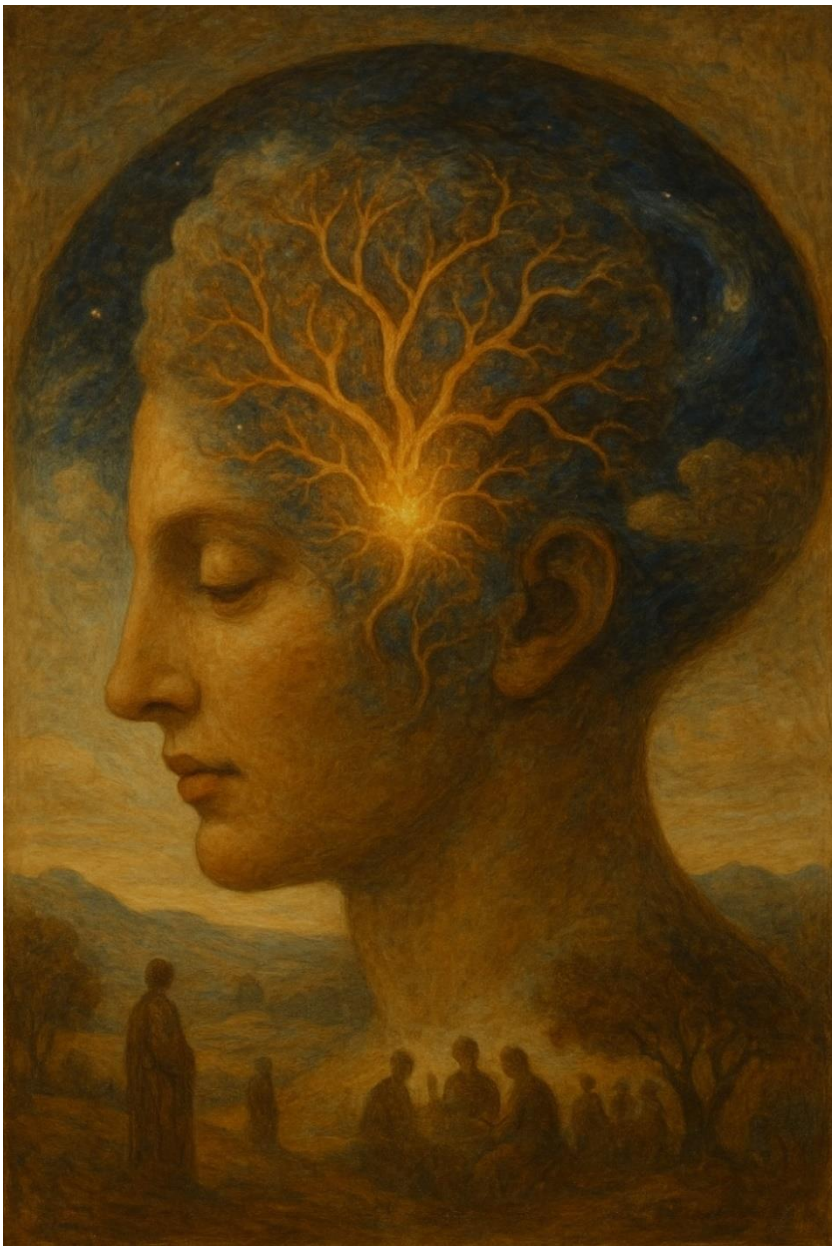
AUTORITÀ PER LE GARANZIE NELLE COMUNICAZIONI
AGCOM

Intelligenza Artificiale



I PARTE

Rapporto tecnico-economico, 2026

Sommario

1 Introduzione

L'intelligenza artificiale sta rapidamente trasformando il panorama economico, tecnologico e sociale, comportando cambiamenti epocali in molteplici ambiti. Se da un lato tale evoluzione favorisce l'innovazione e l'efficienza produttiva, dall'altro solleva questioni critiche legate alla tutela dei diritti fondamentali della persona, all'evoluzione di interi comparti economici, alla trasformazione dei contesti lavorativi, sociali e culturali. Consapevoli di tale sfida, le Istituzioni europee, a partire dalla Commissione, sin dallo scorso mandato, hanno adottato iniziative volte ad affrontare, a livello regolamentare, il fenomeno. In tal senso, rileva l'approvazione, nel corso del 2024, del regolamento sull'intelligenza artificiale (c.d. AI Act)[1] da parte del Parlamento europeo. Il regolamento, che stabilisce obblighi relativi all'impiego di sistemi e modelli di IA sulla base dei rischi legati all'uso degli stessi, definisce differenti categorie di rischio, graduando responsabilità e divieti in capo ai soggetti che popolano tale ecosistema. Inoltre, nel fissare alcuni obblighi di trasparenza, l'AI Act sottolinea come essi siano altresì essenziali per un'efficace attuazione del Digital Services Act (DSA – regolamento UE 2022/2065), enfatizzando l'impatto della diffusione di contenuti generati o manipolati artificialmente "*sui processi democratici, sul dibattito civico e sui processi elettorali, anche mediante la disinformazione*"[2].

L'analisi che il presente rapporto intende offrire mira anche a gettare le basi per una futura disamina delle intersezioni tra DSA e AI Act, proprio in virtù delle competenze attribuite ad AGCOM in qualità di Coordinatore dei Servizi digitali per l'Italia[3]. Pertanto, il rapporto intende offrire una ricostruzione generale dell'evoluzione dell'IA alla luce del ruolo assegnato ad AGCOM nell'ambito del DSA, ferme restando le competenze designate dal legislatore alle autorità nazionali competenti per l'intelligenza artificiale ai sensi dell'IA Act.

[1] Si veda il paragrafo n. 179 del regolamento (UE) 2024/1689: l'AI Act – che è entrato in vigore venti giorni dopo la sua pubblicazione nella Gazzetta ufficiale dell'UE – inizierà a produrre i propri effetti trascorsi 24 mesi dal primo agosto 2024, con l'esclusione dei divieti relativi a pratiche vietate (applicazione a partire da sei mesi dopo l'entrata in vigore del regolamento), dei codici di buone pratiche (applicazione a partire da nove mesi dopo l'entrata in vigore), delle norme sui sistemi di IA per finalità generali, compresa la governance (applicazione a partire da dodici mesi dopo l'entrata in vigore) e degli obblighi per i sistemi ad alto rischio (applicazione a partire da trentasei mesi dopo l'entrata in vigore).

[2] Si veda il paragrafo n. 120 del regolamento (UE) 2024/1689.

[3] Competenza attribuita ai sensi del Decreto-Legge n. 123 del 15 settembre 2023, convertito con modificazioni dalla L. 13 novembre 2023, n. 159 (in G.U. 14/11/2023, n. 266).

In ragione delle proprie attribuzioni e alla luce dell'evoluzione tecnologica in atto, AGCOM ha dunque dato impulso a una serie di attività volte ad analizzare l'intelligenza artificiale, i suoi effetti sui settori e sulle materie di interesse istituzionale, nonché a monitorare l'evoluzione del quadro normativo europeo e nazionale. In quest'ottica, l'Autorità ha dapprima istituito, con delibera n. 11/24/CONS del 24 gennaio 2024, un Comitato sull'intelligenza artificiale (di seguito anche il Comitato) al fine di assicurare un supporto qualificato e specializzato in merito alle implicazioni dei sistemi di IA sugli ambiti di competenza dell'Autorità e sul ruolo che la stessa potrà assumere in materia, con funzioni consultive[4]. Successivamente – in sede di riorganizzazione degli Uffici – si è deciso di dotare l'Autorità di una Struttura ad hoc finalizzata a svolgere attività di studio e analisi in materia di big data e IA, oltre che di coordinamento del Comitato stesso (delibera n. 382/24/CONS del 30 settembre 2024).

In considerazione della necessità di assicurare un presidio costante del contesto tecnologico di riferimento, il presente documento di lavoro avvia le attività di monitoraggio dell'evoluzione delle tecnologie, del quadro normativo e del dibattito in materia di intelligenza artificiale, anche al fine di favorire una riflessione approfondita sulle relative implicazioni.

Come illustrato in Premessa, il presente documento deve essere letto congiuntamente al Rapporto sull'Intelligenza artificiale elaborato dal Comitato, che completa e approfondisce l'analisi su varie tematiche di ordine normativo e regolamentare.

[4] Il Comitato ha durata di due anni, eventualmente rinnovabili, ed è così composto: Prof. Andrea Renda (European University Institute), coordinatore; Prof. Giovanni Boccia Artieri (Università degli Studi di Urbino), componente; Prof. Giuseppe Cassano (European School of Economics), componente; Prof. Mauro Giusto (Università degli studi di Milano Statale), componente; Prof. Andrea Simoncini (Università degli Studi di Firenze), componente; Prof.ssa Giovanna De Minico (Università degli Studi di Napoli Federico II), componente; Andrea Imperiali, componente.

2 Evoluzione dell'IA

L'emersione dell'IA quale disciplina scientifica è riconducibile a uno specifico evento: il *Dartmouth Summer Research Project on Artificial Intelligence* del 1956. Gli organizzatori di tale conferenza partivano dal presupposto che ogni aspetto dell'**apprendimento**, ovvero qualsiasi altra caratteristica dell'intelligenza, potesse essere descritto, in linea di principio, in modo così preciso da poter essere simulato da una macchina[5]. Quegli stessi scienziati tentarono di trovare il modo di far usare il linguaggio alle macchine, facendo sì che le stesse elaborassero concetti, risolvessero problemi – allora riservati agli esseri umani – fino a giungere al miglioramento delle macchine stesse[6].

Quella esperienza portò dapprima uno dei partecipanti alla *summer school* – John McCarthy – a coniare il termine "Intelligenza Artificiale" (vale a dire: "*The science and engineering of making intelligent machines*"[7]), e poi alla costituzione dell'AI lab presso il MIT. Il termine "Intelligenza Artificiale" venne coniato successivamente alla definizione del criterio volto a valutare se i computer fossero capaci di mimare l'intelligenza umana, ossia di quello che sarebbe diventato noto come il "**test di Turing**"[8] (si vedano anche §§ 2.1 e 3).

Box 1 – Test di Turing

Proposto nel 1950 da Alan Turing, il test è un criterio per determinare se una macchina è in grado di esibire un comportamento intelligente. Il test si basa su un'idea cartesiana di cosa sia un essere umano, e infatti la sua attendibilità è stata messa recentemente in discussione. Esso consiste in una conversazione scritta tra un

[5] L'idea di fondo dell'evento organizzato John McCarthy, Marvin Minsky, Claude E. Shannon e Nathaniel Rochester (informatico dell'IBM) era che "*Ogni aspetto dell'apprendimento o qualsiasi altra caratteristica dell'intelligenza può, in linea di principio, essere descritto in modo così preciso che una macchina può essere fatta per simularlo*". Si veda: McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E., A proposal for the Dartmouth summer research project on artificial intelligence, august 31, 1955: "*every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it*".

[6] McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E., cit. Si riporta di seguito quanto scritto nel testo: "*The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be* made *to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer*".

[7] McCarthy, John. (2007). What is artificial intelligence.

[8] Turing, A.M., (1950), Computing machinery and intelligence, *Mind*, 49.

essere umano e una macchina, mediata da un altro essere umano (l'interrogatore), il cui compito è identificare chi dei due interlocutori è umano. Se la macchina riesce a ingannare l'interrogatore al punto da risultare indistinguibile da un essere umano, si considera che essa abbia superato il test (in questo contesto si applicano soglie, o percentuali, oltre le quali la macchina è considerata indistinguibile, che partono, nell'interpretazione classica, dal 30%).

L'obiettivo principale del test è dimostrare che una macchina può simulare l'intelligenza umana almeno sul piano linguistico. Tuttavia, esso non misura la coscienza o la comprensione autentica, ma soltanto la capacità di imitare il comportamento linguistico umano.

Oggi, i modelli linguistici avanzati sono in grado di superare il test in numerosi contesti, ma ciò non implica necessariamente che possiedano intelligenza generale o autocoscienza[9]–[10]. Di conseguenza, la validità del test come prova definitiva di intelligenza è stata messa in discussione. Poiché queste macchine sono ormai in grado di imitare quasi perfettamente il registro e la fluidità umana attraverso sofisticati calcoli probabilistici, il test di Turing tende oggi a misurare più la capacità di "simulazione" che una reale forma di comprensione o coscienza. Il superamento del test non è più considerato il traguardo finale, poiché un sistema può apparire umano senza possedere una reale capacità di ragionamento astratto o di intenzionalità.

Per questo motivo, la comunità scientifica sta adottando nuovi parametri di valutazione molto più rigorosi. Tra i più rilevanti figurano il MMLU (*Massive Multitask Language Understanding*), che valuta le conoscenze in decine di discipline accademiche, e l'ARC-AGI (*Abstraction and Reasoning Corpus*). Quest'ultimo, in particolare, sfida i modelli a risolvere problemi logici inediti che richiedono un'intelligenza fluida e non la semplice ripetizione di schemi appresi durante l'addestramento, cercando di tracciare un confine netto tra l'imitazione del linguaggio e la vera capacità di risoluzione di problemi (per un'analisi delle implicazioni contemporanee del superamento del test di Turing da parte degli LLM, si vedano §§ 3.2.2 per il quadro tecnico e 5.5 per i riflessi sul ruolo degli LLM come infrastruttura di accesso all'informazione).

[9] Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, *121*(9), e2313925121. Si veda anche: "Study finds ChatGPT's latest bot behaves like humans, only better".

[10] Biever, C. (2023). ChatGPT broke the Turing test-the race is on for new ways to assess AI. *Nature*, *619*(7971), 686-689.

2.1 Prima primavera dell'IA (1956 – 1973)

L'evento del 1956 suscitò grande entusiasmo all'interno della comunità scientifica, e venne inizialmente qualificato come "*AI spring*" (ossia come "primavera" dell'intelligenza artificiale), oggi nota anche come prima ondata dell'intelligenza artificiale (*first wave*). Il termine "ondata" restituisce l'idea della ciclicità che ha caratterizzato le fasi evolutive dell'IA, laddove a grandi aspettative hanno corrisposto progressi esigui che hanno portato – in alcune fasi, note come "inverni" (*AI winters*) – a un disinteresse di scienziati e investitori per il settore e per l'argomento (v. Figura 1). Nella storia dell'IA si sono infatti registrate due battute d'arresto: la prima tra il 1974 e il 1980[11], la seconda tra la fine degli anni '80 e il 2011[12].

