ATENEO VENETO

Rivista di scienze, lettere ed arti
Atti e memorie dell'Ateneo Veneto

1 8 1 2

ATENEO VENETO onlus
Istituto di scienze, lettere ed arti
fondato nel 1812
213° anno accademico

Campo San Fantin 1897, 30124 Venezia
tel. 0415224459
http://www.ateneoveneto.org

ATENEO VENETO
Rivista semestrale di scienze, lettere ed arti
Atti e memorie dell'Ateneo Veneto
CCXII, terza serie 24/II (2025)

Autorizzazione del presidente
del Tribunale di Venezia,
decreto n. 203, 25 gennaio 1960
ISSN: 0004-6558
iscrizione al R.O.C. al n. 10161

direttore responsabile
Michele Gottardi
direttore scientifico
Gianmario Guidarelli
segreteria di redazione
Silva Menetto, Carlo Federico Dall'Omo
e-mail
rivista@ateneoveneto.org

Progetto grafico e impaginazione
Livio Cassese

Stampa
Grafiche Veneziane soc. coop.
Spedizione in abbonamento

REGIONE DEL VENETO

Iniziativa regionale realizzata in attuazione
della L.R. n. 17/2019 – art. 32

INDICE

Alessandro Sperduti
UNIVERSITÀ DEGLI STUDI DI PADOVA

Intelligenza Artificiale e medicina

Definire il concetto di intelligenza artificiale (AI) non è compito semplice, sopratutto considerando che la definizione stessa di intelligenza umana è oggetto di dibattito: coinvolge dimensioni cognitive, emotive, sociali e creative difficili da ricondurre a una formulazione univoca. Non sorprende, dunque, che anche la nozione di intelligenza artificiale risulti complessa e sfaccettata. Ridurla alle tecnologie oggi più visibili – come il deep learning o i modelli linguistici di grandi dimensioni – significa trascurare la ricchezza di approcci che ne hanno segnato l'evoluzione. Già in ambito psicologico, autori come Howard Gardner hanno proposto una visione pluralista dell'intelligenza (teoria delle intelligenze multiple[1]), mentre Daniel Goleman ha enfatizzato la dimensione emotiva[2]. In filosofia della mente, il confronto tra cognitivismo e connessionismo ha ulteriormente mostrato la difficoltà di circoscrivere il fenomeno in termini puramente computazionali. Questo pluralismo teorico si riflette inevitabilmente anche nel campo dell'intelligenza artificiale.

Pertanto non sorprende che l'intelligenza artificiale comprenda tradizioni di ricerca differenti. Accanto all'apprendimento automatico, motore dei progressi più recenti, vi sono molti altri ambiti di ricerca, come la risoluzione dei problemi, il ragionamento automatico, la rappresentazione della conoscenza, la pianificazione. Inoltre, la radice concettuale di alcuni di questi ambiti, ed in particolare il ragiona-

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

2 Daniel Goleman, *Emotional Intelligence*, New York, Bantam Books, 1995.

mento automatico, affondano le loro radici nella logica formale e nella filosofia antica, e mirano a formalizzare inferenze e concetti in modo sistematico. La stessa definizione del campo proposta da John McCarthy negli anni Cinquanta intendeva l'AI come «la scienza e l'ingegneria di produrre macchine intelligenti»[3], lasciando volutamente aperta la questione di cosa si dovesse intendere per «intelligente».

Più recentemente, Stuart Russell e Peter Norvig, nel loro manuale *Artificial Intelligence: A Modern Approach*[4], distinguono quattro prospettive: sistemi che pensano come gli umani, che agiscono come gli umani, che pensano razionalmente e che agiscono razionalmente. Tale classificazione mostra come l'AI non sia un paradigma unitario, bensì un insieme di programmi di ricerca parzialmente sovrapposti.

Storicamente, il termine «intelligenza artificiale» si afferma ufficialmente nel 1956, durante il workshop *Dartmouth Summer Research Project*[5] organizzato da John McCarthy, Marvin Minsky, Nathalien Rochester e Claude Shannon, workshop che vide la partecipazione anche di Allen Newell e Herbert A. Simon. L'obiettivo era comprendere come i computer potessero assistere l'essere umano nella risoluzione di problemi complessi, ipotizzando che ogni aspetto dell'apprendimento o dell'intelligenza potesse essere descritto con sufficiente precisione da essere simulato da una macchina.

Sin dagli esordi, uno dei campi privilegiati di applicazione è stato la medicina. Negli anni Sessanta, il programma *ELIZA*, sviluppato da Joseph Weizenbaum[6], simulava un dialogo terapeutico ispirato alla pratica psicoanalitica, dando l'impressione di un'interazione autentica pur basandosi su regole semplici e schematiche. Negli anni Settanta, sistemi esperti come *MYCIN*[7] affrontavano problemi di diagnosi

3 John McCarthy, *What Is Artificial Intelligence?*, Stanford, Stanford University, 2007, p. 1, cfr. anche: //www-formal.stanford.edu/jmc/whatisai.pdf. Si noti che anche se il documento è del 2007, McCarthy attribuisce la definizione alla concezione originaria del campo negli anni '50.

4 Stuart Russell, Peter Norvig, *Artificial Intelligence: A Modern Approach*, 4ª ed., Harlow, Pearson, 2020.

5 John Mccarthy, Marvin Minsky, Nathaniel Rochester, Claude Shannon, *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, Hanover (NH), Dartmouth College, 1955.

6 Joseph Weizenbaum, *ELIZA – A Computer Program for the Study of Natural Language Communication Between Man and Machine*, Communications of the ACM, Volume 9, Issue 1, pp. 36-45, 1966.

7 Edward H. Shortliffe, *Computer-Based Medical Consultations: MYCIN*, New York, Elsevier, 1976.

e terapia antibiotica, formalizzando la conoscenza di specialisti in regole esplicite e introducendo meccanismi di gestione dell'incertezza tramite «fattori di certezza». Analogamente, nello stesso periodo, fu sviluppato *CASNET* uno dei primi sistemi esperti a utilizzare una *rete causale-associativa* per supportare le decisioni mediche in oftalmologia per la diagnosi del glaucoma[8], mostrando come il computer potesse supportare il medico nell'analisi dei dati clinici. Altro esempio, degli anni Ottanta, è DXplain[9], sistema basato sull'AI per fornire suggerimenti di diagnosi differenziale basati su sintomi clinici.

