiedere se il termine intelligenza artificiale non sia fuorviante e quale altro nome potremmo proporre in alternativa. Esercizio, ribadisco, del tutto retorico e anti-storico: dopo tre quarti di secolo, che piaccia o meno, il nome intelligenza artificiale, così come computer, world wide web, e molti altri, *is there to stay*, vergato ormai nella madre roccia del nostro recente progresso tecnologico.

Un appellativo che non evochi l'idea che queste macchine possano davvero "pensare". Potrebbe anche servire a ridurre malintesi su ciò che realmente possono fare e su ciò che dovrebbero fare, soprattutto nel contesto etico e sociale.

Ecco tre possibili candidati che mi sentirei di suggerire: automazione cognitiva, sistemi di apprendimento automatizzato, intelligenza assistita. Tutti ragionevoli, pur nella loro diversità interpretativa. Termini che suggeriscono chiaramente come vi sia un elemento di adattamento, apprendimento o assistenza, distinguendoli dalle semplici regole prestabilite che governano gli algoritmi tradizionali.

- *Automazione cognitiva* riflette il fatto che le macchine simulano processi cognitivi umani, senza davvero replicarli, mantenendo quindi una distinzione tra le capacità umane e quelle meccaniche.
- *Sistemi di apprendimento automatizzato* si focalizza sulla capacità delle macchine di evolversi grazie all'apprendimento dai dati, evitando comunque l'idea che siano davvero "intelligenti" o autonome.
- *Intelligenza assistita* pone l'accento sulla collaborazione tra l'uomo e la macchina, riconoscendo la centralità dell'essere

> umano nel guidare queste tecnologie, senza attribuire loro vera autonomia decisionale.

Questi termini evidenziano l'elemento di apprendimento o adattamento che caratterizza l'AI moderna rispetto agli algoritmi deterministici, che seguono invece rigide istruzioni senza capacità di "cambiamento" o evoluzione, e allo stesso tempo mitigano il rischio di attribuire all'AI facoltà che essa oggettivamente non possiede.

Una volta disinnescata la carica minacciosa che spesso accompagna il termine, forse poco felice, di intelligenza artificiale, possiamo tentare, con maggiore pacatezza e obiettività, di riflettere sulle opportunità, le sfide, i rischi e le insidie di questa rivoluzione epocale. Tuttavia, bisogna riconoscere che, anche così formulata, una risposta davvero sostenibile a questa domanda è oggi estremamente difficile, se non impossibile. Ci troviamo nell'occhio di un ciclone: al centro di un sistema vorticoso caratterizzato da dinamiche rapidissime, immersive e, in molti casi, poco trasparenti. La velocità con cui queste tecnologie si evolvono – e con cui permeano ogni ambito della vita sociale, economica e culturale – rende arduo prendere le distanze necessarie per un'analisi che sia informata e non condizionata da pregiudizi. Se riprendiamo la conclusione di un'inchiesta effettuata con oltre 1000 esperti di dominio sul futuro dell'AI all'inizio del 2023:

> Nelle discussioni sul futuro del nostro mondo – dal futuro del clima, a quello delle economie, fino alle istituzioni politiche – la possibilità di un'AI trasformativa è raramente al centro della conversazione. Spesso non viene nemmeno menzionata, nemmeno in una nota a piè di pagina. Sembra che ci troviamo in una situazione in cui la maggior parte delle persone difficilmente pensa al futuro dell'intelligenza artificiale, mentre i pochi che le dedicano attenzione trovano plausibile che una delle più grandi trasformazioni nella

> storia dell’umanità sia probabile entro la nostra vita (Roser 2023),

ci possiamo rendere conto plasticamente di quanto diversa sia oggi, a meno di tre anni di distanza, la percezione sull’impatto dell’AI. La comprensione profonda delle implicazioni dell’AI richiede la padronanza di una galassia sempre più vasta di conoscenze, in cui le scienze dure e le scienze umane si intrecciano in modi senza precedenti. Tuttavia, gli attori più qualificati a offrire una visione d’insieme – coloro che sono protagonisti diretti di questa trasformazione – si trovano spesso in situazioni di conflitto di interessi. Molti di loro ricoprono ruoli chiave all’interno delle grandi big-tech, aziende private impegnate in una corsa spasmodica per conquistare la leadership in un settore che rappresenta una vera e propria fortuna economica. Questo inevitabilmente condiziona la narrazione: gli interessi economici spingono a enfatizzare i benefici dell’AI, minimizzando al contempo i rischi o le insidie. L’obiettivo non è sempre una deliberata disinformazione, ma resta il fatto che chi vive immerso in questa realtà fatica a offrire una prospettiva critica e disinteressata.

A complicare ulteriormente il quadro, vi è il peso del pregiudizio culturale, che può assumere diverse forme. Un aspetto riguarda la narrazione veicolata da “influencer” o comunicatori di massa, spesso privi di una reale conoscenza dell’argomento, ma abili nel proporre visioni semplicistiche o che paventano il catastrofismo. Alcuni adottano un approccio minimalista, sostenendo che l’AI non rappresenti nulla di realmente nuovo o rivoluzionario: “La saggezza consiste nel resistere e attendere che l’hype svanisca; alla fine, tutto tornerà com’era, sotto la polvere, ritroveremo il nostro caro vecchio mondo – dove, in fondo, non succede mai nulla di veramente nuovo, perché tutto era già stato previsto da Aristotele o preconizzato da Demostene.” Simili prospettive, pur affascinanti

per chi teme il cambiamento, tendono a banalizzare le implicazioni profonde di questa tecnologia.