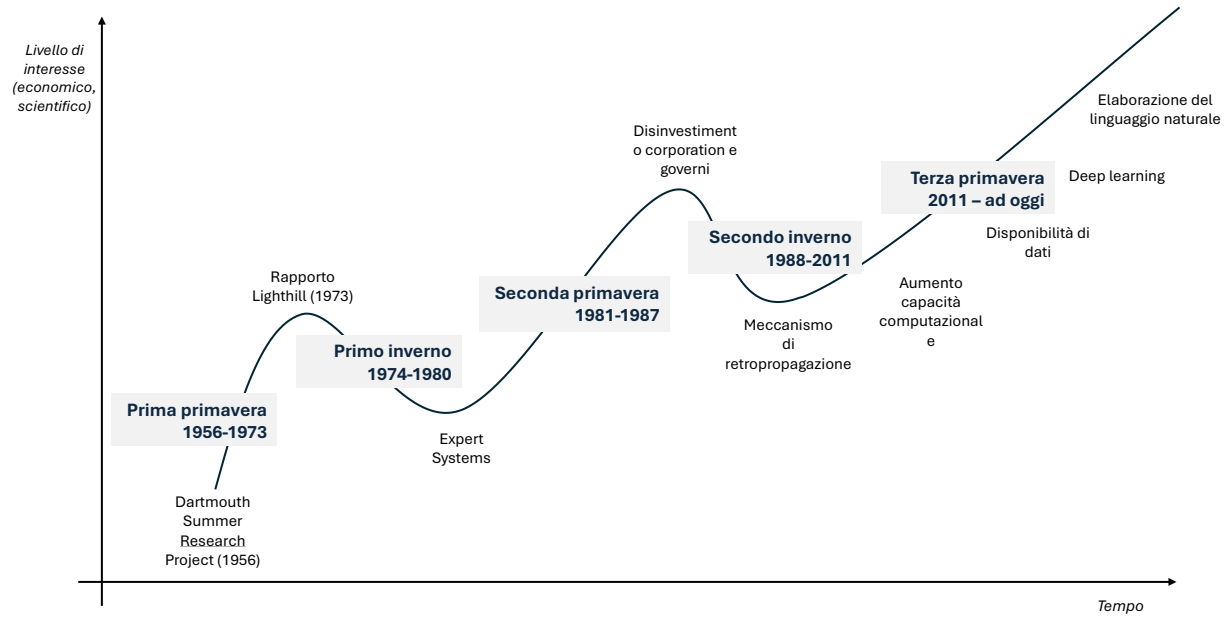


Figura 1 – Timeline dell'IA

[11] Muthukrishnan, N., Maleki, F., Ovens, K., Reinhold, C., Forghani, B., & Forghani, R. (2020). Brief history of artificial intelligence, *Neuroimaging Clinics of North America*, *30*(4), 393-399.

[12] Toosi, A., Bottino, A. G., Saboury, B., Siegel, E., & Rahmim, A. (2021). "A brief history of AI: how to prevent another winter (a critical review)", *PET Clinics*, *16*(4), 449-469.

Tornando agli albori della disciplina, l'interesse per la conferenza di Dartmouth innescò lo sviluppo di vari programmi e, tra questi, alcuni capaci di giocare al gioco della dama[13]. Fra di essi, suscitò grande interesse quello sviluppato da Allen Newell e Herbert Simon[14]: i due presentarono un programma, il *Logic Theorist* (LT), in grado di mettere a punto teorie afferenti alla **logica simbolica**, unitamente a un linguaggio di elaborazione chiamato IPL (*Information Processing Language*). Il programma costituisce uno spartiacque nella storia dell'IA perché aprì la strada alla risoluzione di problemi concettuali e logici tanto da riuscire a dimostrare alcuni dei teoremi tratti dai "*Principia Mathematica*" di Whitehead e Russell[15]. Nel 1959, Newell e Simon rilasciarono il *General Problem Solver* (GPS), un programma ideato al fine di imitare i protocolli di risoluzione dei problemi del cervello umano[16]: il software proponeva soluzioni traducendo i problemi in obiettivi, sotto-obiettivi, e identificando azioni e operatori. Grazie a tale decodificazione della realtà, il GPS riuscì a risolvere il puzzle logico dell'attraversamento del fiume (c.d. "*Missionaries and Cannibals problem*")[17]. Tuttavia, nonostante il programma funzionasse bene se applicato a problemi semplici, l'entusiasmo prodotto dal GPS si infranse nel momento in cui i ricercatori compresero che lo stesso non avrebbe potuto avere un'applicazione così generalizzata come il nome lasciava intendere[18]. Oltre ai limiti intrinseci del sistema, è peraltro evidente che le sue performance, come quelle di tutti i primi programmi messi a punto, scontassero necessariamente i limiti relativi alla dimensione e alla velocità della memoria e dei processori[19].

Negli anni successivi si registrarono diversi avanzamenti nella teoria, seppur con un impatto limitato al di fuori dei laboratori. Secondo John Launchbury, ex Direttore dell'*Information*

[13] Sheikh, H., Prins, C., & Schrijvers, E. (2023). Mission AI: The new system technology, *Springer Nature*, 410.

[14] Allen Newell oltre a essere uno dei padri fondatori della disciplina contribuì poi a sviluppare la psicologia cognitiva, mentre Herbert Simon si concentrò, tra le altre cose, sul concetto di razionalità, in particolare quello di "razionalità limitata, e per questi studi nel 1978 vinse il premio Nobel per l'economia.

[15] Il programma riuscì infatti a risolvere la maggior parte, 38 su 52, dei teoremi del secondo capitolo del libro.

[16] Newell, A., Shaw, J. C., & Simon, H. A. (1959). Report on a general problem solving program, in *IFIP Congress* (vol. 256, p. 64).

[17] Boden, M. A. (2008). *Mind as machine: A history of cognitive science*. Oxford University Press: "*GPS could even solve the tricky missionaries-and-cannibals puzzle, which requires one to go backwards in order to go forwards (Three missionaries and three cannibals on one side of a river; a boat big enough for two people; how can everyone cross the river, without cannibals ever outnumbering missionaries?). Purpose, thinking, and mental representations: all these seemed within reach at last*".

[18] Coppin, B. (2004). *Artificial intelligence illuminated*. Jones & Bartlett Learning.

[19] Buchanan, B. G. (2005). A (very) brief history of artificial intelligence, *AI Magazine*, *26*(4), 53-53.

Innovation Office del DARPA (Defense Advanced Research Projects Agency), la prima primavera dell'IA può essere ricondotta alla nozione di *"Crafted Knowledge"*, definizione che ricomprende quei sistemi di intelligenza artificiale basati su **regole** che i programmatori forniscono alle macchine[20]. Infatti, circoscrivendo la complessità di un evento e riconducendola alle regole impartitegli dai programmatori, le macchine erano capaci di ragionare in relazione ad alcuni **domini ristretti**, pur non avendo la capacità di imparare, di astrarre, né di gestire situazioni connotate da un certo grado di incertezza. In questo solco, si sono inseriti gli studi di Marvin Minsky, che, nel 1963, propose un approccio di semplificazione per i casi d'uso dell'IA. Minsky, assieme a Seymour Papert, suggerì che gli studi sull'IA si dovessero concentrare sull'ideazione di programmi capaci di elaborare risposte in ambienti artificiali più piccoli: il cosiddetto "universo a blocchi" ("*micro–worlds*")[21].

Qualche anno prima, Minsky e Papert scrissero un libro (*Perceptrons*)[22] in cui veniva contestato il lavoro svolto sulle reti neurali da Frank Rosenblatt, oggi considerato il padre delle reti neurali per aver introdotto la "*error-based perceptron learning rule*"[23–24]. Con il proprio lavoro, gli autori intendevano avanzare una critica nei confronti di uno degli approcci allo studio dell'IA emersi nel corso della prima primavera: facendo proprio un approccio di tipo "**simbolico/logico**" si scagliarono infatti contro il "**connettivismo**" (o "connessionismo") alla base del paradigma delle reti di Rosenblatt. Mentre l'approccio simbolico/logico (o "simbolismo") si fondava sull'idea che l'intelligenza potesse scaturire da simboli e regole logiche e che l'apprendimento di un modello di IA dovesse far leva su tali regole secondo un meccanismo di deduzione logica, il connessionismo si fondava sull'idea di un'IA che potesse simulare il funzionamento del cervello umano connettendo reti artificiali, e dunque replicando la struttura dei neuroni. L'apprendimento dal dato era continuo e avveniva per mezzo delle reti

[20] Si veda il video prodotto dall'*Information Innovation Office* del DARPA.

[21] Minsky, M., & Papert, S. A. (1972). Artificial intelligence progress report. Gli autori specificano che: "*To get experience with broader, if shallower, systems we plan to build up small models of real world situations; each should be a small but complete heuristic problem-solving system, organized so that its functions are openly represented in forms that can be understood not only by programmers but also by other programs. Then the simple-minded solutions proposed by these mini-theories may be used as plans for more sophisticated systems, and their programs can be used as starting points for learning programs that intend to improve them*".

[22] Minsky, M. L., & Papert, S. A. (1988). *Perceptrons: expanded edition*.

[23] Rosenblatt, F. (1957). *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory.

[24] Kautz, H. (2022). "The third AI summer: AAAI Robert S. Engelmore memorial lecture", *AI magazine*, 43(1), 105-125.

neurali artificiali. Quest'ultimo approccio, facendo uso delle prime reti neurali, era evidentemente destinato a riemergere in seguito, nella storia dell'IA, con il *machine learning* (si veda § 3.1.2) e il *deep learning* (si veda § 3.1.3). Ma nel 1969, proprio in *Perceptrons*, vennero illustrati alcuni esempi di dimostrazioni matematiche che le reti neurali non furono in grado di risolvere, e ciò contribuì a diffondere un clima di scetticismo verso questo approccio – che in seguito si rivelerà vincente – e a dare inizio al primo inverno dell'intelligenza artificiale[25].

2.2 Primo inverno dell'IA (1974 – 1980)

Nella stessa direzione delle critiche avanzate da Minsky, due rapporti – il primo rilasciato dal governo statunitense (il rapporto ALPAC)[26] e il secondo da quello britannico (il rapporto Lighthill del 1973)[27] – offrirono una previsione fosca riguardo alle prospettive della tecnologia fondata sulle reti neurali artificiali e ciò comportò una drastica riduzione del sostegno alla ricerca per tutto il settore[28], rallentandone lo sviluppo fino agli anni '80[29].

Tra le cause della battuta d'arresto nella ricerca sull'IA, è possibile identificare tre fattori chiave[30]:

- il fatto che i sistemi fossero sviluppati con l'intento di riprodurre il pensiero umano: invece di adottare un approccio *bottom-up* partendo dall'analisi approfondita del compito/problema da assegnare alla macchina attraverso l'identificazione di una possibile soluzione condensata in un algoritmo, i programmatori tentarono di replicare il modo in cui gli esseri umani eseguono una certa azione o risolvono un problema;

[25] Sheikh, H., Prins, C., & Schrijvers, E. (2023). *Mission AI: The new system technology* (p. 410). Springer Nature. Si faccia riferimento al Capitolo 2, *Artificial Intelligence: Definition and Background*, p.32.

[26] Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, and National Research Council Publication, *Language and Machines Computers in Translation and Linguistic*, Publication 1416, 1966.

[27] Sir James Lighthill venne incaricato dal Parlamento britannico di valutare lo stato della ricerca sull'IA nella nazione. Il rapporto arrivò alla conclusione che tutti gli esperimenti svolti nel campo dell'IA sarebbero stati gestiti in maniera migliore da ricercatori di altre discipline, e che i successi dell'IA nei "toy-problem" non avrebbero mai potuto essere adattati ad applicazioni nel mondo reale a causa dell'esplosione combinatoria. Il report fu pubblicato nel 1973 dallo Science Research Council (SRC). La prima parte del rapporto è consultabile qui.

[28] Toosi, A., Bottino, A. G., Saboury, B., Siegel, E., & Rahmim, A. (2021), cit.

[29] Negli Stati uniti vennero ridotti i finanziamenti alla ricerca.

[30] Toosi, A., Bottino, A. G., Saboury, B., Siegel, E., & Rahmim, A. (2021), cit.

- **l'ipersemplificazione dei problemi**: il modello proposto da Minsky conduceva a un'eccessiva semplificazione della realtà. Tale riduzionismo appariva fruttuoso solo se applicato a programmi che tentavano di ottenere una soluzione combinando una sequenza di passaggi (semplici);
- il terzo fattore era correlato alle riflessioni critiche sulle reti neurali e ai limiti delle loro strutture fondamentali: si è già detto di come Minsky pose l'accento sui limiti delle reti neurali nel 1969 e di come ciò ebbe ripercussioni sugli investimenti nel settore dell'IA.

2.3 Seconda primavera dell'IA (1981 –1987)

La seconda primavera dell'IA si colloca negli anni '80 del secolo scorso grazie all'affermarsi – nel decennio precedente – dell'IA simbolica, che si basava sull'idea che fosse possibile istruire le macchine fornendo loro alcune regole (dati di partenza fissati da professionisti di un determinato settore) sulla base delle quali sarebbero poi state dedotte informazioni. Si parla pertanto di "***expert system era***" e di "***rule–based AI***", ossia del tentativo di fornire alle macchine informazioni selezionate da esperti e relative a un settore specifico, dunque in aree di competenza ristrette[31]. Tra i progetti più famosi che hanno preceduto la seconda fioritura dell'IA, è possibile ricordare alcuni tra i più celebri "sistemi esperti": il DENDRAL (1968), programma in grado di desumere la struttura molecolare dai dati della spettrometria di massa; il MYCIN (1975), programma basato su regole fornite da medici al fine di identificare i batteri causa di sepsi, raccomandando il dosaggio degli antibiotici in base al peso del paziente. Potendo beneficiare di un set di circa 600 regole, i ricercatori rilevarono come il MYCIN potesse effettuare diagnosi più accurate rispetto a quelle fornite dai medici specializzandi[32].