Questi sistemi, tuttavia, dipendevano fortemente dall'intervento umano: esperti e programmatori dovevano estrarre, codificare e aggiornare la conoscenza in forma simbolica. Inoltre, mancava una reale capacità di percezione autonoma del mondo esterno. Le aspettative spesso superavano i risultati concreti, dando luogo ai cosiddetti «inverni dell'intelligenza artificiale», periodi di disillusione e riduzione dei finanziamenti, come ricostruito da Nils J. Nilsson nella sua storia dell'AI[10].

Negli anni Novanta, ebbe ampia risonanza un sistema AI per il gioco degli scacchi: *Deep Blue*, sviluppato da IBM, che nel 1997 sconfisse Garry Kasparov grazie a un'enorme capacità di calcolo e a strategie di ricerca esaustiva: analizzava un numero vastissimo di possibili mosse per individuare la più vantaggiosa. In questo caso il problema era formalizzato in modo rigoroso e di nuovo risolto senza necessità di interazione con il mondo fisico. Come documentato da Murray Campbell e colleghi[11], il successo fu il risultato di una combinazione di forza bruta computazionale, euristiche sofisticate e basi di dati di partite storiche.

Accanto a tali approcci si colloca la ricerca operativa, che non si limita a trovare una soluzione, ma mira alla soluzione ottima secondo criteri di efficienza o costo. Qui il problema viene tradotto in termini matematici e affrontato con strumenti di ottimizzazione, spesso a

8 Sholom M. Weiss, Casimir A. Kulikowski, Saul Amarel, Aran Safir, *A model-based method for computer-aided medical decision-making*, Artificial Intelligence, 1978, 11(1–2), pp. 145-172.

9 G. Octo Barnett, James J. Cimino, James A. Hupp, Edward P. Hoffer, *DXplain: an evolving diagnostic decision-support system*, JAMA, 1987, 258(1), pp. 67–74.

10 Nils J. Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements*, Cambridge, Cambridge University Press, 2010.

11 Murray Campbell, A. Joseph Hoane, Feng-Hsiung Hsu, *Deep Blue*, Artificial Intelligence, 2002, 134(1–2), pp. 57–83.

prezzo di elevati costi computazionali. Questo filone, sviluppatosi già durante la Seconda Guerra Mondiale, ha fornito strumenti fondamentali anche per l'AI contemporanea, in particolare nei sistemi di pianificazione automatica e nei problemi di scheduling[12]. Anche in questo caso però, l'intervento umano è fondamentale per pervenire ad una formalizzazione del problema da risolvere e per l'inserimento nel computer dei dati che lo caratterizzano.

Un cambiamento radicale rispetto a questa situazione si verifica nel 2012, quando le reti neurali profonde (*Deep Learning*) dimostrano una superiorità netta nel riconoscimento visivo su larga scala nella competizione ImageNet[13]. La rete neurale *AlexNet* di Alex Krizhevsky, Ilya Sutskever e Geoffrey Hinton[14] segna una svolta epocale, riducendo drasticamente il tasso di errore rispetto ai metodi tradizionali di *computer vision*.

Le reti neurali artificiali sono modelli matematici ispirati, in forma semplificata, al funzionamento del cervello umano, secondo una tradizione che risale al modello di Warren McCulloch e Walter Pitts del 1943[15]. Ogni «neurone» artificiale rappresenta un'astrazione semplificata di un neurone biologico, che riceve input, elabora informazioni tramite un'elaborazione semplice e trasmette output attraverso connessioni ponderate ad altre unità. I neuroni biologici elaborano segnali attraverso dendriti e sinapsi, mentre le reti neurali artificiali simulano questa trasmissione mediante connessioni pesate che modulano il contributo di ciascun input. La combinazione di molte unità semplici permette alla rete di compiere elaborazioni complesse, analogamente a quanto avviene nel cervello umano. L'apprendimento della rete avviene adattando i pesi (parametri liberi) delle connessioni in base a esempi input-output noti, ad esempio una immagine in input che rappresenta un gatto e l'etichetta «gatto» fornita come output de-

12 Carla P. Gomes, *Artificial Intelligence and Operations Research: Challenges and Opportunities in Planning and Scheduling*, The Knowledge Engineering Review. 2000;15(1):1-10.

13 Jia Deng, Wei Dong, Richard Socher, Li-Jiao Li, Kai Li, Fei-Fei Li, *ImageNet: A Large-Scale Hierarchical Image Database*, Miami (FL), IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

14 Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*, Vancouver (BC), Advances in Neural Information Processing Systems (NeurIPS), 2012, pp. 1097–1105.

15 Warren S. Mcculloch, Walter Pitts, *A Logical Calculus of Ideas Immanent in Nervous Activity*, *The Bulletin of Mathematical Biophysics*, Baltimore (MD), The Johns Hopkins Press, 1943, pp. 115–133.

siderato. L'output degli esempi è tipicamente fornito da un umano e per questo si parla di apprendimento *supervisionato*. Per il compito di classificazione di immagini su citato, la rete modifica i pesi per associare correttamente immagini a categorie predefinite (es. «auto», «persona», «animale»).

L'uso di più strati di neuroni, tipico delle reti profonde (*deep neural networks*), aumenta la capacità di rappresentazione della rete. Gli strati inferiori codificano caratteristiche di basso livello (tratti e contorni), gli strati intermedi combinano queste caratteristiche in pattern più complessi (parti dell'oggetto) e gli strati superiori rappresentano entità complete presenti nell'immagine. Questa gerarchia di rappresentazioni è stata osservata sperimentalmente anche nel sistema visivo umano, mostrando una sorprendente analogia tra le reti neurali artificiali e le reti neurali biologiche in termini di organizzazione e funzionalità[16].