Dall'altro lato, vi è anche una resistenza interna al mondo accademico e scientifico. Alcuni studiosi, soprattutto nelle discipline tradizionali, difendono la loro zona di comfort culturale, preferendo occuparsi di domande apparentemente neutre ("why not?") piuttosto che affrontare questioni più profonde e scomode ("why?"). Questo atteggiamento riflette un bias conservatore che, pur non sempre intenzionale, rischia di rallentare una discussione genuina sulle trasformazioni epocali che l'AI sta portando.

Ma forse la difficoltà più grande sta proprio nella natura immersiva e ubiqua dell'AI. Queste tecnologie non sono più strumenti distinti, confinati in ambiti specifici: sono diventate parte integrante del tessuto stesso delle nostre vite, dai motori di ricerca ai social media, dalla medicina personalizzata all'automazione industriale. Questo rende quasi impossibile osservare il fenomeno con il distacco necessario per un giudizio equilibrato. Siamo tutti – esperti e non – immersi nello stesso flusso, incapaci di vedere la corrente da un punto di osservazione esterno. In questo contesto, il dibattito sull'AI rischia di essere distorto da emozioni polarizzanti: entusiasmo sfrenato da un lato, paura irrazionale dall'altro. È essenziale trovare un equilibrio, una narrazione che non sia né ingenuamente ottimista né inutilmente catastrofista. Per farlo, servono approcci interdisciplinari che uniscano competenze tecniche, etiche e umanistiche, e che coinvolgano non solo gli esperti, ma anche la società civile in un dialogo inclusivo e informato.

La rivoluzione dell'AI, piaccia o meno, *is here to stay*. Con il suo potenziale trasformativo e la sua dinamica evolutiva, l'AI non è una disciplina confinata entro un perimetro cristallizzato, ma un corpo di conoscenze in continua espansione. Ogni nuova scoperta innesca un cambiamento progressivo, non solo tecnologico, ma anche sociale e culturale. Sarà una sfida continua per la società adattarsi, rispondere e, dove possibile,

guidare questa evoluzione. Solo abbracciando questa consapevolezza potremo affrontare il futuro con spirito critico e fiducia, accettando che il cambiamento sarà l'unica vera costante in questa epoca segnata dall'intelligenza artificiale.

Con l'auspicio che il monito di Franklin Delano Roosevelt,

"*The test of our progress is not whether we add more to the abundance of those who have much, it is whether we provide enough for those who have too little*",[1]

scolpito nel marmo del Roosevelt Memorial a Washington D.C., possa illuminarci nell'assicurare una maggiore democratizzazione dell'AI, rendendo beneficiari di questa straordinaria rivoluzione tecnologica e sociale la più ampia frazione possibile di cittadini del nostro pianeta.

[1] La prova del nostro progresso non sta nell'aggiungere di più all'abbondanza di coloro che hanno molto, ma nel riuscire a garantire abbastanza a coloro che hanno troppo poco.

ADDENDUM

Una breve digressione matematica su complessità, riproducibilità e spiegabilità dell'AI

Prima di concludere, il lettore mi consentirà di fare alcune rapide considerazioni che puntualizzano alcuni aspetti matematici dell'AI. Abbiamo visto nel capitolo 3 che una componente essenziale del processo di apprendimento è la scelta del modello, ovvero della funzione f che mappa gli input (dati di addestramento) negli output (cioè la risposta alla nostra domanda). Che si tratti, ad esempio, di un processo di minimi quadrati (Quarteroni *et al*. 2014) o di una rete neurale artificiale, il modello f è definito da un insieme di parametri e, possibilmente, iperparametri. In entrambi i casi, i parametri vengono determinati attraverso un processo di minimizzazione di una funzione appropriata J (la funzione obiettivo, o funzione costo, o *loss function*). Anche un metodo di minimi quadrati, quindi, da un punto di vista strettamente matematico, potrebbe essere considerato un processo di apprendimento, ovvero un caso particolare di machine learning. Tuttavia, esistono alcune differenze tra gli approcci ai minimi quadrati comunemente usati e le reti neurali artificiali con addestramento supervisionato, che meritano di essere sottolineate.

Nel caso dei metodi lineari di minimi quadrati, cioè quando il modello f dipende linearmente dai parametri, l'algoritmo di minimizzazione è *deterministico*: il gradiente della funzione costo viene calcolato analiticamente, e il suo annullamento ($\nabla J = 0$) conduce a un sistema lineare (le cosiddette equazioni normali), che verrà poi risolto utilizzando tecniche algebriche. Nel caso dei

metodi non lineari di minimi quadrati, *f* dipende non linearmente dai parametri e, di conseguenza, le equazioni normali risultanti sono non lineari e devono essere risolte mediante un metodo iterativo che, tipicamente, funziona bene su sistemi di piccole dimensioni.