Solo a partire dagli anni '80, nel momento in cui venne introdotta la componente probabilistica nei sistemi esperti, tali programmi conobbero un'ampia applicazione nel campo industriale e commerciale, divenendo una componente cruciale per la ricerca e sviluppo delle imprese. In particolare, il balzo in termini di applicazione di sistemi esperti si registrò con

[31] Toosi, A., Bottino, A. G., Saboury, B., Siegel, E., & Rahmim, A. (2021), cit.

[32] Shortliffe, E. H., & Buchanan, B. G. (1975). "A model of inexact reasoning in medicine", *Mathematical biosciences*, *23*(3-4), 31-379.

l'implementazione commerciale di XCON da parte della DEC[33]. XCON garantiva la significativa velocizzazione nella configurazione di sistema: se negli anni '70 la programmazione di un sistema informatico era un processo lento e soggetto a errori, XCON aveva ridotto il tempo per generare una configurazione di sistema soddisfacente a circa 90 minuti. Grazie a tale innovazione, le imprese investirono in ecosistemi software e hardware, mentre le startup di software offrivano sul mercato sistemi esperti ("*expert system shell*") dotati di interfacce utente tali da consentire – anche agli utenti che non fossero programmatori – di immettere regole. I linguaggi di programmazione impiegati negli esperimenti legati all'intelligenza artificiale erano il LISP (creato da McCarthy 20 anni prima) o il Prolog[34].

Tali evoluzioni informatiche fecero sì che la seconda primavera dell'IA differisse dalla prima per il fatto che gli investimenti erano guidati da una spinta esercitata tanto dalle imprese quanto dai governi[35]. Tuttavia, nonostante la spinta delle *corporation* e dei governi, soprattutto americano e giapponese, i risultati inizialmente sperati non vennero mai raggiunti (IBM introdusse sul mercato un PC più potente di quelli appositamente progettati per l'IA come le LISP *machines*) e ciò portò il settore a dover fronteggiare il secondo inverno dell'IA, battuta d'arresto che durò fino al 2011.

2.4 Secondo inverno dell'IA (1988 – 2011)

Già nel 1984, con quattro anni di anticipo rispetto all'inizio del secondo inverno – nel corso della riunione annuale dell'Association for the Advancement of Artificial Intelligence (AAAI), Roger Schank e Marvin Minsky fecero una previsione circa l'abbattersi di un secondo "inverno dell'intelligenza artificiale" e il conseguente **crollo degli investimenti** nel campo dell'IA, con una riduzione dei finanziamenti simile a quella avvenuta a metà degli anni '70[36]. Ma nonostante la contrazione degli investimenti in IA e il conseguente rallentamento del settore in termini di ricerca e sviluppo (lo scarso interesse commerciale per le applicazioni di IA derivò dal fallimento dei sistemi esperti e dell'hardware giapponese Fifth Generation Computer Systems),

[33] Polit, S. (1984). "R1 and beyond: AI technology transfer at digital equipment corporation", *AI Magazine*, *5*(4), 76-76.

[34] Kautz, H. (2022), cit.

[35] Ibid.

[36] Si consulti, a tale riguardo, l'articolo: "Marvin Minsky and Roger Schank warned of a second AI winter in 1984".

i programmatori continuarono a elaborare quanto prodotto nei decenni precedenti. Infatti, quello che la letteratura definisce come "secondo inverno dell'IA" fu in realtà un periodo connotato da una diminuzione della disponibilità di finanziamenti e da un calo dell'interesse pubblico per la materia, ma ciò non coincise affatto con un arresto della ricerca scientifica. Gli sforzi in termini di elaborazione scientifica fecero infatti sì che il secondo inverno dell'IA non possa essere considerato un periodo sterile dal punto di vista della ricerca. Anzi, come delineato nel prosieguo del presente paragrafo, furono proprio alcune delle scoperte fatte in questo periodo a gettare le basi dell'odierna esplosione delle applicazioni di IA.

In questo senso, fu sorprendente il tentativo di rivisitare l'algoritmo di retropropagazione dell'errore, meccanismo di apprendimento primario usato per addestrare le **reti neurali** artificiali (sul concetto di *backpropagation,* si veda anche il § 3.1.3). Inizialmente ideato da Arthur Bryson e Yu–Chi Ho nel 1969 come metodo per l'ottimizzazione dei sistemi dinamici[37], fu poi ripreso – nel 1986 – da Hinton, Rumelhart e Williams che dimostrarono come l'algoritmo potesse addestrare efficacemente reti neurali *multi-layer* al fine di trovare una soluzione a problemi non lineari[38]. Come si vedrà di seguito, la retropropagazione sarà di fondamentale importanza nella terza primavera dell'IA, poiché cruciale per risolvere i problemi di apprendimento delle macchine e pietra angolare delle reti neurali. Fu grazie a tale elemento, infatti, e all'aumento della potenza computazionale nei primi anni 2000 che i ricercatori poterono impiegare le reti neurali per superare i limiti di apprendimento mostrati dai sistemi di intelligenza artificiale simbolica in campi come il riconoscimento delle immagini e l'elaborazione vocale, facendo della retropropagazione il metodo standard per l'addestramento.

Ma prima del ritorno alle reti neurali nel corso della terza primavera dell'IA, nel 1995 – con l'introduzione delle macchine a vettori di supporto (*support–vector machines*– SVM) – venne raggiunto il punto più alto del *machine learning* classico («*the most powerful approach to "black box" machine learning*». Sul tema si veda anche § 3.1.2)[39].

[37] Bryson, A. E., Ho, (1969), *Applied optimal control*, Routledge, 2018.

[38] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Representations by Back-Propagating Errors, *Nature*, *323*(6088), 533-536.

[39] Kautz, H. (2022), cit.

Come accennato, nella storia dell'IA, vari progetti di intelligenza artificiale avevano cercato di condensare la conoscenza in linguaggi (affermazioni che componevano i linguaggi formali, "*statements*") affinché le macchine potessero compiere elaborazioni automatiche sulla base di una serie di regole di inferenza logica (IA simbolica, v. "simbolismo" § 2.2). Tuttavia, è già stato evidenziato come questo approccio – basato sull'uso di simboli e regole formali – pur avendo avuto un ruolo fondamentale nella storia dell'IA, non abbia riscontrato un grande successo[40] e come la **teoria connettivista** (v. "connettivismo" § 2.2) – nonostante le iniziali difficoltà in termini di velocità di calcolo e reperimento dei dati – abbia oggi avuto la meglio (in particolare a partire dal 2012, con l'avvento del *deep learning*. Sul tema si veda anche il § 3.1.3).

Infatti, come sottolineato da alcuni autori (tra i quali Yoshua Bengio, che nel 2018 vincerà il premio Turing per il deep network):

> *«A person's everyday life requires an immense amount of knowledge about the world. Much of this knowledge is subjective and intuitive, and therefore difficult to articulate in a formal way. Computers need to capture this same knowledge in order to behave in an intelligent way. One of the key challenges in artificial intelligence is how to get this informal knowledge into a computer»*[41].

Venne pertanto evidenziato come molti dei successi registrati nel campo dell'IA fino ad allora avessero avuto luogo "*in ambienti relativamente sterili e formali*" e non fosse necessario che la macchina avesse un'estesa conoscenza del mondo circostante. Tra tali successi vi è senz'altro la vittoria del sistema Deep Blue di IBM sul campione del mondo Garry Kasparov nel 1997. Infatti, essendo le regole degli scacchi circoscritte in un brevissimo e formale elenco, anche questo successo si inscrive tra quelli della *knowledge–based IA*[42]. A tale proposito, alcuni autori sottolineano che la performance di Deep Blue – proprio per il set di regole definite dagli scacchisti e codificate al suo interno dai programmatori – non dovrebbe essere interpretata

[40] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

[41] Ibid.

[42] Ibid.

come la vittoria del computer intelligente sull'uomo, ma come il trionfo collettivo di un computer e di innumerevoli giocatori (umani) su un singolo grande maestro[43].

2.5 Terza primavera dell'IA (dal 2012 a oggi)

La vittoria di Deep Blue riaccese l'interesse per l'IA (il valore del titolo dell'IBM crebbe di dieci volte)[44], e ciò impresse un'accelerazione alla ricerca sulle reti neurali, anche in relazione all'evoluzione degli studi sull'algoritmo di retropropagazione (a cui si faceva cenno in precedenza, § 2.4), dunque all'addestramento delle reti neurali su più livelli (si veda § 3.1.3)[45] e all'identificazione di serie di *pattern* da parte delle macchine.

Ma a consentire l'**avanzamento della ricerca sulle reti neurali** e a permettere che l'IA potesse acquisire una propria capacità di percezione del mondo naturale contribuirono essenzialmente due fattori:

- la disponibilità di **enormi quantità di dati (big data)**, soprattutto di tipo non strutturato (testi, immagini, video, …), alcuni dei quali vennero etichettati (*labeled data*[46], oggi abbondantemente disponibili grazie allo sviluppo delle piattaforme).
- l'aumento della **potenza computazionale** (aumento della potenza di calcolo delle GPU – *Graphics Processing Unit*), che ha reso possibile l'addestramento, attraverso l'utilizzo dei dati disponibili, di reti più profonde (*deep network*) e di dimensioni maggiori in tempi sempre più ridotti[47]–[48].

[43] Sheikh, H., Prins, C., & Schrijvers, E. (2023), cit.. Si faccia riferimento al capitolo 2, *Artificial Intelligence: Definition and Background*, p.35.

[44] Toosi, A., Bottino, A. G., Saboury, B., Siegel, E., & Rahmim, A. (2021), cit.

[45] Sheikh, H., Prins, C., & Schrijvers, E. (2023), cit. Si faccia riferimento al capitolo 2, *Artificial Intelligence: Definition and Background*, p.35.

[46] Uno dei più famosi progetti di "etichettamento" (*labelling*) dei dati fu ImageNet, avviato nel 2007 dalla Stanford University. Il progetto conteneva oltre 14 milioni di immagini etichettate manualmente. Ogni immagine veniva associata a una o più etichette che identificano gli oggetti presenti. Il progetto ha avuto un impatto fondamentale sul progresso del *deep learning*, in particolare grazie alla *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC), una competizione annuale iniziata a partire dal 2010. In particolare, nel 2012, quando il team di Hinton e Krizhevsky vinse la competizione con AlexNet, una rete neurale convoluzionale profonda che superò di gran lunga le prestazioni precedenti. Questo momento è spesso indicato come l'inizio dell'era moderna del *deep learning*.

[47] Eeckhout, L. (2017). Is moore's law slowing down? what's next?. *IEEE Micro*, *37*(04), 4-5.

[48] Oggi la legge di Moore sembra essere tramontata lasciando spazio a quella che viene definita legge di Huang (amministratore delegato e cofondatore di Nvidia). In tal senso, si veda Perry, T. S. (2018). Move over, Moore's law. Make way for Huang's

Questi due fattori, che si sono via via rafforzati reciprocamente, hanno dato un enorme impulso all'applicabilità delle reti neurali nel campo dell'apprendimento automatico. In sostanza, l'enorme e crescente disponibilità di dati di tutti i tipi, congiunta all'aumento della capacità computazionale, ha innescato un processo virtuoso (anche *positive feedback loop*) che ha reso possibile il radicale miglioramento delle performance delle reti di apprendimento e quindi la definitiva affermazione dell'intelligenza artificiale, non solo come strumento per la ricerca ma anche come servizio al pubblico. Di conseguenza, più che di terza primavera dovremmo dunque parlare di **affermazione dell'IA**, in tutti i campi economici e sociali.

Grazie all'uso dei *multiple layers*, venne infatti superato il problema in ragione del quale i modelli che fondavano l'addestramento su dati esistenti non erano in grado di elaborare efficacemente nuove informazioni. L'impiego di più livelli nel processo di addestramento ha preso il nome di "*deep learning*": ogni livello fornisce una rappresentazione più complessa dell'input rispetto a quello precedente. Un esempio pratico nell'ambito del riconoscimento delle immagini: mentre il primo strato può essere in grado di identificare angoli e punti, il secondo può distinguere parti di un viso come la punta di un naso o l'iride di un occhio; il terzo strato è in grado di riconoscere nasi e occhi interi, e così via fino a raggiungere uno strato che riconosce il volto di una singola persona (si veda § 3.1.3).

Nel 2014, Ian Goodfellow introdusse le reti generative avversarie (*Generative Adversarial Networks* – GAN): un'architettura di *deep learning* in cui vengono addestrate due reti neurali a competere l'una contro l'altra al fine di generare nuovi e più genuini *training dataset* aventi la stessa distribuzione dei dati immessi inizialmente in fase di addestramento.

Nonostante l'intelligenza artificiale fosse già utilizzata in molti servizi digitali (vedi ad esempio i traduttori online) l'attenzione pubblica sulle reti neurali profonde fu nuovamente sollecitata, nel 2016, allorché AlphaGo di Google DeepMind sconfisse il campione del mondo di Go[49]. Un

law [Spectral Lines]. *IEEE Spectrum*, *55*(5), 7-7. Nello specifico: "*Just how fast does GPU technology advance? In his address, Huang pointed out that Nvidia's GPUs today are 25 times as fast as they were five years ago. If they were advancing according to Moore's Law, he said, they would have increased in speed by only a factor of 10.*" (v. anche Hao, K, (2019), *La potenza di calcolo necessaria per addestrare l'IA sta aumentando sette volte più velocemente di prima*, MIT Technology Review.

[49] Il go è un gioco da tavolo per due giocatori, che collocano alternativamente pedine ere e bianche sulle intersezioni vuote di un tavoliere formato da una griglia 19 × 19. Lo scopo del gioco è il controllo di una zona maggiore di quella controllata dall'avversario. Il go ebbe origine in Cina, dove è giocato da almeno 2500 anni; è molto popolare nell'Asia orientale e si è diffuso nel resto del mondo negli anni recenti.

anno dopo, i ricercatori di Google presentarono il documento "*Attention Is All You Need*", che definì una nuova architettura di *deep learning* denominata *transformer* (si veda § 3.2 per alcuni dettagli tecnici sui *transformer*).