Grazie alla disponibilità di grandi quantità di dati fornita da *Internet* e alla crescente potenza di calcolo (in particolare tramite GPU), queste reti hanno consentito al computer di «connettersi» al mondo reale: riconoscere volti, oggetti, scene, tracciare movimenti in tempo reale anche in condizioni complesse. Un grosso passo avanti rispetto allo stato dell'arte precedente, perché quando si affronta il gioco degli scacchi di fatto un umano formalizza in maniera matematica il problema e il computer lo risolve senza dover interagire realmente con il mondo esterno. Analogamente, un sistema esperto è tale perché un programmatore umano estrae conoscenza dagli esperti, la codifica in un linguaggio comprensibile dalla macchina e in questo modo permette l'utilizzo della conoscenza estratta a fatica dagli esperti, ma non era mai successo prima di allora che il computer in autonomia potesse percepire direttamente quello che accadeva nel mondo fisico. Ciò ha aperto applicazioni mediche cruciali, ad esempio nello screening delle retinopatie diabetiche e nell'analisi di immagini radiologiche, come ben dimostrato da recenti studi pubblicati su *Nature Communica-*

16 Si veda, ad es., MATTEO RICCI, THOMAS SERRE, *Hierarchical Models of the Visual System*, in *Encyclopedia of Computational Neuroscience*, a cura di Daniel Jaeger, Rüdiger Jung, New York (NY), Springer, 2020.

tions[17] e *The Lancet Digital Health*[18].

Il successo del *Deep Learning* ha però avuto un prezzo: la necessità di enormi quantità di dati etichettati manualmente. Medici, annotatori ed esperti hanno dovuto classificare e segmentare milioni di immagini affinché i modelli apprendessero correttamente. Questo limite ha stimolato la ricerca verso l'apprendimento auto-supervisionato (*self-supervised*) e i cosiddetti *foundation models*, concettualizzati in modo sistematico nel 2021 dal gruppo di ricerca guidato da Percy Liang presso Stanford[19].

L'apprendimento auto-supervisionato permette di utilizzare una parte del dato stesso come supervisione dell'apprendimento. Questo ha permesso di utilizzare dati non etichettati per risolvere uno dei due sotto-problemi, in effetti il più difficile, che l'apprendimento supervisionato deve risolvere. Vediamo di spiegarlo con l'esempio della classificazione delle immagini. Un'immagine digitale è rappresentata da un insieme di *pixel*, ciascuno con valori numerici. Tuttavia, non tutte le possibili combinazioni di valori dei *pixel* generano immagini coerenti o interpretabili, a causa delle leggi fisiche che regolano la luce, la sua riflessione e la sua propagazione. Di conseguenza, nello spazio ad alta dimensionalità di tutte le possibili immagini, esiste un sottospazio relativamente piccolo di immagini valide. Il primo problema nell'apprendimento visivo supervisionato consiste nell'identificare questo sottospazio di immagini plausibili. Una volta riconosciuto, è possibile risolvere il secondo sotto-problema in modo molto più semplice ed efficace, cioè mappare parti di questo sottospazio su categorie specifiche, ad esempio auto, animali o persone. Il riconoscimento del

17 Risa Wolf, Roomasa Channa, Tin Yan Liu, Anum Zehra, Lee Bromberger, Dhruva Patel, Ajaykarthik Ananthakrishnan, Elizabeth Brown, Laura Prichett, Harold Lehmann, Michael Abramoff, *Autonomous Artificial Intelligence Increases Screening and Follow-Up for Diabetic Retinopathy in Youth: The ACCESS Randomized Control Trial*, Nature Communications, Londra, Nature Publishing Group, 2024, vol. 15 (1).

18 Veronica Hernström, Viktoria Josefsson, Hanna Sartor, David Schmidt, Anna-Maria Larsson, Solveig Hofvind, Ingvar Andersson, Aldana Rosso, Oskar Hagberg, Kristina Lång, *Screening Performance and Characteristics of Breast Cancer Detected in the Mammography Screening with Artificial Intelligence Trial (MASAI): a Randomised, Controlled, Parallel-Group, Non-Inferiority, Single-Blinded, Screening Accuracy Study*, Londra, Elsevier (The Lancet Digital Health), 2025, vol. 7, pp. 175–183.

19 Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, Stanford (CA), Stanford University, Center for Research on Foundation Models (CRFM), 2021. DOI: 10.48550/arXiv.2108.07258.

sottospazio è complesso perché lo spazio delle variabili è altamente dimensionale e il numero delle possibili combinazioni dei *pixel* è enorme; pertanto, è necessario un ampio volume di dati per apprendere le configurazioni valide. Tradizionalmente, questo richiedeva immagini etichettate manualmente, per guidare l'apprendimento e implementare funzioni di classificazione.

Recentemente, grazie alla disponibilità di miliardi di immagini non etichettate su *Internet*, è stato possibile apprendere direttamente la distribuzione delle immagini nello spazio dei *pixel* senza la necessità che un umano le etichettasse. Questo ha portato alla nascita dei modelli fondativi (*foundation models*), capaci di catturare il sottospazio di dati valido e generare rappresentazioni utili al computer per realizzare compiti di mappatura in vari domini, riducendo la necessità di supervisione umana. Infatti, per poter apprendere la mappatura desiderata a partire dalle rappresentazioni fornite da un modello fondativo occore una quantità molto minore di dati etichettati da un umano.