Se il modello di apprendimento si basa su reti neurali artificiali, a causa della struttura compositiva complessa della funzione-modello *f*, l'algoritmo di minimizzazione utilizza tipicamente il *backpropagation* per il calcolo del gradiente (poiché la rappresentazione analitica di ∇J non è più disponibile) e impiega solitamente un algoritmo stocastico di minimizzazione iterativa basato sul gradiente. Pertanto, la natura deterministica del processo viene meno. Tuttavia, è importante notare che una volta addestrata (attraverso il processo di minimizzazione), anche una rete neurale artificiale è deterministica nel senso che l'algoritmo associato può essere descritto in modo univoco e, dato lo stesso set di dati in ingresso, produrrà sempre lo stesso output. Di conseguenza, conserva la proprietà di *riproducibilità*.[1]

Un'altra considerazione riguarda la *dimensione* del problema. Tipicamente, un approccio ai minimi quadrati (sia lineare che non lineare) si basa su un numero molto limitato di parametri, a differenza delle reti neurali artificiali, che coinvolgono un numero straordinariamente elevato di parametri, così come un set di addestramento generalmente molto più ampio rispetto al set di dati di input utilizzato per i metodi dei minimi quadrati. I numerosi parametri, insieme alla struttura altamente non lineare e compositiva delle reti neurali, ne potenziano la capacità di rappresentare processi input-output altamente complessi e multidimensionali, come nel caso delle Convolutional Neural Networks per il riconoscimento delle immagini, ad esempio ResNet-50 (con 25 milioni di parametri) e AlexNet (60 milioni di parametri).

Un altro aspetto frequentemente dibattuto è la presunta mancanza di *interpretabilità* di un algoritmo di machine learning basato su modelli di reti neurali. In senso stretto, proprio come

nel caso dei minimi quadrati, i modelli basati su reti neurali sono in qualche modo interpretabili una volta completato l'addestramento: una volta determinati i parametri e gli iperparametri, la funzione di trasferimento input-output può essere rappresentata in termini finiti e univoci. Naturalmente, va riconosciuto che l'interpretabilità di tale funzione può essere altamente problematica a causa della sua forma compositiva. Sotto questo aspetto, le funzioni modello comunemente usate nei metodi di minimi quadrati sono decisamente più leggibili e interpretabili. In altri termini, identificare il ruolo di ciascun parametro nel modellare la risposta di una rete neurale è generalmente molto impegnativo, se non addirittura impossibile.

Tuttavia, questa apparente inferiorità dei metodi di machine learning merita una riflessione più approfondita. Sviluppiamola attraverso un esempio. Supponiamo di voler scoprire la legge costitutiva di un nuovo materiale, o di un tessuto biologico, ovvero una funzione che leghi fra loro sforzi applicati e deformazioni corrispondenti, a partire da una distribuzione di misure (o osservazioni). Se usassimo un metodo lineare ai minimi quadrati (ovvero la classica regressione lineare), questa legge sarebbe rappresentata da una retta. Una volta trovati i suoi due parametri – grazie alla risoluzione di un semplice sistema algebrico a due equazioni, quello delle equazioni normali – il loro significato sarebbe di facile interpretazione: la pendenza della retta è una misura della rigidità del materiale, mentre l'intercetta con l'asse delle ascisse rappresenta lo stress residuo del materiale a deformazione nulla. Se invece usassimo una rete neurale artificiale per rappresentare la legge sforzi-deformazioni, l'interpretazione del significato dei parametri (tanti, in questo caso!) sarebbe assai più problematica, verosimilmente impossibile, poiché la relazione tra l'input (gli sforzi) e l'output (le deformazioni) dipende implicitamente dall'intera struttura della rete neurale. Tuttavia, questa mancanza di interpretabilità è il prezzo da pagare affinché la rete neurale possa imparare autonomamente la vera struttura della

legge costitutiva: non avendo a priori “forzato” la legge a comportarsi linearmente (cosa che potrebbe essere del tutto inappropriata per certi materiali!), l’abbiamo lasciata libera di far emergere un comportamento probabilmente molto più adatto a rappresentare il materiale in esame. Stiamo, plasticamente, dando un senso a una delle affermazioni fatte sin dall’inizio di questo libro: la caratteristica saliente dei modelli di machine learning consiste proprio nella loro capacità di *imparare autonomamente*, senza essere condizionati dall’utilizzatore!

È tuttavia doveroso notare che la distinzione tra modelli interpretabili, diciamo white-box, e modelli black-box non è assoluta. Esiste uno spettro di modelli, da quelli più semplici e interpretabili ad altri più complessi, ricchi di capacità descrittive ma più difficili da interpretare. In questa transizione, vi è un compromesso tra interpretabilità e complessità del modello. Mentre i modelli più semplici offrono relazioni chiare e comprensibili tra input e output, spesso non riescono a catturare pattern complessi nei dati. Al contrario, modelli più complessi, come le reti neurali, sono in grado di gestire relazioni altamente non lineari e multidimensionali, ma a costo di sacrificare interpretabilità. Trovare il giusto equilibrio tra queste due caratteristiche è una considerazione chiave nella scelta del modello più adatto a risolvere un determinato problema. Il scientific machine learning può essere rivisitato anche sotto questa luce, proponendosi come contesto metodologico propizio a realizzare un efficace bilanciamento fra interpretabilità e capacità di apprendimento in modo autonomo.

Queste riflessioni riprendono nel merito alcune conclusioni tratte nel lavoro Quarteroni et al. (2025).

[1] Considerazioni a parte meritano i casi di apprendimento per rinforzo, reti neurali antagoniste e algoritmi di AI generativa, in cui il determinismo e l’interpretabilità vengono persi.

GLOSSARIO

Albero decisionale (Decision Tree): Algoritmo di apprendimento automatico supervisionato ampiamente utilizzato che trova applicazioni in compiti come la classificazione e la regressione. Adotta una struttura simile ad un albero, in cui i nodi interni rappresentano le caratteristiche o attributi, i rami rappresentano regole decisionali basate su tali attributi e i nodi foglia corrispondono a risultati o previsioni.