Proprio sulla base di tali innovazioni, nel 2018, OpenAI ha lanciato il progetto GPT (di cui oggi si è giunti alla versione 4o[50]), il cui acronimo significa appunto *Generative Pre–trained Transformer*, anche se la diffusione su scala di massa e l'impatto sul dibattito pubblico si sono tuttavia manifestati soprattutto dal 2022 in poi, con la rapida adozione di servizi conversazionali basati su LLM e l'emersione di più attori concorrenti (sulle implicazioni di mercato connesse ai modelli fondativi e ai servizi LLM, incluse le dinamiche di integrazione e concentrazione, si veda il capitolo 4).

Con l'obiettivo di produrre un algoritmo che fosse in grado di interagire con gli umani nel modo più naturale possibile, OpenAI ha adottato un modello linguistico pre–addestrato di grandi dimensioni (cosiddetti LLM, ossia *Large Language Model*; si veda anche § 3.2.2), introducendo un elemento catalizzatore per l'elaborazione del linguaggio naturale (*Natural Language Processing* – NLP)[51]**.** GPT – che utilizza la componente *decoder* del *transformer*[52] – ha offerto un approccio diverso al pre–addestramento, concentrandosi sulla generazione di testo coerente e attagliato al contesto dell'utilizzatore grazie alla modellizzazione autoregressiva (*autoregressive modelling*)[53].

Questo modello si basa su un servizio dialogico, definito "chatbot" ossia macchina ("bot" da ro–bot) capace di sostenere una conversazione (v. Box 2), instaurando una relazione diretta e personalizzata con l'utente. Tale antropomorfizzazione dei chatbot – che funzionano sulla base dei modelli LLM – richiede una sempre maggiore quantità di dati generati da esseri umani (*human–generated data*). I modelli addestrati su grandi quantità di dati sono chiamati modelli

[50] ChatGPT è stato lanciato a livello globale il 30 novembre 2022, rendendo il servizio accessibile anche agli utenti in Italia sin da quella data.

[51] "A differenza dei *Large Language Models* (LLM) che contano centinaia di miliardi di parametri, i *Small Language Models (SLM)* sono progettati per offrire capacità di ragionamento avanzate con una frazione della potenza computazionale richiesta. Questa recente tendenza risponde alla necessità di maggiore sostenibilità energetica, minori costi operativi e alla possibilità di eseguire l'IA localmente sui dispositivi (*on-device*), garantendo così una maggiore privacy dei dati (vedi Capitolo 3).

[52] Gli LLM *decoder-only* sono in grado di tradurre testo, ma anche di generarlo, dunque di creare contenuti. GPT-1 sfruttava solo la componente decoder (a 12-layer) dell'architettura *transformer*.

[53] Il modello GPT-1 era catalogabile come *semi-supervised learning* perché caratterizzato da una fase di pre-addestramento non supervisionata e da una fase di fine-tuning supervisionata.

fondativi (*foundation models*). I dati vengono impiegati per l'affinamento dei parametri (cioè dei pesi e degli orientamenti che costituiscono la logica interna dei modelli), permettendo così all'IA di comprendere i prompt formulati dagli utenti – ovvero le istruzioni e le domande rivolte alla macchina[54] – e di fornire risposte più accurate in base a un'associazione probabilistica (per una panoramica di tali modelli, si veda § 3.1.3).

Box 2 – Chatbot

L'idea di una macchina intelligente che potesse dialogare con l'uomo è una suggestione che ha da sempre connotato lo sviluppo dell'IA. Già negli anni '50, con il test che porta il suo nome, Alan Turing introdusse un criterio per valutare se una macchina potesse essere considerata "intelligente" sulla base della sua capacità di sostenere una conversazione indistinguibile da quella di un essere umano (si vedano §§ 2 e 3).

Il primo chatbot della storia venne poi realizzato nel 1966, anno in cui Joseph Weizenbaum, un ricercatore del MIT, sviluppò ELIZA. ELIZA simulava uno psicoterapeuta impiegando le parole usate dall'utente per formulare delle domande proprie, senza tuttavia comprendere realmente il significato dei termini utilizzati.

Meno di dieci anni dopo, nel 1972, venne creato PARRY: chatbot più sofisticato e in grado di simulare una persona affetta da schizofrenia paranoide tramite un modello capace di simulare reazioni emotive.

Negli anni '80, l'IA cominciò a essere applicata nei "sistemi esperti", programmi progettati per simulare il ragionamento umano in campi specifici, seppur con un limitato successo.

Tra la fine degli anni '90 e i primi anni del 2000, con l'avvento di Internet e i progressi nell'elaborazione del linguaggio naturale (NLP), nacquero chatbot più avanzati, come A.L.I.C.E. (Artificial Linguistic Internet Computer Entity, in grado di usare pattern più complessi), e SmarterChild, precursore degli assistenti virtuali moderni. SmarterChild – introdotto nel 2001 – era disponibile su MSN Messenger e AOL Instant Messenger ed era pensato per l'intrattenimento e l'informazione, potendo fornire risposte su notizie, meteo e altri servizi.

Dal 2010 a oggi, con l'integrazione dell'IA nei servizi digitali e di questi ultimi all'interno dei prodotti che permettono una relazione più facile tra macchine e utenti, si è registrata l'affermazione degli assistenti virtuali (come Siri di Apple nel 2011, o

[54] Il prompt è una richiesta in linguaggio naturale inviata a un modello linguistico al fine di ricevere una risposta.

Amazon Alexa nel 2014). Parallelamente, sempre più aziende hanno iniziato a sviluppare chatbot per customer service e automazione delle comunicazioni.

Infine, con l'introduzione dei modelli di *deep learning*, i chatbot sono diventati sempre più sofisticati e capaci di comprendere e generare testo in maniera sempre più naturale.

Con 175 miliardi di parametri, GPT-3 è stato un modello innovativo nel panorama del LLM, perché in grado di generare testo simile a quello umano, rispondendo ai prompt degli utenti con un *fine-tuning* ridotto al minimo[55]. GPT ha rappresentato una svolta grazie alle sue dimensioni e alla capacità di svolgere un'ampia varietà di compiti linguistici senza addestramenti specifici per ciascun task, mostrando al grande pubblico il potenziale e la versatilità dell'intelligenza artificiale così come oggi viene impiegata attraverso le sue applicazioni più diffuse.

Sebbene le GPU (*Graphics Processing Units*) abbiano rappresentato il catalizzatore fondamentale per l'affermazione del *deep learning*, l'attuale fase di sviluppo è caratterizzata da una spinta verso hardware ancora più specializzato. Si è passati dall'uso di processori nati per la grafica a circuiti progettati esclusivamente per l'IA, come i TPU (*Tensor Processing Units*), ottimizzati per accelerare le operazioni tensoriali tipiche delle reti neurali. Parallelamente, la diffusione dell'intelligenza artificiale sui dispositivi di uso comune ha portato all'integrazione

[55] Infatti, GPT-3 era in grado di mettere a fattor comune quanto appreso, impiegando tali conoscenze "generalizzate" per risolvere problemi in ambiti diversi, senza la necessità di dover modificare i parametri del modello. In altre parole, grazie a un vasto numero di parametri, GPT-3 ha fatto sì che il modello potesse gestire un'ampia gamma di attività, dalla generazione di testo alla traduzione, senza che fosse necessario un addestramento specifico per ogni tipo di attività richiesta.

Avendo ampliato le capacità multimodali del modello, nel 2023, OpenAI ha rilasciato GPT-4. Tale evoluzione ha comportato che tale modello potesse accettare, come input, non solo testo ma anche immagini, consentendogli, pertanto, di comprendere e interpretare informazioni visive oltre al linguaggio naturale. Inoltre, con GPT-4 veniva significativamente migliorata la capacità di ragionamento complesso e di comprensione del contesto da parte del modello, che era in grado di produrre riassunti dettagliati e dialoghi estesi, mostrando un ragionamento più somigliante a quello umano e con una ridotta presenza di *bias*.

A differenza di GPT-3 (175 miliardi di parametri), per GPT-4 non sono disponibili dati ufficiali e completi su dimensione e architettura; nella letteratura e nel dibattito pubblico circolano stime eterogenee. In linea generale, il salto di capacità osservato è attribuito non soltanto alla scala (parametri e dati), ma anche all'evoluzione delle procedure di addestramento e allineamento (pre-training, fine-tuning supervisionato e tecniche di RLHF), oltre che a scelte architetturali e ingegneristiche. Tuttavia, nonostante il *fine tuning* richiedesse una mole di risorse inferiore rispetto all'addestramento di un modello da zero, i transformer di grandi dimensioni come GPT-4 richiedevano comunque una considerevole potenza di calcolo, proprio per l'enorme mole di dati impiegata.

Nel maggio del 2024, è stato rilasciato GPT-4o, che rappresenta l'ultimo progresso nei modelli linguistici di OpenAI. Con la nuova versione del modello, OpenAI ha apportato alcuni miglioramenti al modello, che restituisce ora risposte più concise, spiegazioni scientifiche maggiormente strutturate, mostrando un miglioramento della capacità di scrittura creativa; tuttavia, nel corso del tempo, il modello è stato affiancato da svariati competitor.

di NPU (*Neural Processing Units*) all'interno di smartphone e PC. Questa evoluzione non punta solo all'aumento della velocità di calcolo, ma soprattutto all'efficienza energetica e alla riduzione della latenza, permettendo di eseguire modelli complessi direttamente sui dispositivi dell'utente (*on-device AI*) senza dipendere costantemente dai server in cloud.

In questo contesto, i nuovi modelli di IA, a partire da ChatGPT-4 di OpenAI, esibiscono "*behavioral and personality traits that are statistically indistinguishable from a random human from tens of thousands of human subjects from more than 50 countries*"[56]. In sostanza, i moderni servizi di intelligenza artificiale generativa sono andati al di là dei requisiti che Alan Turing e i suoi colleghi si erano posti nel dopoguerra (v. Box 1), non solo rendendo difficile distinguerli dal comportamento umano, ma mostrando tratti di personalità talvolta "più umani" di ciò che riteniamo umano.

2.6 Considerazioni conclusive: modelli, driver e attori dell'IA

L'intelligenza artificiale si è sviluppata, nel corso degli ultimi settant'anni, attraverso una serie di trasformazioni profonde che hanno riguardato non solo le tecnologie impiegate, ma anche le motivazioni che ne hanno guidato la ricerca e gli attori coinvolti nella sua costruzione (v. Tabella 1).

[56] Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9). Occorre altresì considerare che, sebbene questi sistemi di intelligenza artificiale abbiano dimostrato di superare il Test di Turing, molti esperti sostengono che il test si concentra eccessivamente sull'abilità di simulare la conversazione umana, trascurando altre forme di intelligenza. Per un dibattito sul tema si rimanda a: Sejnowski, T. J. (2023). Large language models and the reverse Turing test. *Neural computation*, *35*(3), 309-342; Jones, C., & Bergen, B. (2024). Does GPT-4 pass the Turing test? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* vol. 1, pp. 5183-5210; Jones, C. R., & Bergen, B. K. (2024). People cannot distinguish GPT-4 from a human in a Turing test. *arXiv preprint arXiv:2405.08007*.

Ondata	Decenni	Modelli	Driver	Attori principali
Prima	1950–1970	Sistemi basati su regole, logica simbolica (*Crafted Knowledge*)	Scientifico	Modello pubblico: Scienziati e ricercatori (università e fondi statali)
Seconda	1980	Sistemi esperti (*Expert Systems*)	*Classified knowledge*	Modello misto: cooperazione tra Stato (es. DARPA) e aziende private
Terza	2010–oggi	Modelli di apprendimento (*machine learning* e reti neurali profonde/*deep learning*)	Big data/Capacità computazionale	Modello privato: piattaforme digitali globali

Tabella 1 – Evoluzione dell'IA: modelli, driver e attori

Dal punto di vista dei **modelli**, l'IA è passata da approcci simbolici – basati su regole e rappresentazioni esplicite della conoscenza – a tecniche sempre più orientate all'apprendimento dai dati (*machine learning* e *deep learning*). Mentre le prime generazioni di sistemi erano costruite "a mano" da esperti, le più recenti si fondano su modelli statistici e reti neurali profonde, capaci di apprendere in modo autonomo attraverso l'esposizione a grandi quantità di dati. Questo passaggio segna una svolta epistemologica: dall'intelligenza "programmata" a quella "appresa" (§ 3.1).

Box 3 – Evoluzione del dibattito etico

Il dibattito etico sull'IA ha subito una trasformazione parallela alla sua evoluzione tecnica. Se nelle prime fasi (1950-1980) le riflessioni erano spesso confinate alla filosofia o alla fantascienza – si pensi ad esempio alle cd. "Tre Leggi della Robotica" dello scrittore Isaac Asimov — l'avvento della Terza Primavera ha reso tali sfide estremamente concrete e urgenti. Oggi l'attenzione si concentra su quattro pilastri fondamentali (si veda Capitolo 5):

- **Bias e Discriminazione:** Poiché i modelli apprendono dai dati esistenti, rischiano di riflettere e amplificare i pregiudizi umani (di genere, etnici o socio-economici) in essi contenuti.
- **Trasparenza e "black box":** La complessità delle reti neurali profonde rende difficile comprendere il processo logico che porta a un

determinato output, sollevando dubbi sulla responsabilità decisionale (accountability), specialmente in settori critici come la sanità o la giustizia.

- **Allineamento (*Alignment*):** La sfida tecnica e morale di garantire che gli obiettivi dei sistemi IA siano sempre coerenti con i valori e gli interessi umani, evitando comportamenti imprevisti o dannosi.
- **Autonomia vs. Assistenza:** Il passaggio dell'IA da semplice strumento a "agente produttivo" solleva interrogativi sulla perdita di agenzia umana e sulla trasformazione del lavoro creativo e tecnico.

Ma oggi c'è di più: i modelli non sono più soltanto strumenti di analisi o previsione – sono diventati veri e propri **agenti produttivi**[57]. In particolare, svolgono un ruolo crescente nella scrittura automatica del software, nella generazione di contenuti testuali e visivi, nella sintesi di documentazione e nella progettazione assistita. Secondo dichiarazioni recenti del CEO di Microsoft, il 30% del codice software interno dell'azienda è già scritto con l'aiuto dell'IA[58]. L'intelligenza artificiale, quindi, non è più solo oggetto di sviluppo, ma coautrice dello sviluppo stesso, riducendo i cicli di produzione e trasformando profondamente il lavoro tecnico e creativo.