Quanto descritto per le immagini è valido per qualunque tipo di dato. Ad esempio, vale anche per il testo. Nel caso del linguaggio, ad esempio, il modello viene addestrato a prevedere come potrebbe continuare una frase. Questo è un esempio concreto di auto-apprendimento, dove il segnale di apprendimento non viene fornito da un annotatore umano, ma è ricavato direttamente dai dati stessi. Cerchiamo di capirlo con un esempio concreto. Si consideri una parola iniziale, ad esempio "oggi", e si analizzi l'insieme delle frasi che possono iniziare con tale termine e le possibili continuazioni. Anche limitandosi a un sottoinsieme ridotto, si osserva che a uno stesso prefisso possono seguire sviluppi differenti: "oggi ho cucinato la pasta", "oggi ho cucinato con piacere", "oggi è una bellissima giornata", "oggi è una tappa importante". La presenza di prefissi comuni con continuazioni diverse evidenzia che il linguaggio naturale può essere descritto come un insieme di sequenze caratterizzate da differenti gradi di plausibilità. Da questo punto di vista, una descrizione adeguata del linguaggio consiste nell'associare una probabilità a ciascuna possibile sequenza di parole. Alcune combinazioni risultano prive di senso e quindi hanno probabilità nulla; altre sono rare e presentano probabilità bassa; altre ancora sono frequenti e quindi più probabili. Se tali probabilità

vengono stimate correttamente, è possibile generare frasi coerenti selezionando le continuazioni più plausibili. Operativamente, si può procedere estraendo un campione di frasi da una collezione di documenti. Supponiamo di selezionare 10.000 frasi. Si può contare quante di esse iniziano con "oggi": ad esempio 83 su 10.000. Analogamente, si può verificare quante proseguono con "oggi ho", ad esempio 33, e così via. In questo modo si stimano le probabilità dei diversi prefissi. Un'analisi più informativa consiste nel calcolare la probabilità di una parola successiva dato un prefisso noto. Se 83 frasi iniziano con "oggi" e, tra queste, 33 continuano con "ho", 31 con "è" e 19 con "mi", il rapporto tra ciascuna frequenza e il totale dei prefissi "oggi" fornisce una stima della probabilità di ciascuna possibile continuazione. Lo stesso procedimento può essere iterato per sequenze più lunghe, ad esempio "oggi ho", stimando la probabilità delle parole che possono seguire, come "cucinato", "fatto", "visto". Ripetendo sistematicamente questo processo su un ampio corpus, si ottiene una stima della distribuzione delle sequenze linguistiche e si rende possibile la generazione automatica di frasi plausibili. Tuttavia, questo approccio basato su conteggi espliciti presenta un problema di scalabilità. Il numero di parole nel dizionario è molto elevato e le combinazioni possibili crescono rapidamente con la lunghezza delle frasi. Una tabella completa contenente tutte le probabilità condizionate richiederebbe una quantità di memoria impraticabile.

Per superare questo limite si ricorre alle reti neurali profonde. Invece di memorizzare esplicitamente tutte le probabilità, si addestra una rete neurale su grandi quantità di documenti. Alla rete viene fornito un prefisso – ad esempio il simbolo di inizio frase o una sequenza di parole – e il modello produce in output una distribuzione di probabilità su tutte le parole del vocabolario. Se viene selezionata una parola, questa viene aggiunta al prefisso e il processo si ripete. In questo modo, l'informazione statistica presente nei dati non è archiviata in tabelle di grandi dimensioni, ma è compressa nei parametri della rete neurale. Il modello è quindi in grado di fornire le probabilità condizionate per qualunque prefisso e di generare sequenze linguistiche coerenti, riproducendo le regolarità apprese dai documenti di addestramento. Analizzando miliardi di testi, impara quali sequenze di parole

sono plausibili e quali no. Alcune combinazioni sono molto frequenti, altre rare, altre ancora prive di senso. Apprendendo queste regolarità statistiche, il sistema costruisce un modello del linguaggio. Analogamente, nelle immagini, porzioni mancanti possono essere ricostruite a partire dal contesto: si forniscono in input alla rete neurale immagini con porzioni mascherate e si chiede all'apprendimento di ricostruirle, in modo da ottenere l'immagine completa.

Dal punto di vista tecnologico, a supporto del trattamento efficiente ed efficace di una grossa mole di sequenze di testo, un grosso passo avanti è stato fatto con la concezione della architettura Transformer[20]. L'architettura Transformer è particolarmente adatta alla modellazione di sequenze ed è stata introdotta originariamente nel contesto della traduzione automatica. Il suo elemento distintivo è il meccanismo di attenzione, che consente di apprendere le correlazioni tra le diverse parole di una frase senza ricorrere a strutture ricorrenti tradizionali. Nel processo di elaborazione, il testo non viene trattato direttamente come una sequenza di parole intere, ma viene suddiviso in unità elementari denominate token. I token possono corrispondere a parole complete, radici, prefissi, suffissi o frammenti di parola. Questa scelta non è arbitraria: si basa su considerazioni linguistiche e statistiche che derivano dallo studio della struttura e dell'evoluzione dei linguaggi naturali. La tokenizzazione consente di rappresentare in modo efficiente il vocabolario e di gestire parole rare o nuove combinando unità più piccole già note al modello. Attraverso l'addestramento su grandi corpora testuali, il Transformer apprende a stimare le dipendenze tra token anche quando tali dipendenze non sono localmente adiacenti. Si consideri, ad esempio, la frase inglese: "The animal didn't cross the street because it was too tired." In questo caso, il pronome "it" deve essere correttamente associato a "the animal" e interpretato in relazione all'aggettivo "tired". La corretta risoluzione di questa coreferenza richiede la capacità di modellare relazioni a distanza all'interno della frase. Il meccanismo che rende possibile tale operazione è l'autoattenzione (*self-attention*). In questo schema, ogni token della sequenza viene messo in relazione con tutti gli altri token della stessa sequenza. Il modello assegna pesi differenti a tali relazioni,

20 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, *Attention Is All You Need*, in *Advances in Neural Information Processing Systems*, 30, Curran Associates, Inc., 2017, pp. 5998–6008.

determinando quali parole siano più rilevanti per l'interpretazione di ciascun elemento. In questo modo, la rappresentazione di ogni parola non è isolata, ma dipende dal contesto globale della frase.

Questo approccio facilita l'apprendimento delle distribuzioni di probabilità sulle sequenze linguistiche, poiché consente di integrare informazioni contestuali ampie in modo diretto ed efficiente. L'architettura Transformer costituisce pertanto la base dei moderni modelli di linguaggio di grandi dimensioni, dai quali derivano sistemi capaci di comprendere e generare testo in linguaggio naturale, come *ChatGPT*, *Gemini, Claude* e *DeepSeek*.