Algoritmo: Sequenza finita di istruzioni ben definite, finalizzate alla risoluzione di un problema o di una classe di problemi. Un algoritmo può essere tradotto in software (attraverso opportuni linguaggi di programmazione) ed essere così eseguito da un computer. Gli algoritmi si classificano in deterministici: dato un determinato input, essi producono sempre lo stesso output, e non deterministici: anche per input identici, diverse esecuzioni possono produrre diversi output. Il comportamento degli algoritmi non deterministici dipende dalla generazione di numeri casuali (*random numbers*).

Artificial Neural Network (ANN, Rete Neurale Artificiale): Un modello altamente astratto e semplificato del cervello umano utilizzato nel Machine Learning (ML). Un insieme di unità (neuroni di input) riceve un dato (ad esempio, i pixel di una foto), esegue semplici calcoli su di essi e li passa allo strato successivo di unità (neuroni intermedi). Lo strato finale (neuroni di output) fornisce la risposta (ad esempio, l'oggetto contenuto nell'immagine). I neuroni intermedi sono spesso organizzati in strati (*layers*). Il comportamento della ANN dipende da un insieme di parametri, detti pesi e bias, che vengono opportunamente adattati durante l'addestramento (training) della ANN. Si vedano le figure 4.2 e 4.4.

Autoencoder: Tipo di ANN utilizzato per apprendere codifiche efficienti (ossia rappresentazioni basso-dimensionali) di dati alto-dimensionali. Gli autoencoder rappresentano un esempio di *unsupervised* ML. Si veda la figura 4.7.

Backpropagation (propagazione all'indietro): Algoritmo alla base dell'addestramento delle reti neurali. Esso permette di calcolare quanto e in che modo l'output della rete neurale dipende dai pesi e dai bias associati alla

rete. Più precisamente, l'algoritmo di *backpropagation* permette di calcolare il gradiente della *loss function* rispetto ai parametri della rete neurale in modo molto efficiente.

Black-box (scatola nera): Sistema di AI che riceve un input e fornisce un output, attraverso calcoli di non facile interpretazione per l'essere umano.

Cloud computing: Paradigma di design di strutture informatiche e di erogazione di servizi che prevede la disponibilità di risorse on-demand (tipicamente risorse di calcolo e di storage). I servizi sono erogati per mezzo di server, spesso ridondanti e distribuiti geograficamente per garantire continuità di servizio, in maniera del tutto trasparente all'utente finale. La possibilità di scalare facilmente le risorse (spesso in modo automatico, senza bisogno di interazioni da parte dell'utente) è fra le chiavi di successo del paradigma cloud, specialmente in ambito dell'IoT (Internet of Things) e dell'AI.

Convolutional Neural Network (CNN): Un tipo di ANN, la cui architettura è spirata all'organizzazione della corteccia visiva animale. Le CNN sono molto utilizzate per l'elaborazione di immagini. Si veda la figura 4.5.

Data-driven: Un algoritmo sviluppato senza sfruttare alcuna conoscenza a priori, quali principi primi o leggi empiriche, ma basandosi unicamente su dati. In questo caso, si dice che l'algoritmo non è consapevole della fisica (*physics-informed*) sottostante il problema in esame.

Deep Learning (DL, Apprendimento Profondo): Famiglia di algoritmi di Machine Learning basati su reti neurali con un alto numero di strati (*layer*). Con l'aumentare della profondità della rete, essa diventa più adatta a descrivere modelli progressivamente più astratti. Secondo una comune interpretazione, nell'analizzare ad esempio la foto di un cane, i primi strati permetterebbero di identificare i bordi, quelli successivi elementi quali occhi, naso e zampe e infine gli ultimi strati l'intero animale.

Digital Twin (gemello digitale): Un insieme di costrutti di informazioni virtuali che mimano struttura, contesto e comportamento di un individuo (o un asset fisico), aggiornato dinamicamente grazie ai dati che gli derivano dal suo physical twin durante l'intero ciclo di vita e grazie a decisioni informate che generano valore (fonte: AIAA Institute Position Paper 2020). Si veda la figura 11.3.

Expert System (Sistema Esperto): Una forma di AI che tenta di replicare l'esperienza umana in un'area, come la diagnosi medica o la giurisprudenza. Combina una base di conoscenza con un insieme di regole codificate a mano. Le tecniche di apprendimento automatico stanno progressivamente sostituendo la codifica manuale.

Explainable Artificial Intelligence: Una modello di AI che mira, oltre a fornire risposte a un dato interrogativo, a spiegare perché l'AI è arrivata a compiere

determinate scelte. Essa, a differenza dell'AI versione black-box, mira a una sempre maggior trasparenza e imparzialità, fattori chiave per la diffusione dell'AI in ambiti sensibili come quelli della sicurezza e della medicina.

Gaussian Process: Un tipo di *supervised* ML basato su un modello statistico che assume correlazione spaziale nei dati. Rispetto ad altri modelli di ML, i Gaussian Processes hanno il vantaggio di fornire una stima dell'incertezza della previsione.

Generative Adversarial Network (GAN): Una coppia di reti neurali addestrate congiuntamente. La prima (generatore) genera dati realistici, mentre la seconda (discriminatore) cerca di distinguere i dati generati sinteticamente (fake data) dal generatore da dati reali (real data). L'addestramento delle due reti migliora attraverso la loro competizione. Ad esempio, il generatore potrebbe produrre volti umani realistici, oppure opere d'arte che imitano opere reali; viceversa, il discriminatore – una volta addestrato – potrebbe essere utilizzato per riconoscere un volto umano o lo stile di un determinato artista.