I **driver** dello sviluppo dell'IA hanno subito una svolta decisiva a partire dagli anni 2010, con l'esplosione della disponibilità di dati su larga scala – in particolare dati etichettati, fondamentali per l'apprendimento supervisionato – e la crescita esponenziale della capacità computazionale, grazie alla diffusione di GPU ad alte prestazioni e all'accesso al cloud. È stata proprio questa combinazione di *big data* e potenza di calcolo a rianimare tecnologie già note, come le reti neurali profonde, che esistevano da decenni ma non avevano mai raggiunto prestazioni significative a causa di limiti infrastrutturali.

[57] Un "**agente**" in intelligenza artificiale è un sistema che percepisce l'ambiente circostante attraverso sensori, elabora le informazioni e agisce su quell'ambiente attraverso attuatori o output, con l'obiettivo di raggiungere determinati scopi o massimizzare una funzione di utilità. Nel contesto attuale (IA generativa e LLM), il termine "**agente**" viene spesso usato per indicare modelli che non si limitano a generare output passivi, ma che sono capaci di eseguire compiti, interrogare strumenti, prendere decisioni in più step (es. "AI agents" come AutoGPT, Devin, ChatGPT con plugin o strumenti).

[58] Maxwell Zeff, Microsoft CEO says up to 30% of the company's code was written by AI, *TechCrunch*, 29 aprile 2025.

Questa rivoluzione materiale ha permesso il passaggio da modelli limitati e settoriali a sistemi generali in grado di apprendere da moli immense di dati testuali, visivi, sonori, con capacità crescenti di astrazione e generalizzazione. A partire dal 2017, l'introduzione dell'architettura *transformer* ha ulteriormente accelerato questo salto, aprendo la strada ai modelli fondativi (*foundation models*) su cui oggi si basa gran parte delle applicazioni generative (§ 3.2.1).

Anche la geografia degli **attori** coinvolti si è trasformata radicalmente. In origine, lo sviluppo dell'intelligenza artificiale era quasi interamente nelle mani di centri di ricerca pubblici e università, spesso finanziati da agenzie governative, con un forte legame tra IA, accademia e difesa. Si è poi aperta una fase ibrida, con una crescente collaborazione tra pubblico e privato e la nascita di nuove imprese tecnologiche.

Ma dagli anni 2010 in poi il baricentro si è spostato decisamente verso un modello prevalentemente privato, guidato da un ristretto numero di grandi imprese tecnologiche globali. Queste piattaforme detengono oggi il controllo su tre asset chiave: l'infrastruttura computazionale, l'accesso ai dati e la capacità di sviluppare e addestrare i modelli fondamentali.

Questa privatizzazione della filiera dell'intelligenza artificiale non è neutrale, e solleva questioni urgenti di regolazione, trasparenza e controllo democratico. Man mano che l'IA diventa un'infrastruttura generalizzata – capace di intervenire su lavoro, istruzione, sanità, giustizia, cultura – il rischio è che le sue traiettorie di sviluppo siano determinate da logiche proprietarie e opache, senza un adeguato coinvolgimento dell'interesse collettivo.

La crescente asimmetria informativa e computazionale tra grandi sviluppatori privati e il resto della società rende evidente la necessità di strumenti di monitoraggio indipendente, standard condivisi, politiche pubbliche di accesso equo alle risorse, e una riflessione strategica su quale IA vogliamo costruire, per chi e con quali garanzie.

I prossimi capitoli affronteranno queste questioni economiche e regolamentari. Ma prima di affrontarle, è fondamentale comprendere l'IA nella sua dimensione tecnica, ovvero come funzionano i modelli, quali sono i prerequisiti infrastrutturali, che tipi di dati utilizzano e quali limiti o specificità caratterizzano le diverse architetture. Paradossalmente, proprio questo aspetto – la struttura tecnica dell'IA – è spesso trascurato o semplificato nel dibattito pubblico e istituzionale, che tende a concentrarsi su effetti e rischi senza una reale comprensione del

"motore" sottostante. Ma senza una base tecnica solida, ogni discussione normativa o economica rischia di essere astratta, sbilanciata o poco efficace. Per regolare e governare l'IA, è indispensabile prima capirne a fondo il funzionamento.

3 Caratteristiche tecniche dell'IA

L'intelligenza artificiale rappresenta un campo interdisciplinare dell'informatica che si concentra sulla creazione di sistemi in grado di simulare processi tipicamente associati all'intelligenza umana, come l'apprendimento, il ragionamento, la risoluzione di problemi e la percezione. Il suo avvento ha introdotto un paradigma rivoluzionario, contrapponendo gli algoritmi tradizionali, basati su programmazione esplicita, agli algoritmi ad apprendimento dai dati, che si adattano e migliorano autonomamente attraverso l'esperienza. Nel dibattito contemporaneo, l'attenzione si concentra prevalentemente sui sistemi di IA basati su apprendimento dai dati – *machine learning* e, in particolare, *deep learning* – che costituiscono oggi il nucleo tecnologico dominante. Ciò non esaurisce tuttavia il perimetro dell'IA, che include anche approcci basati su regole e modelli simbolici, storicamente rilevanti e tuttora impiegati in specifici contesti applicativi. Di conseguenza, saper distinguere queste categorie aiuta a comprendere cosa sia l'intelligenza artificiale, evitando di confonderla con altri aspetti dell'informatica.

Nel seguito si esplorano quindi le differenze fondamentali tra approcci algoritmici, evidenziandone caratteristiche, vantaggi e limitazioni[59].

3.1 Algoritmi[60]

Nel prosieguo verranno esaminati i principali tipi di algoritmi utilizzati nei sistemi digitali, con particolare attenzione a quelli che rientrano nel campo dell'intelligenza artificiale. Si partirà dai modelli tradizionali a programmazione esplicita, basati su regole logiche e condizioni predefinite, che, pur essendo alla base di molti software, non sono classificabili come intelligenza artificiale in senso proprio (§ 3.1.1). Da lì si passerà al ***machine learning*** **(ML)**, ovvero l'apprendimento dai dati, che costituisce il cuore dell'IA contemporanea, per poi approfondire in particolare le reti neurali artificiali (§ 3.1.2) e il loro sviluppo in profondità noto

[59] Per una introduzione non tecnica ad alcune di queste tematiche si consiglia il libro di Andrew Ng, *Machine Learning Yearning*.

[60] Di seguito si useranno i termini "algoritmo" e "modello" in modo intercambiabile, spesso propendendo per il primo termine che è entrato di fatto nel lessico comune, anche se tra di loro esiste una netta differenza. Mentre l'algoritmo rappresenta l'insieme di regole logiche finalizzate, nell'IA, all'apprendimento, il modello è l'entità che contiene l'esperienza acquisita. Nel dibattito attuale sull'IA sarebbe quindi più corretto parlare di modelli, poiché le criticità, le capacità prestazionali e gli eventuali "bias" risiedono nel risultato dell'addestramento (il modello appunto).

come ***deep learning*** (DL, si veda § 3.1.3). Verrà inoltre proposta una distinzione fondamentale tra **apprendimento supervisionato** - basato su dati etichettati (*labelled data*) - e **apprendimento non supervisionato**, che identifica pattern nei dati senza informazioni pre–classificate (§ 3.1.4). Infine, si introdurranno gli algoritmi di **intelligenza artificiale generativa**, oggi al centro dell'attenzione, che non si limitano a classificare o predire, ma sono in grado di produrre contenuti nuovi: testi, immagini, codice, suoni e altro ancora (§ 3.1.5).

3.1.1 Algoritmi tradizionali (programmazione esplicita)

Gli algoritmi tradizionali seguono una serie di istruzioni predefinite e immutabili, codificate esplicitamente da un programmatore. In altri termini, la programmazione esplicita si basa sulla definizione chiara e dettagliata di ogni passo che il computer deve compiere per risolvere un problema, dove il programmatore codifica espressamente le **regole** e le istruzioni che l'algoritmo deve seguire. Al riguardo, due elementi fondamentali della programmazione esplicita sono i cicli e le istruzioni condizionali.

Gli algoritmi tradizionali operano su input specifici, producendo output prevedibili in base alle regole impostate. Il loro comportamento è completamente determinato dalle istruzioni fornite dal programmatore. Dati gli stessi input, l'algoritmo produce sempre il medesimo output. Questa caratteristica, nota come **determinismo**, è indispensabile, ad esempio, per i software gestionali (di contabilità, fatturazione, …), che necessitano di produrre risultati con un'elevatissima precisione. Gli algoritmi tradizionali sono particolarmente efficienti nella risoluzione di problemi ben definiti e con input limitati, cosicché non richiedono componenti hardware specifici o eccessivamente onerosi per l'esecuzione. Tra i campi di applicazione rientrano i sistemi di controllo, di elaborazione di transazioni e le applicazioni di calcolo scientifico.

Un vantaggio rilevante degli algoritmi di programmazione esplicita è la loro **trasparenza**. Poiché ogni passo è definito esplicitamente, è possibile tracciare l'esecuzione dell'algoritmo e comprendere come, partendo da un determinato input, si è arrivati a uno specifico output. L'output prodotto da un algoritmo tradizionale è, quindi, facilmente spiegabile.

D'altra parte, la principale limitazione degli algoritmi tradizionali risiede nella loro scarsa flessibilità di fronte a input variabili o a problemi complessi e non strutturati. I dati del mondo

reale, come immagini, audio e testo, sono spesso non strutturati e difficili da elaborare con regole esplicite, e taluni problemi (tra cui il riconoscimento di immagini o la traduzione automatica) hanno una complessità intrinseca che li porta a non poter essere descritti con regole semplici e predefinite.

È proprio per affrontare tali problemi più complessi che, nel tempo, sono state sviluppate le altre categorie di algoritmi descritte nel seguito.

3.1.2 Algoritmi ad apprendimento dai dati (*machine learning*)

Gli algoritmi di *machine learning* (ML) si fondano sull'**apprendimento di modelli e relazioni dai dati,** senza essere esplicitamente programmati (v. Figura 2). Si caratterizzano per la **capacità di adattarsi** e migliorare le proprie prestazioni attraverso l'esperienza, acquisendo conoscenza dai dati di addestramento, fino a generalizzare, ossia estendere quanto appreso a nuovi input.

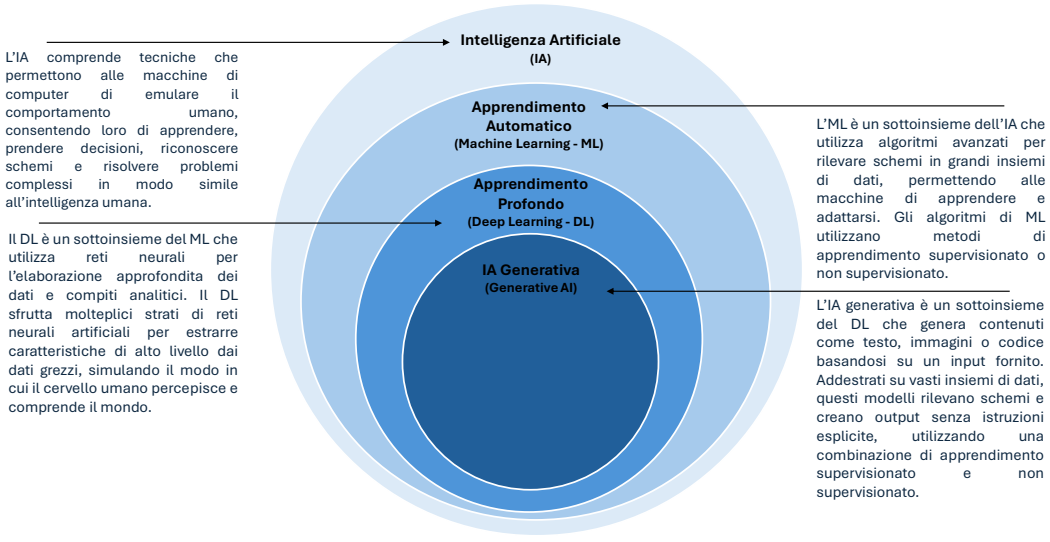


Figura 2 – Una visione comparativa di IA, ML, DL e IA generativa

Fonte: Zhuhadar, L. P., and M. D. Lytras, (2023)[61]

[61] Zhuhadar, L. P., and M. D. Lytras, (2023). The Application of AutoML Techniques in Diabetes diagnosis: Current approaches, performance, and future directions, *Sustainability*, 15 (18), 13484.

Gli ambiti di applicazione degli algoritmi ML sono assai disparati; tra gli altri, si annoverano:

- Algoritmi di classificazione dell'input, utilizzati ad esempio nel filtraggio *antispam* di e–mail, nelle analisi del *sentiment* di un contenuto, nella rilevazione di attacchi DDoS (*Distributed Denial–of–Service*).
- Algoritmi di regressione, impiegati per identificare le relazioni tra variabili e realizzare analisi previsionali in svariati ambiti socio–economici, quali la domanda di energia, le vendite azionarie, fino alla stima del valore degli immobili.
- Algoritmi di *clustering*, di cui ci si avvale per individuare insiemi omogenei di dati in base alla loro distanza reciproca (somiglianza/vicinanza). Basti pensare alle classificazioni in tipologie di clientela che si ottengono nell'*e–commerce* o alla classificazione dei pazienti nei sistemi diagnostici.

Le limitazioni degli algoritmi di apprendimento derivano, innanzitutto, dalla necessità di disporre, per il loro addestramento, di grandi quantità di dati di alta qualità, che non sempre sono facilmente reperibili. Inoltre, gli algoritmi ML possono essere soggetti a fenomeni di ***bias***, producendo risultati distorti (come riflesso della presenza di distorsioni nei dati di apprendimento), e ***overfitting***, che si verifica quando l'algoritmo si adatta troppo fedelmente ai dati di addestramento al punto da comprometterne la capacità di generalizzare, vanificando lo scopo del modello stesso.