Quando questi modelli vengono addestrati su scala molto ampia – con enormi quantità di dati e parametri – emergono capacità sorprendenti. Una di queste è l'apprendimento in contesto: il sistema può adattarsi a un nuovo compito semplicemente osservando alcuni esempi forniti nel prompt di input, senza necessità di un ulteriore addestramento formale.

I moderni sistemi di intelligenza artificiale generativa sono basati su questo principio. Essi non sono programmati per ogni singola attività, ma sfruttano una conoscenza generale acquisita durante il pre-addestramento e la applicano in modo flessibile a compiti diversi: rispondere a domande, riassumere testi, analizzare immagini o integrare più modalità informative.

In sintesi, il passaggio dai modelli specializzati ai modelli fondativi rappresenta un cambiamento di paradigma: dall'apprendere compiti specifici con supervisione diretta, all'apprendere la struttura generale dei dati e riutilizzarla in modo adattivo.

Questa importante capacità dell'AI generativa ha, ovviamente, attirato l'attenzione anche in ambito medico. Uno studio prospettico[21], randomizzato e controllato, condotto tra il 2024 e il 2025, ha analizzato l'effetto dell'impiego di GPT-4 nel supporto a decisioni cliniche complesse. Alla ricerca hanno partecipato 92 medici: metà ha utilizzato GPT-4 in combinazione con le consuete fonti di riferimento, mentre l'altra metà ha fatto affidamento esclusivamente su risorse tradizionali. I risultati hanno mostrato che i medici supportati da GPT-4 hanno ottenuto punteggi medi superiori di circa 6,5 punti percentuali rispetto al gruppo di controllo. Un dato particolarmente interessante

21 Ethan Goh, Ryan J. Gallo, Eric Strong, *GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial*, *Nature Medicine*, 2025, 31, pp. 1233–1238.

è che le prestazioni di GPT-4 utilizzato autonomamente sono risultate paragonabili a quelle dei medici che lo impiegavano come strumento di supporto, suggerendo che, in contesti ben delimitati, forme di assistenza decisionale quasi autonome basate sull'intelligenza artificiale potrebbero essere tecnicamente realizzabili. L'uso del sistema ha tuttavia comportato un lieve aumento del tempo necessario per analizzare ciascun caso clinico; secondo i ricercatori che hanno svolto lo studio, tale rallentamento sarebbe dovuto a un processo di valutazione più approfondito da parte dei medici. Nel complesso, i risultati indicano che l'IA generativa può contribuire a migliorare la qualità del processo decisionale clinico, anche se il suo impatto sembra manifestarsi soprattutto sul piano qualitativo più che su quello della pura efficienza operativa.

Bisogna però fare attenzione alle fonti informative fornite in input ai modelli AI. Infatti, uno studio pubblicato su *The Lancet Digital Health* da Mahmud Omar e colleghi[22] ha analizzato la vulnerabilità dei Large Language Model alla disinformazione medica inserita nei prompt, cioè nei testi che gli utenti forniscono ai modelli come istruzioni. La ricerca ha valutato 20 modelli linguistici, sia generalisti sia specificamente addestrati per uso medico, utilizzando circa 3,4 milioni di prompt tratti da note di dimissione ospedaliere, vignette cliniche simulate e post sui social media, tutti contenenti informazioni mediche deliberatamente false.

I ricercatori hanno esaminato la capacità dei modelli di riconoscere la disinformazione nelle raccomandazioni generate e di individuare fallacie logiche nel ragionamento. I risultati hanno mostrato che le prestazioni variano tra i modelli: il modello generalista GPT-4o è risultato il meno suscettibile alla disinformazione e il più accurato nell'identificazione delle fallacie, mentre i modelli ottimizzati specificamente per la medicina hanno ottenuto prestazioni complessivamente inferiori rispetto ai modelli generalisti. Lo studio evidenzia quindi che gli LLM possono essere influenzati dalla disinformazione medica, soprattutto quando questa viene presentata con un tono autorevole. Allo stesso tempo rappresenta uno dei primi benchmark strut-

22 Omar Mahmud, Sorin Vera, Wieler Lothar H., Charney Alexander W., Kovatch Patricia, Horowitz Carol R., Korfiatis Panagiotis, Glicksberg Benjamin S., Freeman Robert, Nadkarni Girish N., Klang Eyal, *Mapping the susceptibility of large language models to medical misinformation across clinical notes and social media: a cross-sectional benchmarking analysis*, The Lancet Digital Health, Volume 8, Issue 1, 2026.

turati su larga scala per valutare il comportamento dei modelli di linguaggio di fronte a informazioni mediche false.

Affinché tali sistemi possano essere realmente efficaci nel supporto decisionale, è necessario progettare modalità di interazione adeguate tra l'utente umano e il modello. Ciò implica lo sviluppo di interfacce e flussi di lavoro che permettano al medico di sfruttare in modo efficace le informazioni prodotte dal sistema. In altre parole, il problema non riguarda esclusivamente la qualità del modello linguistico, ma anche l'ergonomia e la progettazione dell'interazione uomo–macchina.

Un ulteriore ambito di ricerca riguarda l'utilizzo dei modelli di linguaggio come agenti artificiali capaci di simulare il comportamento di diversi specialisti. In un esperimento condotto presso l'Università di Stanford[23] è stato sviluppato un laboratorio virtuale nel quale più modelli di linguaggio collaborano simulando ruoli differenti all'interno di un team di ricerca. In questo contesto vengono rappresentate diverse competenze: uno specialista responsabile dell'individuazione del problema scientifico, un immunologo, un esperto di *machine learning*, un biologo e un revisore critico incaricato di individuare possibili errori logici o metodologici, assumendo una funzione simile a quella di un "avvocato del diavolo".

Questi agenti artificiali interagiscono tra loro per affrontare problemi di ricerca complessi. In uno degli esperimenti, ricercatori umani hanno incaricato il laboratorio virtuale di progettare nanocorpi, ossia frammenti di anticorpi in grado di legarsi al virus SARS-CoV-2, responsabile della pandemia di COVID-19. Il sistema ha generato 92 candidati nanocorpi, più del 90% dei quali si è dimostrato efficace contro il virus quando successivamente testato in laboratorio. Questo risultato evidenzia come sistemi basati su modelli di linguaggio possano essere utilizzati anche come strumenti di supporto alla ricerca scientifica.