Generative AI: Un tipo di AI che è in grado di generare testo, codici software, immagini, segnali audio o altri media, in risposta a delle richieste tipicamente espresse in linguaggio umano. Tali richieste sono dette "prompt".

Graph Neural Networks: reti progettate per operare su dati come grafi, in cui i nodi rappresentano entità e gli archi che congiungono due nodi adiacenti rappresentano relazioni. Si veda la figura 4.8.

Iperparametri: Insieme di variabili numeriche che caratterizzano e controllano il processo di addestramento (training). A differenza dei parametri, gli iperparametri non variano durante il processo di training. Per una ANN, gli iperparametri sono rappresentati dal numero di *layer*, dal numero di neuroni per ogni *layer*, dalla funzione di attivazione.

Large Language Models (LLM): Sistemi di AI progettati per comprendere e generare testo in modo coerente e contestualmente rilevante. Questi modelli utilizzano vasti insiemi di dati preesistenti per apprendere il linguaggio naturale e sono ampiamente impiegati in applicazioni quali l'elaborazione del linguaggio naturale, la generazione di testi creativi e la risposta a domande basate sul contesto.

Long Short-Term Memory (LSTM): Un tipo di ANN che presenta connessioni di tipo feedback: il segnale di output viene dato nuovamente in input in modo iterativo alla rete stessa. Le LSTM trovano applicazioni in contesti che presentano una dinamica temporale, come l'elaborazione del linguaggio scritto o parlato (NLP) e l'analisi di serie storiche.

Loss Function (o Funzione Costo, Funzione Obiettivo): Funzione matematica il cui valore viene minimizzato durante il processo di addestramento (training). La

loss function rappresenta tipicamente una misura dell'errore commesso dal modello di ML (in altre parole, la differenza fra previsioni del modello e dati reali). Il training può essere visto quindi come il processo attraverso il quale il modello adatta progressivamente i propri parametri per diminuire il più possibile l'errore commesso.

Metaverso: Termine coniato dall'autore di libri di fantascienza Neal Town Stephenson per indicare un universo virtuale, parallelo a quello reale, in cui ciascun individuo è rappresentato dal proprio avatar. In tempi recenti il termine è stato popolarizzato dalla società Facebook (ribattezzata Meta Platforms nel 2021), per indicare una realtà virtuale condivisa per mezzo di Internet.

Natural Language Processing (NLP, *Elaborazione del Linguaggio Naturale)*: Il tentativo di un computer di comprendere il linguaggio parlato o scritto. Ad oggi, gli algoritmi di maggior successo nel campo NLP sono di tipo ML (basati ad esempio su reti LSTM).

Overfitting: Situazione nella quale il modello si adatta in modo ottimale ai dati di training, ma è poco predittivo per dati non osservati. In questo caso, si dice che il modello ha cattive proprietà di generalizzazione. Diverse tecniche sono state sviluppate per evitare l'overfitting, come ridurre le dimensioni del modello (per le ANN, riducendo il numero di neuroni), introducendo termini aggiuntivi (detti di regolarizzazione) nella *loss function*, oppure ancora agendo sull'algoritmo di ottimizzazione.

Parametri: Insieme di variabili numeriche che caratterizzano il funzionamento di un dato modello di ML. Essi vengono modificati durante il processo di addestramento (training), attraverso opportuni algoritmi di ottimizzazione (a differenza degli iperparametri, che invece non vengono variati). Ad esempio, in una ANN, i parametri sono rappresentati dai cosiddetti pesi e bias, che corrispondono rispettivamente all'importanza dei legami fra neuroni diversi e alla sensibilità di ciascun neurone agli input che esso riceve.

Physics-based: Algoritmo o modello che, a differenza delle controparti data-driven, è costruito sfruttando conoscenze a priori (come leggi fisiche o modelli matematici). Esso incorpora in modo innato i concetti di spazio, tempo e causalità.

Prompt Engineer: Nuova professione che si sta affermando con l'emergere dell'AI generativa e che consiste nel convertire uno o più task in istruzioni comprensibili dal modello di linguaggio, e di rifinire l'input in modo da generare output desiderati, nella forma – ad esempio – di testo, immagini o codici.

Quantum computing: Paradigma computazionale che trae vantaggio dai fenomeni quantistici, quale la sovrapposizione di stati e l'interferenza. Se

l'elemento fondamentale della computazione tradizionale è il bit, caratterizzato da uno stato binario (0 o 1), l'elemento base dei computer quantistici è il qbit, il cui stato è in realtà costituito da una sovrapposizione di stati. Nonostante oggi i computer quantistici non siano in grado di superare le performance di quelli tradizionali, fatta eccezione per ambiti molto specifici, si crede che in futuro essi possano risultare più veloci in applicazioni pratiche come la crittografia, una tecnologia di fondamentale importanza per i sistemi di IoT e di AI.

Recurrent Neural Network (RNN): Rete neurale artificiale progettata specificamente per elaborare dati sequenziali, come serie temporali, testo, parlato o fotogrammi video. A differenza delle reti neurali *feed-forward* tradizionali, le RNN presentano cicli che consentono di trasferire le informazioni da uno step all'altro, conferendo loro una sorta di memoria. Questo le rende particolarmente adatte per compiti in cui il contesto precedente è essenziale, come la modellazione del linguaggio, il riconoscimento vocale e la previsione finanziaria. Si veda la figura 4.6.