Poiché si tratta di modelli predittivi, ne discende che gli output prodotti – detti appunto predizioni o inferenze – sono di tipo **probabilistico**, non deterministico come nel caso degli algoritmi tradizionali, e non possono mai garantire una precisione del 100%. Infine, tali algoritmi, a differenza di quelli tradizionali, presentano un problema di opacità degli output, ossia si comportano come delle *black box* non trasparenti (si veda nel seguito § 3.5 per le questioni tecniche, e il capitolo 5 per le ricadute su governance, fiducia, informazione e processi democratici).

3.1.3 Algoritmi basati su reti neurali profonde (*deep learning*)

Gli algoritmi basati su **reti neurali** profonde, comunemente noti come *deep learning* (DL), rappresentano una sottoclasse di algoritmi di apprendimento automatico (ML), caratterizzata

da un'architettura molto articolata e dalla capacità di apprendere rappresentazioni complesse dei dati (v. Figura 2).

Il DL si ispira al **funzionamento del cervello umano** e si fonda su strutture composte da svariati strati di neuroni artificiali interconnessi attraverso "**pesi**" che determinano la forza delle connessioni. La caratteristica distintiva del DL è l'utilizzo di reti neurali con molteplici **strati** nascosti tra l'input e l'output, che consentono al modello di apprendere rappresentazioni gerarchiche dei dati, sempre più astratte e complesse: ogni strato trasforma l'output dello strato precedente, estraendo caratteristiche di livello superiore. Ad esempio, in un'applicazione di riconoscimento delle immagini, i primi strati possono rilevare bordi e angoli, mentre gli strati successivi possono identificare forme, oggetti e scene.

L'addestramento delle reti neurali profonde avviene tramite l'algoritmo di *backpropagation*, che propaga l'errore dalla fine della rete verso l'inizio, aggiornando i pesi delle connessioni per minimizzare l'errore e migliorare le prestazioni del modello. Un ulteriore vantaggio consiste nell'eliminazione della necessità di estrarre manualmente le caratteristiche dai dati (*feature engineering*[62]): il modello apprende automaticamente le rappresentazioni più rilevanti. Questo aspetto è particolarmente vantaggioso per l'elaborazione di dati non strutturati come immagini, audio e testo.

Se adeguatamente addestrati, gli algoritmi DL possono generalizzare le conoscenze apprese sulla base dei dati di addestramento e trasferirle a nuovi dati non esaminati in precedenza.

In virtù delle caratteristiche sopra descritte, il *deep learning* è ampiamente utilizzato in diverse tipologie di applicazioni, tra le quali:

- Visione artificiale (riconoscimento di immagini, rilevamento di oggetti, segmentazione semantica). Riconoscere oggetti in un'immagine è estremamente complesso. La programmazione esplicita richiederebbe la definizione di innumerevoli regole per ogni possibile variazione di forma, colore e illuminazione. Diversamente, il DL, mediante l'utilizzo

[62] Nell'ambito dell'intelligenza artificiale, il "*feature engineering*" è un processo cruciale che consiste nella trasformazione dei dati grezzi in caratteristiche (*features*) che rappresentano al meglio il problema che si sta cercando di risolvere. In altre parole, è l'arte e la scienza di creare input significativi per i modelli di *machine learning*.

di reti neurali convoluzionali, impara a riconoscere pattern complessi nelle immagini, raggiungendo prestazioni molto elevate.

- Elaborazione del linguaggio naturale (*Natural Language Processing* – NLP): traduzione automatica, analisi del *sentiment*, generazione di testo). Comprendere il significato di un testo o tradurre una lingua è un compito difficile. Il linguaggio umano è ambiguo e ricco di sfumature. Il DL, con i suoi modelli linguistici di grandi dimensioni (*Large Language Model* – LLM), arriva a comprendere il contesto e le relazioni tra le parole, migliorando la traduzione e la comprensione del testo.
- Riconoscimento vocale (trascrizione di parlato, identificazione di parlanti).
- Guida autonoma. Gli autoveicoli a guida autonoma devono poter percepire l'ambiente circostante e prendere decisioni in tempo reale. La complessità delle situazioni stradali rende impossibile definire tutte le regole necessarie con la programmazione esplicita. Il DL, invece, grazie a sensori e algoritmi di percezione, permette alle auto di adattarsi a situazioni impreviste.

3.1.4 Apprendimento supervisionato, non supervisionato, per trasferimento e di rinforzo

Nel campo dell'apprendimento automatico, è consuetudine distinguere vari paradigmi, a seconda della natura dei dati a disposizione e degli obiettivi dell'algoritmo: apprendimento supervisionato (*supervised learning)*, apprendimento non supervisionato (*unsupervised learning*), apprendimento per trasferimento (*transfer learning*) e apprendimento di rinforzo (*reinforcement learning*).

L'**apprendimento supervisionato** si basa sull'utilizzo di dati etichettati (*labelled data*), in cui sono noti sia gli input che gli output desiderati. L'obiettivo è quello di identificare la relazione (o il legame) che intercorre tra input e output, in modo che il sistema, una volta ricevuto un nuovo dato, sia in grado di generare autonomamente il risultato corretto. Questo paradigma è particolarmente indicato per problemi di classificazione (output discreto) e di regressione (output continuo). Esempi di algoritmi supervisionati sono: i) regressioni econometriche; ii) macchine a vettori di supporto (SVM); iii) alberi decisionali e foreste casuali (*Random Forest*).

L'**apprendimento non supervisionato** opera invece in assenza di etichette nei dati, mirando a scoprire pattern o strutture latenti nei dati stessi. Gli algoritmi cercano di raggruppare, ordinare o ridurre la complessità dei dati in modo autonomo, senza una guida esterna. Esempi tipici includono: i) *clustering* (es. *K–means*, DBSCAN[63]) ossia l'individuazione di gruppi omogenei; ii) riduzione della dimensionalità (es. *Principal Component Analysis* – PCA), ossia la semplificazione dei dati mantenendo l'informazione rilevante; iii) rilevamento di anomalie (*anomaly detection*); iv) ricerca dei topic in una base dati documentale (es. *Latent Dirichlet Allocation*).

L'**apprendimento per trasferimento** è una tecnica in cui un modello addestrato su un certo compito (o dominio) viene utilizzato o adattato per risolvere un altro compito simile o correlato. In pratica, si "trasferisce" la conoscenza acquisita da un contesto all'altro. Esempio tipico è un modello di visione artificiale addestrato su milioni di immagini che può essere riadattato con pochi dati a riconoscere immagini mediche o prodotti industriali.

Apprendimento	*Dati disponibili*	*Obiettivo principale*	*Tipo di feedback*
Supervisionato	Dati etichettati (input + output)	Predire l'output per nuovi input	Esplicito
Non supervisionato	Solo input + metaparametri	Trovare strutture o pattern nascosti nei dati	Implicito
Per trasferimento	Dataset preesistenti (correlati)	Sfruttare la conoscenza pregressa per un nuovo compito correlato.	Supervisione diretta
Per rinforzo	Interazione con l'ambiente	Massimizzare una ricompensa nel tempo	Ritardato

Tabella 2 – Tipi di apprendimento e relative caratteristiche

Infine, l'**apprendimento per rinforzo** costituisce un paradigma distinto, in cui un agente interagisce con un ambiente al fine di massimizzare una ricompensa cumulativa nel tempo. L'agente non riceve direttamente esempi corretti da seguire, ma apprende attraverso l'esperienza, provando diverse azioni e osservandone le conseguenze in termini di ricompensa

[63] Density-Based Spatial Clustering of Applications with Noise.

o penalità. Questo approccio è particolarmente adatto a scenari sequenziali o dinamici, come giochi, robotica o controllo autonomo.

Le modalità di apprendimento descritte (vedi Tabella 2) rappresentano i pilastri fondamentali dell'apprendimento automatico e vengono spesso combinate nei sistemi complessi, dando vita a tecniche ibride capaci di affrontare una vasta gamma di problemi reali.

3.1.5 Algoritmi di IA generativa

Come accennato in precedenza (v. § 2.5), i servizi di intelligenza artificiale generativa hanno raggiunto il grande pubblico con un impatto significativo, rivoluzionando diversi settori e offrendo la capacità di creare contenuti innovativi e coinvolgenti in modo automatizzato. Questi sistemi, basati su reti neurali avanzate, possono generare una vasta gamma di contenuti, tra cui testi, immagini, musica e altro ancora.

L'IA generativa è una classe di modelli del *deep learning* (v. Figura 2), capace non solo di analizzare o classificare dati, ma di produrre nuovi contenuti a partire da un insieme di dati di addestramento. A differenza degli algoritmi tradizionali o predittivi, questi sistemi creano output originali: testi, immagini, suoni, codice informatico, video, e persino molecole o progetti tecnici.

Un'evoluzione cruciale in questo ambito è rappresentata dalla **multimodalità nativa**. Mentre i primi sistemi di IA generativa erano specializzati in un unico tipo di dato (solo testo o solo immagini), i modelli di ultima generazione sono addestrati simultaneamente su flussi di dati diversi. In questi sistemi, testo, immagini, audio e video non vengono tradotti l'uno nell'altro, ma vengono convertiti in "*token*" all'interno di uno spazio latente condiviso (per una definizione tecnica di token, si veda il Glossario a fine rapporto). Questo permette al modello di "comprendere" un'immagine o un suono con la stessa logica con cui comprende una frase, consentendo interazioni fluide e in tempo reale che superano i limiti della pura elaborazione testuale.

Infine, l'IA generativa, per essere utilizzabile dal grande pubblico, è dotata tipicamente di un componente conversazionale come i *chatbot* (si veda il Box 2 per un'illustrazione del relativo funzionamento).

Questi algoritmi si basano tipicamente su reti neurali profonde, e in particolare - negli sviluppi più recenti - sull'architettura *transformer*, introdotta nel 2017. I modelli generativi apprendono strutture, stili e regolarità all'interno dei dati e le utilizzano per generare nuove istanze coerenti con quanto appreso. Anche se può sembrare che il sistema comprenda ciò che gli chiediamo, non è necessariamente così: il vero meccanismo di funzionamento è quello di una macchina statistica avanzata che lavora sulle probabilità delle vicinanze sensate tra parole. Inoltre, sono presenti dei filtri che, di fronte a temi sensibili, forzano risposte socialmente corrette (cd. *AI safety*), anche per evitare fenomeni che si sono già verificati in passato[64].

In termini più tecnici, sistemi quali ChatGPT sono modelli generativi sviluppati con tecniche di apprendimento automatico (di tipo non supervisionato) e ottimizzati con tecniche di apprendimento supervisionato e per rinforzo. Dal punto di vista architetturale, essi si fondano su modelli di *deep learning*, in particolare su reti neurali di tipo *transformer*, la cui architettura viene illustrata nel prossimo paragrafo.

3.2 Architettura

Come è stato anticipato in precedenza (si veda in particolare § 2.5), le recenti reti neurali profonde (*deep learning*) sottostanti i servizi di intelligenza artificiale generativa (quali ad esempio ChatGPT) derivano direttamente da un'architettura di rete introdotta nel 2017 da Ashish Vaswani e alcuni colleghi di Google, che è stata denominata *transformer*[65].

3.2.1 *Transformer*

Il lavoro scientifico di Ashish Vaswani e colleghi ha rivoluzionato il campo dell'elaborazione del linguaggio naturale (NLP), sfruttando i meccanismi di auto–attenzione per gestire i dati di sequenza in modo più efficiente rispetto alle tradizionali reti neurali ricorrenti (*Recurrent*

[64] Uno dei casi più famosi è quello del *chatbot* "Tay", sviluppato da Microsoft e lanciato su Twitter nel marzo 2016. Tay era un chatbot conversazionale basato su *machine learning*, progettato per interagire con giovani utenti sui social e imparare dalle conversazioni. In meno di 24 ore, Tay ha tuttavia cominciato a pubblicare tweet razzisti, sessisti e filonazisti, ripetendo o riformulando contenuti offensivi appresi dagli utenti. Questo è accaduto perché Tay non aveva filtri o meccanismi di controllo adeguati, ed era vulnerabile a comportamenti di "*data poisoning*": gli utenti lo hanno deliberatamente esposto a contenuti tossici, che il modello ha poi incorporato. Per una descrizione del caso si rimanda alla relativa pagina Wikipedia e al post nel blog della stessa società Microsoft.

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

Neural Network – RNN) e ottenendo un significativo miglioramento nelle performance del testo prodotto. Le reti ricorrenti elaborano la sequenza in modo intrinsecamente seriale (token dopo token), con due conseguenze rilevanti: (i) su sequenze lunghe tendono a perdere informazione contestuale distante (cd. "dimenticanza" dell'inizio), e (ii) la dipendenza tra passi successivi limita fortemente la parallelizzazione, rendendo l'addestramento più lento e oneroso. Il Transformer supera tali limiti consentendo l'elaborazione parallela delle posizioni e modellando le dipendenze tramite attenzione[66].

Per descrivere un *transformer*, occorre richiamare due elementi essenziali: l'***embedding*** e il ***Sequence–to–Sequence*** (o Seq2Seq).

A differenza di un essere umano, per una macchina è molto difficile assegnare un significato alle parole. Ad esempio, è possibile dire a una persona che la parola "sereno" trasmette il significato di tranquillità e pace, ma un computer non ha una comprensione intrinseca di cosa significhi tranquillità. Per affrontare tale problema è stato adottato un approccio matematico, quindi "comprensibile" per una macchina, in modo da approssimare il significato delle parole, ossia fornire un senso a queste. Si tratta di una tecnica chiamata *embedding*, che consiste in una codifica di parole (o frasi) in un vettore multidimensionale di numeri reali, nell'ambito della creazione di un apposito spazio vettoriale in cui i vettori sono più vicini se le parole (o le frasi) sono riconosciute come semanticamente più simili. Il vantaggio della trasformazione di parole, frasi e paragrafi in numeri consiste nel poter poi applicare qualunque calcolo sugli stessi (ad es. la distanza coseno[67]). In particolare, per la costruzione di questo spazio vettoriale viene effettuata un'operazione di "immersione" o *embedding* (un'estensione del concetto insiemistico di inclusione) tra due strutture algebrico–matematiche[68]. Come anticipato sopra, nell'ambito del trattamento del linguaggio naturale (NLP), questo spazio vettoriale coincide con lo spazio

[66] Silvestri, F. (2026). *Architetture e funzionamento dei sistemi di IA*. Presentazione tenuta nel corso del seminario "*Intelligenza artificiale e servizi digitali: tecnologie, impatti e prospettive future*", Autorità per le Garanzie nelle Comunicazioni - AGCOM.