Nel contesto medico esistono, inoltre, piattaforme che integrano più modelli linguistici in un unico sistema. Un esempio è Polaris 3.0[24], che combina diversi modelli di linguaggio naturale per consentire interazioni coordinate tra differenti componenti intelligenti. Tuttavia,

23 Stanford University, *AI in Biomedical Research: Opportunities and Challenges*, FreeThink, Capitolo 5, AI Index Report 2025, Stanford Institute for Human-Centered AI (HAI), 2025. Disponibile online: https://hai.stanford.edu/assets/files/hai_ai-index-report-2025_chapter5_final.pdf

24 Hippocratic AI, *Polaris 3.0: A 4.2 Trillion Parameter Suite of 22 LLMs Enhancing Patient Safety and Experience*, Palo Alto, Hippocratic AI, 2025.

oltre al linguaggio, riveste un ruolo fondamentale anche la visione artificiale. Tecniche avanzate consentono oggi non solo di analizzare immagini mediche, come precedentemente discusso, ma anche di generare immagini sintetiche[25] o produrre video artificiali[26]. Negli ultimi anni sono stati sviluppati numerosi modelli generativi dedicati alla diagnosi basata su immagini, e tali sistemi hanno richiesto quantità di calcolo sempre più elevate. L'aumento della dimensione dei dataset e della complessità dei modelli ha infatti comportato una crescita significativa delle risorse computazionali necessarie. In generale, l'incremento dei dati disponibili consente di migliorare le prestazioni dei modelli, ma richiede parallelamente una maggiore capacità di calcolo.

Tra gli esempi recenti e più rilevanti di applicazione dell'intelligenza artificiale nel campo biologico si trovano modelli sviluppati per la progettazione di molecole e per la simulazione biologica. Un caso significativo riguarda i sistemi sviluppati da DeepMind per la generazione di nuovi ligandi proteici ad alta affinità, capaci di legarsi selettivamente a specifiche molecole bersaglio. Parallelamente sono stati sviluppati modelli computazionali per la simulazione di porzioni di sistemi biologici complessi, come piccole reti neuronali del cervello.

L'intelligenza artificiale è già integrata in numerosi sistemi utilizzati nella pratica sanitaria. Studi condotti a livello europeo[27], basati su indagini tra diversi attori del settore sanitario, mostrano che l'AI è impiegata in molteplici ambiti, tra cui la ottimizzazione delle risorse e dell'efficienza del flusso di lavoro, snellimento dei compiti amministrativi, miglioramento dell'accuratezza diagnostica, creazione di piani di trattamenti personalizzati, analisi predittiva sui risultati dei pazienti, miglioramento del coinvolgimento e dell'aderenza al piano terapeutico, affrontare i gap delle competenze nella forza lavoro sanitaria, garantire un equo accesso all'assistenza sanitaria. Un indicatore

25 Jinzhuo Wang Et Al., *Self-improving generative foundation model for synthetic medical image generation and clinical applications*, *Nature Medicine*, 2025, 31(2), pp. 609–617.

26 Hadrien Reynaud, Qingjie Meng, Mischa Dombrowski, Arijit Ghosh, Thomas Day, Alberto Gomez, Paul Leeson, Bernhard Kainz, *Echo-Net-Synthetic: Privacy-preserving Video Generation for Safe Medical Data Sharing*, Medical Image Computing and Computer Assisted Intervention –27th International Conference, Marrakesh, Morocco, October 6–10, 2024, Proceedings, Part VII.

27 European Commission – Directorate-General For Health And Food Safety, Eeig, Open Evidence, PwC, *Study on the deployment of AI in healthcare*, *Publications Office of the European Union*, Lussemburgo, 2025, ISBN 978-92-68-28758-3, DOI 10.2875/2169577.

significativo di questa diffusione è rappresentato dal numero crescente di dispositivi medici che incorporano componenti di intelligenza artificiale approvati dalla Food and Drug Administration negli Stati Uniti, che negli ultimi anni ha mostrato una crescita importante[28].

Anche dal punto di vista della produzione scientifica si osserva un aumento significativo degli studi che includono riferimenti all'intelligenza artificiale[29]. Analizzando la distribuzione geografica delle pubblicazioni emerge che l'Italia presenta un numero di contributi comparabile a quello di altri paesi europei, indicando una partecipazione attiva della comunità scientifica nazionale in questo ambito[30].

Un ulteriore sviluppo rilevante riguarda l'approccio multimodale, che integra diverse tipologie di dati, come testo e immagini. Un esempio importante è rappresentato dal modello CLIP[31], che combina rappresentazioni visive e linguistiche all'interno dello stesso sistema. Grazie a questa integrazione, il modello è in grado di analizzare immagini e produrre descrizioni testuali coerenti. Nel contesto medico, un sistema di questo tipo può analizzare automaticamente una radiografia e produrre un'interpretazione testuale simile a quella formulata da un radiologo, identificando la presenza di anomalie come masse sospette o possibili indicazioni di patologie[32]. Lo stesso approccio può essere applicato all'analisi di immagini microscopiche, come i vetrini istologici, consentendo non solo la descrizione generale dell'immagine, ma anche l'individuazione di specifiche regioni di interesse e la classificazione delle anomalie osservate.

28 Adrien Laurent, *AI Medical Devices: 2025 Status, Regulation & Challenges*. https://intuitionlabs.ai/pdfs/ai-medical-devices-2025-status-regulation-challenges.pdf

29 Rotem Lahat, Noa Berick, Majd Hajouj, Tali Teitelbaum, Isaac Shochat, *AIAIAI: AI insights on amassing influence in AI-related publications – an AI-assisted retrospective analysis into AI-related publication*, *BMJ Health Care Informatics*, 2025, 32(1), e101244.

30 Abdulhadi M. Alotaibi Et Al., *Bibliometric analysis of artificial intelligence applications in cardiovascular imaging: trends, impact, and emerging research areas*, *Annals of Medicine & Surgery*, 2025, 87(4), pp. 1947–1968.