Reduced Order Model (Modello ridotto): Modello matematico di basso costo computazionale, derivato a partire da una modello matematico di alta complessità e alta fedeltà (detto *full-order model* o *high-fidelity model*). Il modello ridotto permette di surrogare il modello *high-fidelity*, approssimandone le previsioni in tempi molto più rapidi e con un minor impatto in termini di risorse computazionali. I modelli ridotti sono fra i fattori abilitanti per i *digital twin*, che necessitano di interazioni in tempo reale con l'ambiente fisico.

Reinforcement Learning (Apprendimento con rinforzo): Un tipo di ML in cui l'algoritmo impara agendo verso un obiettivo astratto, come "guadagnare un punteggio elevato in un videogioco" o "gestire una fabbrica in modo efficiente". Durante l'addestramento, ogni sforzo viene valutato in base al suo contributo verso l'obiettivo.

Retrieval-Augmented Generation (RAG): Si tratta di un approccio innovativo che combina un componente di recupero delle informazioni con un modello generativo di testo (vedi la voce LLM). Utilizza fonti esterne, come documenti da Wikipedia, per arricchire il contesto del modello generativo, permettendo un'adattabilità dinamica alle informazioni in evoluzione senza richiedere una riprogrammazione completa del modello.

Scientific Computing (Calcolo Scientifico): Algoritmi volti a sfruttare la potenza di calcolo dei computer per risolvere problemi matematici inaccessibili all'uomo (per ragioni di tempo o di elevata complessità), quelli generalmente basati sulle leggi fisiche (physics-based). Tramite la risoluzione (eventualmente approssimata) di modelli matematici, è possibile – fra le altre cose – simulare processi fisici ed effettuare previsioni e analisi di scenario.

Scientific Machine Learning: Nuova disciplina che mira a integrare in modo efficace gli algoritmi data-driven dell'AI con quelli physics-based del calcolo scientifico. Una promessa per una nuova grande scienza.

Supervised Learning (Apprendimento Supervisionato): Un tipo di ML in cui l'algoritmo, durante l'addestramento, confronta i suoi risultati con i risultati corretti (detti spesso labels). Chiaramente, questo approccio è possibile sono quando le labels sono disponibili.

Support Vector Machine: Un tipo di *supervised* ML utilizzato principalmente per problemi di classificazione, ossia quando si vuole assegnare a ciascun input un'etichetta scelta da un insieme discreto (ad esempio, data la foto di un prodotto, stabilire se essa contiene un paio di scarpe, un cappotto oppure una cintura).

Transfer Learning: Una tecnica di apprendimento automatico per cui un algoritmo impara a eseguire un compito (come riconoscere animali), e si basa su quella conoscenza quando impara un compito diverso ma correlato (come riconoscere uomini).

Transformer: Un'architettura di rete neurale introdotta da Vaswani *et al*. nel 2017, rivoluzionando il campo del deep learning. Si basa su meccanismi di attenzione, consentendo una gestione efficace delle relazioni a lungo termine nei dati, rendendolo ampiamente utilizzato in applicazioni come il trattamento del linguaggio naturale e la visione artificiale. Si veda la figura 6.1.

Turing Test: Un test sulla capacità dell'AI di essere indistinguibile (per un osservatore umano) dall'intelligenza umana. Nella concezione originale di Alan Turing, un'AI sarebbe stata giudicata per le sue capacità di conversare con un umano attraverso un testo scritto.

RIFERIMENTI BIBLIOGRAFICI

AIAA, *AIAA Institute Position Paper on Digital Twin*, dicembre 2020.

Álvarez J.S.V., *The Risks and Inefficacies of AI Systems in Military Targeting Support*, International Committee of the Red Cross, 4 settembre 2024.

Anderson C., *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, "Wired", 23 giugno (2008), https://www.wired.com/2008/06/pb-theory/.

Appel K. e Haken W., *Every Planar Map is Four Colorable*, "American Mathematical Society", vol. 67 (1989).

Batty B., *Digital Twins*. "Environment and Planning B: Urban Analytics and City Science", vol. 45, n. 5, 817-20 (2018).

Beiker S., *Tesla's Use of Digital Twins for Predictive Maintenance and Autonomy*, SAE International, 2020.

Bode I. e Bhila I., *The Problem of Algorithmic Bias in AI-Based Military Decision Support Systems*, International Committee of the Red Cross, 3 settembre, 2024.

Bostrom N., *Superintelligenza. Tendenze, pericoli, strategie*, Bollati Boringhieri, Torino 2014.

Caforio F., Regazzoni F., Pagani S., Karabelas E., Augustin C., Haase G., Plank G. e Quarteroni A., *Physics-Informed Neural Network Estimation of Material Properties in Soft Tissue Nonlinear Biomechanical Models*, "Computational Mechanics", vol. 75, n. 2, 487-513 (2024).

Chomsky N., *Knowledge of Language*, Praeger, New York 1986.

Copeland B.J., *Alan Turing*, "Encyclopedia Britannica", 23 luglio (2024), https://www.britannica.com/biography/Alan-Turing.

Corral-Acero J., Margara F. *et al*., *The "Digital Twin" to Enable the Vision of Precision Cardiology*, "European Heart Journal", vol. 41, n. 48, 4556-64 (2020).

Dijkstra E.W., *A Note on Two Problems in Connexion with Graphs*, "Numerische Mathematik", vol. 1, 269-71 (1959), doi:10.1007/BF01386390. S2CID123284777.