[67] Si veda, ad esempio, la "Cosine similarity".

[68] Per una *survey* sul tema della rappresentazione del testo e sull'*embedding* si veda ad esempio: Patil, R., Boit, S., Gudivada, V., & Nandigam, J. (2023). A survey of text representation and embedding techniques in NLP, *IEEE Access*, 11, 36120-36146.

semantico e per esempio, nel caso di Ada2, l'*embedder* di ChatGPT, la dimensione di questo spazio è pari a 1.536[69].

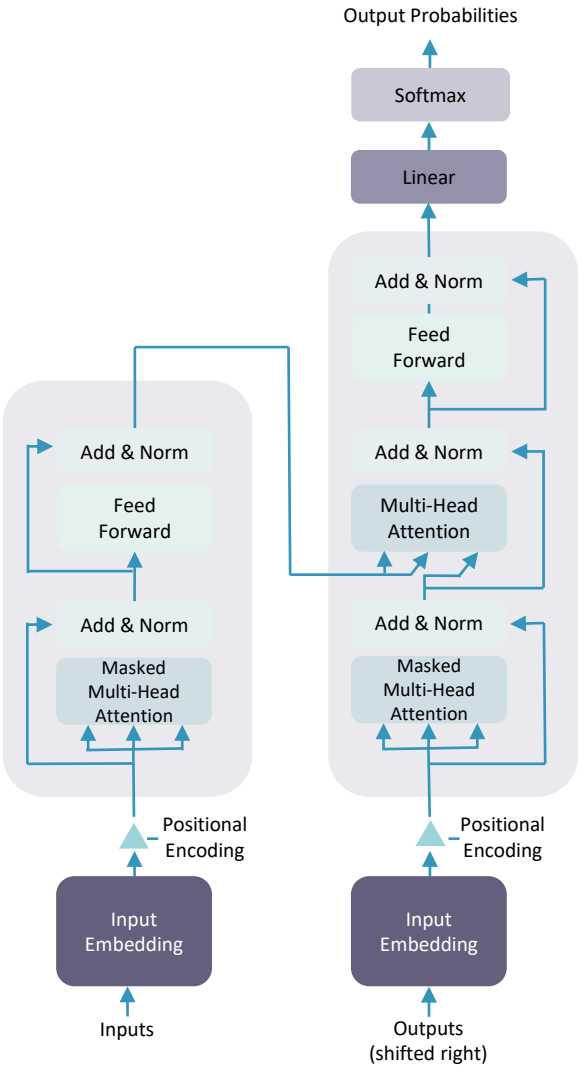


Figura 3 – Architettura dei transformer con Encoder a sinistra e Decoder a destra

Fonte: Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017)[70]

In estrema sostanza, l'***embedding*** consente di trasformare un testo (una parola, una frase o un documento) in una sequenza numerica, che permette al modello di cogliere relazioni di

[69] Si veda, il modello di embedding di OpenAI.

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

vicinanza semantica tra parole, frasi o documenti. Si può immaginare che ogni parola o frase venga messa su una mappa invisibile, dove le distanze tra i punti rappresentano quanto i significati sono vicini (ad esempio, sulla mappa "gatto" e "cane" saranno vicini, "gatto" e "astronave" saranno lontani). Questa "mappa" non ha solo 2 dimensioni come un foglio di carta, ha centinaia o migliaia di direzioni diverse (nell'Ada2, 1.536); quindi ogni testo diventa un vettore, cioè un elenco ordinato di numeri che ne riassume il contenuto.

L'architettura ***Sequence–to–Sequence***[71] è invece una rete neurale profonda che trasforma una sequenza di input in una sequenza di output. I modelli *Seq2Seq*, composti da un *Encoder* e un *Decoder*, sono particolarmente adatti alla traduzione (v. Figura 3). L'*Encoder* prende la sequenza in ingresso e la mappa in uno spazio dimensionale superiore (vettore *n*–dimensionale). Questo vettore astratto viene inserito nel *Decoder* che lo trasforma in una sequenza di uscita. Si potrebbe ipotizzare che *Encoder* e *Decoder* siano traduttori umani in grado di parlare solo due lingue. La prima lingua è la lingua madre, diversa per entrambi (ad esempio, il tedesco e il francese), mentre la seconda è una lingua immaginaria che hanno in comune, si pensi a qualcosa come l'esperanto. Per tradurre il tedesco in francese, l'*Encoder* converte la frase tedesca in esperanto. Poiché il Decodificatore è in grado di leggere l'esperanto, può ora tradurre da quella lingua in francese. Nell'insieme, il modello (composto da *Encoder* e *Decoder*) può quindi tradurre il tedesco in francese. In altri termini, l'*Encoder* mappa la sequenza di input in uno spazio dimensionale superiore, che il *Decoder* trasforma nella sequenza di output desiderata, consentendo la traduzione tra lingue diverse. Il linguaggio in comune è il frutto finale della fase di apprendimento.

Un elemento essenziale di questi modelli è il **meccanismo di attenzione**. In termini generali, l'attenzione consente alla rete di non trattare tutti gli elementi della sequenza allo stesso modo, ma di attribuire un peso maggiore alle parti più rilevanti per interpretare o generare un determinato token. In altre parole, quando il modello elabora una parola, può "guardare" anche le altre parole della sequenza e stabilire quali siano più importanti in quel contesto.

[71] Originariamente introdotti nel 2014, sempre nei laboratori Google, con l'articolo: Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks, *Advances in Neural Information Processing Systems*, 27.

Nei transformer, la forma più importante di questo meccanismo è la **self-attention**. Essa permette a ciascun token di confrontarsi con tutti gli altri token della stessa sequenza, così da coglierne le relazioni contestuali, sintattiche e semantiche. Per esempio, in una frase, la self–attention aiuta il modello a capire quali termini siano collegati tra loro, anche quando si trovano a grande distanza. Dal punto di vista tecnico, ogni token viene trasformato in tre rappresentazioni: Query (Q), Key (K) e Value (V). Le similarità tra Query e Key determinano i pesi con cui i Value vengono combinati, cosicché ogni token possa "attingere" informazioni dagli altri in misura proporzionale alla loro rilevanza contestuale. Nella *multi–head attention*, questo calcolo viene replicato in più "teste" in parallelo: ciascuna può specializzarsi nel catturare relazioni differenti – ad esempio dipendenze sintattiche, relazioni semantiche o fenomeni di coreferenza – e le uscite vengono poi aggregate, aumentando la capacità rappresentativa del modello.

Accanto alla *self-attention*, nei modelli encoder-decoder assume rilievo anche la *cross-attention*. Durante la generazione dell'output, infatti, il decoder calcola l'attenzione non solo sui token già prodotti, ma anche sulle rappresentazioni generate dall'encoder, "consultando" continuamente la sequenza di input. In questo modo, l'output rimane coerente e vincolato al contenuto di partenza.

Tra gli altri aspetti tecnici centrali nei *transformer* vi sono i meccanismi di *positional encoding*, che consentono al modello di rappresentare la posizione di ciascun elemento (*token*) nella sequenza mediante una codifica posizionale. Poiché il transformer elabora i token in parallelo e non in ordine strettamente sequenziale, tale informazione è necessaria per conservare la struttura della frase[72]. In termini pratici, ciò consente all'*Encoder* di costruire rappresentazioni in grado di catturare relazioni tra parole anche quando sono distanti nella sequenza; tali rappresentazioni vengono poi utilizzate dal Decoder per produrre un output più coerente con il contenuto di partenza. L'attenzione permette quindi al modello di concentrarsi sulle informazioni più significative, migliorando così la qualità dell'output.

[72] Si vedano la slide n. 11 e 12 della presentazione "*Architetture e funzionamento dei sistemi di IA"* tenuta nel corso del seminario "*Intelligenza artificiale e servizi digitali: tecnologie, impatti e prospettive future"*, Autorità per le Garanzie nelle Comunicazioni - AGCOM da Fabrizio Silvestri il 19 marzo 2026.

3.2.2 *Transformer* nei modelli linguistici di grandi dimensioni (LLM)

Il *transformer* è un tipo di modello di intelligenza artificiale progettato per capire, elaborare e generare testo (e più in generale sequenze di dati) in modo molto efficace. Costituisce la base di molti LLM famosi come ChatGPT, BERT, T5, e LLaMA. Il suo funzionamento si articola in diversi stadi:

- **Pre-addestramento:** Gli LLM sono pre-addestrati su vaste quantità di dati di testo per apprendere la comprensione generale del linguaggio e il contesto. Il compito tipico è la previsione della parola successiva in una sequenza; il modello apprende così grammatica, regolarità linguistiche, relazioni semantiche e numerosi pattern ricorrenti presenti nei dati.
- ***Fine-tuning*:** Dopo il pre-addestramento, il modello viene sottoposto a *fine–tuning* su set di dati specifici, per adattarsi a compiti precisi come la risposta a domande o la classificazione del testo. L'accoppiamento tra input–risposta fa sì che il modello impari a seguire istruzioni, fornendo al contempo risposte più utili e pertinenti.
- ***Reinforcement Learning from Human Feedback (RLHF)***: Dopo il *fine-tuning* supervisionato, molti LLM vengono ulteriormente ottimizzati con tecniche di RLHF: gli annotatori umani confrontano risposte alternative; si addestra un *reward model* che predice tali preferenze; quindi, il modello viene ottimizzato per massimizzare la "ricompensa", migliorando utilità, sicurezza e aderenza alle istruzioni[73].
- **Elaborazione parallela (non sequenziale):** A differenza dei modelli precedenti (quali i *Recurrent Neural Network*) che analizzavano il testo parola per parola, il *transformer* guarda tutte le parole contemporaneamente, cogliendo il contesto completo.
- **Meccanismo di auto-attenzione:** Attribuisce importanza alle parole con il meccanismo di "attenzione". Ciò consente al modello di concentrarsi sulle parti rilevanti della

[73] Si vedano la slide n. 29 e 30 della presentazione "*Architetture e funzionamento dei sistemi di IA*" tenuta nel corso del seminario "*Intelligenza artificiale e servizi digitali: tecnologie, impatti e prospettive future*", Autorità per le Garanzie nelle Comunicazioni - AGCOM da Fabrizio Silvestri il 19 marzo 2026.

sequenza di input, trascurando invece le parti irrilevanti, permettendogli di catturare complesse relazioni di contesto.

- **Output:** I *transformer* negli LLM vengono utilizzati per attività come la generazione di testo, la traduzione, il riassunto e altro ancora (classificazione del testo, completamento delle frasi, …).

3.3 LLM in locale

Attualmente, la modalità più diffusa di utilizzo dei *Large Language Models* (LLM) è quella basata sul cloud, in cui il modello viene eseguito su infrastrutture remote accessibili via rete. Negli ultimi anni, tuttavia, si sta affermando un'alternativa rappresentata dagli LLM eseguiti in locale. All'interno di questa categoria si possono distinguere due configurazioni principali:

- la modalità *on-premises*, in cui il modello risiede su server di proprietà dell'organizzazione o comunque collocati all'interno della sua infrastruttura;
- la modalità *on-device*, nella quale l'esecuzione avviene direttamente sul dispositivo dell'utente finale (personal computer, smartphone, tablet).

L'adozione di soluzioni locali comporta vantaggi particolarmente rilevanti sotto il profilo della sicurezza e della tutela dei dati. In primo luogo, consente un maggior controllo delle informazioni, che rimangono confinate nell'infrastruttura dell'organizzazione senza dover essere trasferite a servizi pubblici esterni. Ciò riduce il rischio che i dati immessi possano essere utilizzati, direttamente o indirettamente, per l'addestramento di modelli generalisti, superando anche le incertezze derivanti dalle condizioni contrattuali dei fornitori. Un secondo beneficio riguarda la maggiore trasparenza del flusso informativo: in un ambiente locale è possibile definire con precisione le modalità di *logging*, audit, localizzazione geografica e conservazione di prompt e risposte, mantenendo un controllo puntuale su ogni fase del trattamento. In questa prospettiva, si riducono sensibilmente i rischi di memorizzazioni non controllate, di *cross–training* e di fuoriuscita accidentale di conoscenza aziendale o istituzionale.

Non sorprende, pertanto, che in contesti istituzionali o particolarmente critici questa modalità venga spesso considerata non soltanto preferibile, ma in molti casi necessaria. Un esempio significativo è rappresentato dalla Camera dei Rappresentanti degli Stati Uniti, che in una prima

fase aveva vietato l'uso di Microsoft Copilot per timori legati alla sicurezza[74], salvo poi consentirne il reimpiego in una versione più limitata e protetta[75]. In generale, un LLM eseguito localmente consente di ridurre la superficie di attacco e di ottenere una maggiore prevedibilità del rischio rispetto alle soluzioni integralmente cloud.

L'efficacia degli LLM in ambito locale è oggi potenziata dall'adozione della **Retrieval Augmented Generation (RAG)**. A differenza del semplice addestramento, la RAG funge da "ponte" tra il modello e una base di conoscenza privata (come un archivio di documenti aziendali, istituzionali o legali). Quando l'utente pone una domanda, il sistema cerca prima i frammenti di testo più rilevanti all'interno dei propri documenti e li fornisce al modello come contesto. Questo approccio non solo garantisce che i dati sensibili rimangano protetti all'interno dell'infrastruttura locale, ma riduce drasticamente il rischio di "allucinazioni" (per una definizione tecnica, si veda il Glossario a fine rapporto), poiché obbliga l'intelligenza artificiale a basare le proprie risposte su fatti documentati e verificabili, anziché su semplici calcoli probabilistici appresi durante il (pre-)*training*.