31 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, *Learning Transferable Visual Models From Natural Language Supervision (CLIP)*, Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021, pp. 1–17.

32 Iustin Sîrbu, Iulia-Renata Sîrbu, Jasmina Bogojeska, Traian Rebedea, *GIT-CXR: End-to-End Transformer for Chest X-Ray Report Generation*, *Information*, 2025, 16(7), 524.

Le analisi prospettiche presentate in un recente studio della Commissione Europea[33] indicano che l'impiego dell'intelligenza artificiale nel settore sanitario è destinato ad ampliarsi in numerosi ambiti applicativi. Tra le principali aree di sviluppo si includono la robotica chirurgica, i sistemi di supporto alla prognosi, il monitoraggio remoto dei pazienti, l'ottimizzazione dei flussi di lavoro clinici e altre applicazioni orientate alla gestione dei processi sanitari e all'assistenza medica. Questi scenari evidenziano una progressiva integrazione dell'intelligenza artificiale all'interno delle infrastrutture sanitarie e delle pratiche cliniche.

Parallelamente alla diffusione di tali tecnologie emergono tuttavia alcune criticità di carattere sociale e sistemico. In particolare, l'adozione su larga scala di modelli di intelligenza artificiale implica che eventuali errori presenti nel modello possano produrre effetti su scala globale. Quando un sistema viene utilizzato in contesti diffusi o integrato in piattaforme condivise, un errore sistematico può propagarsi rapidamente e influenzare un numero molto elevato di decisioni cliniche o operative.

Questa dinamica differisce da quella dei sistemi tradizionali utilizzati esclusivamente in contesti locali. In tali situazioni, eventuali errori tendono a produrre conseguenze circoscritte al singolo ambiente di utilizzo e non si propagano necessariamente su scala più ampia. Nel caso dei modelli di intelligenza artificiale distribuiti globalmente, invece, il rischio riguarda la possibilità che un problema strutturale del sistema venga replicato in molteplici contesti applicativi.

Per questo motivo è fondamentale prestare particolare attenzione alla qualità dei dati utilizzati per l'addestramento dei modelli e alla presenza di possibili distorsioni o bias nei dataset, come ben evidenziato dallo studio di Obermeyer e colleghi[34]. Se i dati di addestramento non rappresentano adeguatamente la variabilità delle popolazioni o delle condizioni cliniche, il modello può produrre risultati sistematicamente distorti. Diversi studi hanno evidenziato la presenza di tali problemi in vari ambiti della medicina. Tra gli esempi più

33 Per informazioni complete e aggiornate sulla politica comunitaria in ambito applicativo dell'intelligenza artificiale nel settore dell'assistenza sanitaria cfr. la pagina web del sito della Commissione Europea: https://health.ec.europa.eu/ehealth-digital-health-and-care/artificial-intelligence-healthcare_it .

34 Ziad Obermeyer, Brian Powers, Christine Vogeli, Sendhil Mullainathan, *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, in *Science*, vol. 366, n. 6464, 2019, pp. 447–453.

frequentemente citati figurano i sistemi AI utilizzati in dermatologia[35], nei quali dataset poco rappresentativi possono portare a prestazioni differenti tra gruppi di pazienti; applicazioni in cardiologia[36]; sistemi utilizzati per l'assegnazione delle risorse sanitarie[37]; e strumenti di analisi nella medicina genomica[38]. In tutti questi casi, la qualità e la rappresentatività dei dati di addestramento risultano determinanti per garantire l'affidabilità e l'equità delle decisioni supportate dall'intelligenza artificiale. Questo conferma che l'AI può amplificare disuguaglianze preesistenti se non sottoposta a *debiasing* attivo durante l'addestramento.

Sul piano regolatorio, sebbene il numero di dispositivi medici approvati dalla FDA sia in crescita, la letteratura recente si concentra sulla «fragilità» dell'AI. Kelly e colleghi[39] avevano già invocato validazioni prospettiche. Oggi, la discussione si è spostata sulla AI Governance, come ben rappresentato nel lavoro di Palmieri e colleghi[40].

La visione dell'Europa rispetto all'uso dell'AI, in generale e in ambito medico in particolare, pone al centro la persona, rendendo fondamentale l'aspetto etico nell'adozione e nell'uso dell'intelligenza artificiale. A questo scopo, l'*AI Act*[41] definisce sette requisiti chiari, che coprono aree come trasparenza, sicurezza, accuratezza, supervisione umana, protezione dei dati, responsabilità e prevenzione di discriminazioni. L'*AI Act* è un documento molto ampio e complesso, con nume-

35 Andrzej Grzybowski, Kai Jin, Hong Wu, *Challenges of artificial intelligence in medicine and dermatology*, *Clinical Dermatology*, 2024, 42 (3), pp. 210–215.

36 Dennis Amponsah, Ravi Thamman, Elizabeth Brandt, Christopher James, Kathleen Spector-Bagdady, Chien-Ming Yong, *Artificial Intelligence to Promote Racial and Ethnic Cardiovascular Health Equity*, *Current Cardiovascular Risk Reports*, 2024, 18(11), pp. 153–162.

37 Richard J. Chen, Judy J. Wang, Drew F. K. Williamson, Tiffany Y. Chen, Jana Lipkova, Ming Y. Lu, Sharifa Sahai, Faisal Mahmood, *Algorithmic fairness in artificial intelligence for medicine and healthcare*, *Nature Biomedical Engineering*, 2023, 7, pp. 719–742.

38 Leslie A. Smith, James A. Cahill, Ji-Hyun Lee, Kiley Graim, *Equitable machine learning counteracts ancestral bias in precision medicine*, *Nature Communications*, 2025, 16(1), 2144.

39 Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, Dominic King, *Key challenges for delivering clinical impact with artificial intelligence*, *BMC Medicine*, 2019, 17, 195.

40 Sofia Palmieri, Christopher T. Robertson, I. Glen Cohen, *New Guidance on Responsible Use of AI*, *JAMA*, 2026, 335(3), pp. 207–208.