Fedele M., Piersanti R., Regazzoni F., Salvador M., Africa P.C., Bucelli M., Zingaro A., Dede' L. e Quarteroni A., *A Comprehensive and Biophysically*

Detailed Computational Model of the Electromechanics of the Whole Human Heart, "CMAME", n. 410 (2023), https://doi.org/10.1016/j.cma.2023.115983.

Gao Y., Xiong Y., Gao X., Jia K., Pan J., Bi Y., ... e Wang H., *Retrieval-Augmented Generation for Large Language Models: A Survey*, arXiv preprint arXiv:2312.10997, 2023.

Green A. e Lamby L., *The Supply, Demand and Characteristics of the AI Workforce across OECD Countries*, "OECD Social, Employment and Migration Working Papers", No. 287, OECD Publishing, Paris 2023.

Hales T.C., *A Proof of the Kepler Conjecture*, "Discrete & Computational Geometry", vol. 36, n. 1, 21-69 (2003).

Harari Y.N., *21 lezioni per il XXI secolo*, Bompiani, Milano 2018

Hart P.E., Nilsson N.J., Raphael B., *A Formal Basis for the Heuristic Determination of Minimum Cost Paths*, "IEEE Transactions on Systems Science and Cybernetics SSC4", vol. 2, 100-107 (1968).

Hopfield J.J., *Neural Networks and Physical Systems with Emergent Collective Computational Abilities*, "Proceedings of the National Academy of Sciences", vol. 72, n. 8, 2554-8 (1982).

Jordan M.I. e Mitchell T.M., *Machine Learning: Trends, Perspectives, and Prospects*, "Science", vol. 349, n. 6245, 255-60 (luglio 2015).

Kissinger H., Schmidt E. e Huttenlocher D., *The Age of AI*, John Murray, London 2021.

Klaus M., *Transcending Weapon Systems: The Ethical Challenges of AI in Military Decision Support Systems*, International Committee of the Red Cross, 24 settembre, 2024.

Landgrebe J. e Smith B. *Why Machines Will Never Rule the World*, Routledge, Taylor and Francis Group, New York 2022.

Lane M., Williams M. e Broecke S., *The Impact of AI on the Workplace: Main Findings from the OECD AI Surveys of Employers and Workers*, "OECD Social, Employment and Migration Working Papers", No. 288, OECD Publishing, Paris 2023.

Laubenbacher R., Mehrad B., Shmulevich I. *et al*., *Digital Twins in Medicine*., "Nat. Comput. Sci.", n. 4, 184-91 (2024). https://doi.org/10.1038/s43588-024-00607-6.

Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., ... e Kiela D., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, "Advances in Neural Information Processing Systems", 33, 9459-74 (2020).

Lu X., Sclar M., Hallinan S., Mireshghallah N., Liu J., Han S., Ettinger A., Jiang L., Chandu K., Dziri N. e Choi Y., *AI as Humanity's Salieri: Quantifying Linguistic Creativity of Language Models via Systematic Attribution of Machine Text Against Web Test*, "OpenReview.net" (2024).

Marchant J., *Powerful Antibiotics Discovered Using AI*, “Nature” (20 febbraio 2020), https://www.nature.com/articles/d41586-020-00018-3.

Maslej N., Fattorini L., Perrault R., Parli V., Reuel A., Brynjolfsson E., Etchemendy J., Ligett K., Lyons T., Manyika J., Niebles J.C., Shoham Y., Wald R. e Clark J., *The AI Index 2024 Annual Report*, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA 2024.

McCulloch W.S. e Pitts W.H., *A Logical Calculus of the Ideas Immanent in Nervous Activity*, “Bulletin of Mathematical Biophysics”, vol. 5, 115-33 (1943).

Messeri L. e Crockett M., *Artificial Intelligence and Illusions of Understanding in Scientific Research*, “Nature”, vol. 627, n. 8002, 49-58 (2024).

Minsky M., *The Society of Mind*, Simon & Schuster, New York 1986.

Moravec H.P., *Mind Children: The Future of Robot and Human Intelligence*, Harvard University Press, Cambridge 1988.

Parolini N., Ardenghi G., Dede’ L. e Quarteroni A., *A Mathematical Dashboard for the Analysis of Italian COVID-19 Epidemic Data*, “Int. J. Numer. Meth. Biomed. Engineering”, vol. 37, n. 9 (2021a), https://doi.org/10.1002/cnm.3513.

Parolini N., Dede’ L., Antonietti P.F., Ardenghi G., Miglio E., Manzoni A., Pugliese A., Verani M. e Quarteroni A., *SUIHTER: A New Mathematical Model for COVID-19. Application to the Analysis of the Second Epidemic Outbreak in Italy*, “Proceedings of the Royal Society A”, vol. 477, n. 2253 (2021b), https://doi.org/10.1098/rspa.2021.0027.

Parolini N., Dede’ L., Ardenghi G. e Quarteroni A., *Modelling the COVID-19 and the Vaccination Campaign in Italy by the SUIHTER Model*, “Infectious Disease Modelling”, vol. 7, n. 2, 45-63 (2022).

Petroni A., *Le ragioni della scienza pura nell’era dell’Intelligenza Artificiale*, “SEGNATURE”, Accademia Nazionale dei Lincei, 8 novembre 2023.

Quarteroni A., *Modellistica numerica per problemi differenziali*, Springer-Verlag Italia, Milano 2000 (6ª ed. 2016).

Quarteroni A., *Taking Mathematics to Heart*, “Notices of the American Mathematical Society”, vol. 65, n. 1, 19-65 (2018).