41 Unione Europea, *Regolamento (UE) 2024/1689 – Artificial Intelligence Act*, *Gazzetta Ufficiale dell'Unione Europea*, L 305/1, 2024.

rosi articoli e allegati. La sua lettura risulta difficile anche per esperti, e presenta alcune incongruenze rispetto alla normativa europea esistente sulla definizione di dato (come il GDPR[42]), perché le concezioni del dato e della sua gestione variano. In particolare, mentre il GDPR si concentra sulla protezione dei dati personali, l'*AI Act* introduce una valutazione dei rischi legati all'uso dei sistemi AI, considerando il potenziale impatto sui diritti fondamentali e sulla sicurezza delle persone. L'*AI Act* individua diversi livelli di rischio nell'uso dei sistemi di intelligenza artificiale, classificandoli in: *rischio inaccettabile*, che riguarda sistemi vietati perché possono minacciare i diritti fondamentali o la sicurezza delle persone; il *rischio elevato*, che include sistemi soggetti a obblighi normativi severi, come certificazioni e valutazioni di conformità prima della messa in commercio; il *rischio di trasparenza*, che riguarda sistemi che richiedono informazioni chiare agli utenti sull'uso dell'AI, ma con minori vincoli tecnici; infine, il *rischio minimo o nullo*, che si riferisce a sistemi con impatto limitato, per i quali non sono previsti obblighi particolari. Al momento, benché esistano già alcune controversie giuridiche in cui l'*AI Act* è entrato nel dibattito giuridico[43], non esistono sentenze giurisprudenziali definitive sull'interpretazione dell'*AI Act*; di conseguenza, le interpretazioni rimangono variabili tra esperti e istituzioni. Solo il magistrato, attraverso eventuali casi concreti, potrà fornire chiarimenti vincolanti sulle applicazioni pratiche. In ambito medico, questi livelli di rischio sono particolarmente significativi, perché l'uso dell'AI può influenzare decisioni cliniche, gestione dei pazienti e sicurezza sanitaria, rendendo essenziale valutare e monitorare costantemente l'accuratezza, la trasparenza e l'impatto etico dei sistemi implementati.

Concludo con alcune considerazioni finali. Nella storia dell'umanità, l'uomo ha sempre cercato di potenziare le proprie capacità attraverso strumenti fisici. Questa è la prima volta che il potenziamento riguarda la sfera cognitiva, un fenomeno mai avvenuto prima. Essendo software, questi strumenti cognitivi possono essere facilmente replicati nello spazio e nel tempo, rendendo i modelli, in linea di principio, potenzialmente «eterni»: possiamo cambiare l'hardware, ma il

42 Unione Europea, *Regolamento (UE) 2016/679 del Parlamento Europeo e del Consiglio – GDPR (General Data Protection Regulation)*, *Gazzetta Ufficiale dell'Unione Europea*, L 119/1, 2016.

43 Tar Lazio, Sezione III-bis, sentenza 2 febbraio 2026, n. 1895; Corte di Giustizia dell'Unione Europea – richiesta di pronuncia pregiudiziale (C-806/24).

software può essere duplicato e distribuito senza limiti. Il potenziamento consiste nell'acquisire conoscenze e competenze derivanti da uno o più individui, attraverso i loro prodotti, scritti, studi o segnali interpretati. Il *Deep Learning* viene applicato a dati di varia natura, consentendo di raccogliere questa conoscenza e distribuirla attraverso servizi *cloud* in tutto il mondo dove è presente *Internet*. Questo significa che eventuali difetti nei modelli si propagano ovunque, rendendo fondamentale la prudenza e l'attenzione nello sviluppo e nell'uso di tali sistemi. Non è possibile evitare completamente gli usi intenzionalmente malevoli, poiché questi ci saranno sempre. È invece necessario ridurre i rischi inconsapevoli, quelli ancora non riconosciuti all'interno dei sistemi basati su AI. Per questo motivo, è fondamentale mantenere l'uomo al centro dello sviluppo dell'intelligenza artificiale, garantendo una supervisione continua di questi strumenti, soprattutto in ambito medico.

ABSTRACT

L'intelligenza artificiale (AI) è un campo multidisciplinare difficile da definire in modo univoco, poiché riflette la complessità del concetto di intelligenza umana. Fin dalle origini ha integrato approcci diversi, dal ragionamento logico ai sistemi esperti fino al machine learning. In medicina, le prime applicazioni negli anni Sessanta e Settanta miravano a supportare diagnosi e decisioni cliniche, ma richiedevano una complessa codifica manuale della conoscenza. Una svolta è avvenuta con il deep learning e le reti neurali profonde, capaci di apprendere direttamente dai dati e analizzare immagini e testi su larga scala. Lo sviluppo dei foundation models e dell'apprendimento auto-supervisionato ha ampliato ulteriormente le applicazioni, dall'analisi automatica di immagini mediche al supporto alle decisioni cliniche e alla ricerca scientifica. Restano tuttavia criticità rilevanti, tra cui dipendenza dalla qualità dei dati, rischio di bias e vulnerabilità alla disinformazione. La diffusione globale di questi sistemi può amplificare eventuali errori, rendendo essenziali regolamentazione, governance e supervisione umana per un uso etico e responsabile dell'AI in medicina.

Artificial intelligence (AI) is a multidisciplinary field that is difficult to define precisely, as it reflects the complexity of the concept of human intelligence itself. Since its origins, AI has integrated diverse approaches, from logical reasoning and expert systems to machine learning. In medicine, early applications in the 1960s and 1970s aimed to support diagnosis and clinical decision-making, but they required extensive manual encoding of expert knowledge.

A major shift occurred with the emergence of deep learning and neural networks, which can learn directly from data and analyze images and text at large scale. The development of foundation models and self-supervised learning has further expanded AI applications, including automated medical image analysis, clinical decision support, and assistance in scientific research. However, significant challenges remain, including dependence on data quality, the risk of bias, and vulnerability to misinformation. The global deployment of AI systems can also amplify errors at scale, making regulation, governance, and human oversight essential to ensure ethical and responsible use of AI in medicine.