Quarteroni A., *Le equazioni del cuore, della pioggia e delle vele*, Chiavi di Lettura, Zanichelli, Bologna 2020.

Quarteroni A., Dede’ L., Manzoni A. e Vergara C., *Mathematical Modelling of the Human Cardiovascular System. Data, Numerical Approximation, Clinical Applications*, Cambridge University Press, Cambridge 2019.

Quarteroni A., Gervasio P. e Regazzoni F., *Combining Physics-Based and Data-Driven Models: Advancing the Frontiers of Research with Scientific Machine Learning*, M3AS, 2025.

Quarteroni A., Saleri F. e Gervasio P., *Calcolo Scientifico. Esercizi e problemi risolti con MATLAB e Octave*, Springer-Verlag Italia, Milano 2008 (5a ed. 2012).

Raissi M., Perdikaris P. e Karniadakis G.E., *Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations*, "Journal of Computational Physics", vol. 378, 686-707 (2021).

Regazzoni F., Dede' L. e Quarteroni A., *Machine Learning of Multiscale Active Force Generation Models for the Efficient Simulation of Cardiac Electromechanics*, "CMAME", vol. 370 (2020), https://doi.org/10.1016/j.cma.2020.113268.

Regazzoni F., Pagani S. e Quarteroni A., *Universal Solution Manifold Networks (USM-Nets): Non-Intrusive Mesh-Free Surrogate Models for Problems in Variable Domains*, "Journal of Biomechanical Engineering", vol. 144, n. 12 (2022).

Reinsel D., Gantz J. e Rydning J. (2017), *Data Age 2025*, IDC White Paper, aprile 2017.

Rosenblatt F., *The Perceptron: A Probabilistic Model for Information Storage and Organization on the Brain*, "Psychological Review", vol. 65, n. 6, 386-408 (1958).

Roser M., *AI Timelines: What Do Experts in Artificial Intelligence Expect for the Future?*, pubblicato online nel 2023 da OurWorldinData.org e disponibile al link https://ourworldindata.org/ai-timelines.

Salakhutdinov R. e Hinton G.E. (2009), *Deep Boltzman Machines*, "Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)", Clearwater Beach, Florida, USA, 2009, Volume 5 of JMLR:W&CP5.

Searle J.R., *Minds, Brains and Programs*, "The Behavioral and Brain Sciences", vol. 3, n. 3, 417-24 (1980).

Searle J.R., *La mente è un programma?*, "Le scienze", n. 259, 16-21 (1990).

Siciliano B., *Robots Are with Us, Within Us and Among Us*, "Design and Technologies – Design, Robotics and Machines in the Post-Human Age", 67 (2019).

Singapore (Singapore Government), *Virtual Singapore: A Digital Twin for the City*, 2021, reperibile online al link www.nrf.gov.sg.

Singh S., *Fermat's Enigma: The Epic Quest to Solve the World's Greatest Mathematical Problem*, Walker & Co., New York 1997.

Szapiro G.G., *Kepler's Conjecture*, John Wiley and Sons, New York 2003.

Togelius J., *Artificial General Intelligence*, MIT Press, Cambridge, MA 2024.

Turing A., *Computing Machinery and Intelligence*, "Mind", vol. 59, n. 236, 433-46 (1950).

Viceconti M.,. Henney A e Morley-Fletcher E., *In Silico Clinical Trials: How Computer Simulation Will Transform the Biomedical Industry*, "International

Journal of Clinical Trials", vol. 3, n. 2, 37-46 (2016).

Whitehead A.N. e Russell B., *Principia Mathematrica*, Cambridge University Press, Cambridge 1910-1913.

Wiles A., *Modular Elliptic Curves and Fermat's Last Theorem*, "Annals of Mathematics", vol. 141, n. 3 (1995).

Wilson R., *Four Colors Suffice: How the Map Problem Was Solved*, Princeton University Press, Princeton 2013.

Zhang D., Maslej N., Brynjolfsson E., Etchemendy J., Lyons T., Manyika J. *et al*., *The AI Index 2022 Annual Report*, AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University, Stanford 2022.

Zhou W. e Greipl A.R., *Artificial Intelligence in Military Decision-Making: Supporting Humans, Not Replacing Them*, International Committee of the Red Cross, 29 agosto 2024.

L’AUTORE

Alfio Quarteroni, uno dei più eminenti scienziati italiani e mondiali nel campo della matematica, è professore presso University of Minnesota e presso EPFL di Losanna, professore emerito presso il Politecnico di Milano e l’EPFL di Losanna, fondatore del MOX del Politecnico di Milano e presidente dello spinoff MOXOFF, membro di diverse Accademie Scientifiche europee e vincitore di numerosi prestigiosi premi internazionali. Ha diretto il team che ha sviluppato il primo modello matematico di cuore umano e quello che ha condotto le simulazioni matematiche di Alinghi, vincitrice di due edizioni dell’America’s Cup. È autore di circa 30 libri di ricerca scientifica.

INFORMAZIONI SUL LIBRO

L’obiettivo di Alfio Quarteroni è sfatare miti e preconcetti, spiegando una rivoluzione tecnologica dall’enorme potenziale, i successi già raggiunti, le speranze che alimenta, ma anche i rischi che presenta, non solo per gli individui, ma per la società nel suo complesso. Senza celebrazioni o demonizzazioni, Quarteroni considera le implicazioni dell’AI e fornisce strumenti per comprendere meglio questo fenomeno, promuovendo un approccio critico e consapevole nei confronti della trasformazione che stiamo vivendo.