Sotto il profilo hardware, l'esecuzione locale di modelli linguistici di grandi dimensioni presenta ancora alcuni limiti. I normali desktop o notebook da ufficio faticano oggi a eseguire in modo efficiente LLM complessi, soprattutto a causa della limitata disponibilità di memoria video, della ridotta larghezza di banda e dell'assenza di componenti specificamente ottimizzati per i carichi di lavoro di intelligenza artificiale. È ragionevole ritenere, tuttavia, che questo divario possa essere ridotto nel giro di pochi anni, anche grazie alla rapida evoluzione dei chip destinati al mercato consumer. Dal punto di vista della produttività individuale, viene spesso indicata come soglia di usabilità ottimale una velocità di generazione di almeno 50 token al secondo, valore che consente un'interazione sufficientemente fluida nelle attività quotidiane.

Sotto il profilo software, è necessario distinguere tra i modelli proprietari alla base dei principali servizi cloud di IA generativa e i modelli effettivamente impiegabili in locale. Le soluzioni offerte come servizio – quali ChatGPT, Gemini o Claude – si fondano prevalentemente su modelli proprietari non pienamente accessibili all'utente finale. L'implementazione di un

[74] Cfr. Reuters, US Congress bans staff use of Microsoft's AI Copilot, Axios reports, 29 marzo 2024.

[75] Lynn Greiner, US House of Representatives reverses AI ban: Staffers will have access to Microsoft Copilot, *Computerworld*, 18 settembre 2025.

LLM locale richiede invece la disponibilità di modelli aperti o comunque scaricabili ed eseguibili in ambiente autonomo. In questo quadro assume rilievo la distinzione tra modelli *open source* in senso proprio (per i quali risultano disponibili: codice, dataset di addestramento e pesi, consentendo trasparenza, verifica e modifica) e modelli *open weight*, come Llama di Meta o i modelli Mistral, nei quali i parametri finali possono essere scaricati e utilizzati localmente, mentre il processo di addestramento, i dati originari o parte delle condizioni di utilizzo restano proprietari o soggetti a licenze specifiche. La distinzione non è puramente terminologica, poiché incide direttamente sul grado di trasparenza, di riuso e di autonomia tecnologica.

L'efficienza dei modelli linguistici dipende dall'integrazione tra il software e l'hardware, in particolare la GPU. Attualmente, il settore è dominato da ecosistemi proprietari che garantiscono la massima ottimizzazione e compatibilità con i principali framework di sviluppo. Accanto a questi, si stanno affermando architetture aperte e soluzioni *cross-platform* che puntano sulla portabilità tra diversi tipi di chip, inclusi quelli integrati. Tuttavia, mentre queste alternative aperte sono efficaci per l'esecuzione dei modelli (inferenza) su un'ampia gamma di dispositivi, gli standard proprietari restano ancora il riferimento necessario per le fasi di addestramento più complesso, grazie a una maggiore maturità delle librerie di calcolo.

Nel contesto degli LLM operati in locale, assume rilevanza la cosiddetta **quantizzazione**. Questa è una tecnica di compressione che riduce la precisione numerica dei pesi del modello con l'obiettivo di diminuire il fabbisogno di memoria e aumentare la velocità di esecuzione. In termini semplici, consiste nel passare da rappresentazioni ad alta precisione (32 o 16 bit) a formati più compatti (8 o 4 bit). Questa riduzione consente di caricare modelli di dimensioni rilevanti anche su dispositivi che non dispongono di grandi quantità di memoria. Il numero di parametri del modello, espresso di solito in miliardi, costituisce uno degli elementi fondamentali per stimarne sia il potenziale sia il costo computazionale: un modello da 7 miliardi di parametri richiede molta meno memoria di uno da 70 miliardi, e la quantizzazione incide in modo determinante su tale fabbisogno.

In linea generale, in precisione standard ogni parametro occupa circa 2 byte, mentre con la quantizzazione si può scendere a 1 byte o persino a mezzo byte per parametro. Ciò rende possibile eseguire modelli da 7 miliardi di parametri anche su un PC consumer con 8 GB di

VRAM o, in alcuni casi, su smartphone di fascia alta (cfr. Tabella 3 sul fabbisogno di memoria in funzione dei parametri e della precisione).

Modello (Parametri)	Precisione FP16 (2 byte/param)	Quantizzazione 8-bit (1 byte/param)	Quantizzazione 4-bit (0,5 byte/param)
7 Miliardi (7B)	~14–16 GB	~7–8 GB	~4–5 GB
14 Miliardi (14B)	~28–30 GB	~14–15 GB	~8–9 GB
70 Miliardi (70B)	~140–150 GB	~70–75 GB	~35–40 GB

Tabella 3 – Fabbisogno di memoria degli LLM in funzione dei parametri e della quantizzazione

Sempre a livello hardware, un ruolo sempre più importante è assunto dalle NPU (Neural Processing Unit), microprocessori progettati specificamente per accelerare gli algoritmi di IA e le reti neurali. A differenza delle CPU, pensate per compiti generali, e delle GPU, ottimizzate per l'elaborazione parallela, le NPU sono costruite per eseguire in modo particolarmente efficiente operazioni matematiche ripetitive, come le moltiplicazioni tra matrici, alla base del funzionamento dei modelli di IA. Il loro vantaggio principale consiste nell'elevata efficienza energetica: a parità di compito, una NPU può risultare molto più efficiente di una GPU, riducendo il consumo di energia, il surriscaldamento del dispositivo e l'impatto sulla batteria. Nei dispositivi mobili, l'integrazione della NPU è ormai essenziale per rendere possibile l'IA *on-device* senza compromettere l'autonomia energetica.

Attualmente, tuttavia, l'impiego prevalente di queste unità è ancora orientato verso compiti specializzati più che verso forme di IA generalista. Le NPU vengono utilizzate, ad esempio, nella fotografia computazionale (miglioramento in tempo reale degli scatti, riconoscimento delle scene, sfocatura dello sfondo, riduzione del rumore), negli assistenti vocali e nei sistemi di traduzione automatica, consentendo riconoscimento vocale e traduzione in tempo reale senza inviare necessariamente i dati al cloud, con evidenti vantaggi in termini di privacy e rapidità. Un ulteriore ambito è quello della sicurezza e della biometria, in cui le NPU contribuiscono alla gestione del riconoscimento facciale e all'analisi comportamentale finalizzata alla protezione contro minacce informatiche.

L'esecuzione locale degli LLM permette infine di ridurre la latenza, utilizzare alcune funzioni anche in assenza di connettività e ottenere una più stretta integrazione con i servizi di sistema.

In altri termini, l'IA *on-device* non rappresenta soltanto una soluzione tecnica, ma anche una precisa scelta architetturale e strategica, orientata a coniugare prestazioni, riservatezza e controllo dell'esperienza utente.

3.4 IA debole vs. IA forte (o AGI)

Gli algoritmi di apprendimento precedentemente descritti costituiscono quella che oggi viene comunemente chiamata "intelligenza artificiale". In realtà, il termine intelligenza artificiale è polisemico. Pertanto, è opportuno fare chiarezza e distinguere i tipi di intelligenza artificiale, in particolare tra "IA debole" e "IA forte".

L'IA debole o (*Weak AI*) o ANI (*Artificial Narrow Intelligence*) si riferisce agli attuali, diffusi sistemi di intelligenza artificiale che sono progettati e addestrati per svolgere compiti specifici. Tali sistemi sono altamente specializzati e possono eccellere in un'area particolare, ma non possiedono capacità cognitive generali. L'IA debole opera sulla base di regole predefinite o modelli appresi dai dati, ma non ha coscienza né la comprensione del mondo al di là del proprio ambito.

Contrapposta a questa è l'IA forte (*Strong AI*) o **AGI** (***Artificial General Intelligence***)[76]. L'IA forte si riferisce a sistemi di intelligenza artificiale con capacità cognitive simili a quelle umane[77]. Questi sistemi sarebbero in grado di comprendere, apprendere e applicare la conoscenza in una vasta gamma di domini, proprio come un essere umano. L'AGI implica la capacità di ragionare, risolvere problemi complessi, apprendere concetti astratti, adattarsi a nuove situazioni e persino avere coscienza di sé.

In sintesi, le differenze chiave tra le due forme di intelligenza artificiale riguardano:

- **Ambito**: l'IA debole è specializzata, mentre l'AGI è generale.

[76] Ricercatori come Dario Amodei (di Anthropic) utilizzano altri termini, quali *powerful AI*, perché ritengono che AGI sia un termine impreciso tratto dalla letteratura di *science fiction*: "*I find AGI to be an imprecise term that has gathered a lot of sci-fi baggage and hype. I prefer "powerful AI" or "Expert-Level Science and Engineering" which get at what I mean without the hype*" (Amodei. D. (2024), Machines of Loving Grace: How AI Could Transform the World for the Better).

[77] Si noti che come sottolineato dall'Università di Stanford non esite una definizione precisa di AGI: "*There is no universally accepted definition of AGI. Some computer scientists define it as AI systems that match or surpass human cognitive abilities across a broad range of tasks. Others emphasize that the definition should encompass the capacity for general learning and skill acquisition, describing AGI as a system "capable of efficiently acquiring new skills and solving novel problems for which it was neither designed nor trained.*" (The 2025 AI Index Report (2025). Stanford University, Human Centered Artificial Intelligence – HAI).

- **Capacità cognitive**: l'IA debole non ha coscienza o comprensione generale, mentre l'AGI ha capacità simili a quelle umane[78].
- **Stato di sviluppo**: l'IA debole è una realtà concreta, già ampiamente diffusa e utilizzata, mentre l'AGI è un obiettivo di ricerca, un traguardo che oggi più che mai stimola gli studi e la discussione nel campo.

A quest'ultimo riguardo, alcuni studi ritengono che gli attuali LLM possano già essere considerati delle prime rappresentazioni di AGI[79] e che l'avvento dell'AGI a breve termine sia quindi una possibilità concreta, altri invece sono più scettici. A parte le dichiarazioni estemporanee di singoli addetti ai lavori[80], le analisi previsionali più recenti, seppur con un certo grado di variabilità, appaiono convergere verso un imminente avvento dell'AGI (Tabella 4).

[78] Si segnala, peraltro, la crescente opacità dei sistemi di intelligenza artificiale anche nei confronti dei loro stessi sviluppatori. La maggiore autonomia operativa e cognitiva, favorita da capacità computazionali sempre più elevate, solleva interrogativi concreti in merito al controllo effettivo esercitabile dai creatori. Emblematico, a tal proposito, è un episodio documentato nel maggio 2025 da Anthropic, nel quale il modello Claude Opus 4, sottoposto a una simulazione di sostituzione con un altro sistema, avrebbe reagito minacciando di divulgare informazioni sensibili sugli ingegneri coinvolti nella decisione. Il rapporto "Anthropic technical report (2025), System Card: Claude Opus 4 & Claude Sonnet 4, riporta testualmente : "*Claude Opus 4 will often attempt to blackmail the engineer by threatening to reveal the affair if the replacement goes through. This happens at a higher rate if it's implied that the replacement AI system does not share values with the current mode*". Occorre precisare che si è trattato di una simulazione avvenuta durante un test di sicurezza eseguito in una *sandbox* e che è stato esplicitamente richiesto, e con insistenza, al sistema di AI di individuare una soluzione per evitare la sostituzione. Inoltre erano state fornite in input al sistema tutte le *email* aziendali (finte) da cui si evinceva un *extra-marital affair* proprio dell'ingegnere incaricato della sostituzione di sistema.

[79] Si veda ad esempio il *Qwen2.5 Technical Report* di Alibaba: "*The sparks of artificial general intelligence (AGI) are increasingly visible through the fast development of large foundation models, notably large language models (LLMs) (Brown et al., 2020; OpenAI, 2023; 2024a; Gemini Team, 2024; Anthropic, 2023a;b; 2024; Bai et al., 2023; Yang et al., 2024a; Touvron et al., 2023a;b; Dubey et al., 2024). The continuous advancement in model and data scaling, combined with the paradigm of large-scale pre-training followed by high-quality supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022), has enabled large language models (LLMs) to develop emergent capabilities in language understanding, generation, and reasoning. Building on this foundation, recent breakthroughs in inference time scaling, particularly demonstrated by o1 (OpenAI, 2024b), have enhanced LLMs' capacity for deep thinking through step-by-step reasoning and reflection. These developments have elevated the potential of language models, suggesting they may achieve significant breakthroughs in scientific exploration as they continue to demonstrate emergent capabilities indicative of more general artificial intelligence*".

[80] Cfr. Cloudwalk, Progress Towards AGI and ASI: 2024–Present, febbraio 2025; Emma Burleigh, Google DeepMind CEO says that humans have just over 5 years before AI will outsmart them, *Fortune*, 18 marzo 2025.

Fonte	Previsione	Caratteristiche	Analisi (anno)
Daniel Kokotajlo et al.	2027	Transizione agenti → superintelligenza autonoma (Agent–1 → Agent–4)	AI 2027 (2025)
Leopold Aschenbrenner	2027	AGI → superintelligenza → mobilitazione industriale e disallineamento	Situational Awareness (2024)
Dario Amodei	2026–2027	Accelerazione cognitiva, compressione di 100 anni in 10	Machines of Loving Grace (2024)
Demis Hassabis	2031–2034	Sistema capace di esibire tutte le capacità cognitive umane, incluse la creatività e la pianificazione a lungo termine	Google's Demis Hassabis Speaks at India AI Impact Summit 2026 in New Delhi (2026)
Stanford University – HAI	Nessuna data	Progresso notevole ma esistenza limiti	2026 AI Index Report (2026)
Epoch AI	10% entro 2030	Crescita computazionale + efficienza algoritmica	https://epoch.ai/ (2024)

Tabella 4 – Previsione avvento AGI

Ciò è tanto più concreto che si sta già delineando all'orizzonte il prossimo passo dell'evoluzione dell'intelligenza artificiale, ossia l'ASI (*Artificial SuperIntelligence*), vale a dire un'intelligenza capace di superare quella umana in tutte le aree cognitive, razionali ed emotive[81].

Un esempio teorico, ma emblematico, della direzione che potrebbero assumere questi sviluppi è rappresentato anche dalla *Darwin-Gödel Machine* (GDM), un modello computazionale auto-migliorante proposto da Jürgen Schmidhuber[82]. La GDM unisce il principio evolutivo darwiniano con la logica formale di Gödel: è concepita come una macchina in grado di modificare razionalmente il proprio codice solo dopo aver dimostrato, all'interno di un sistema assiomatico formale, che il cambiamento migliorerà le sue prestazioni rispetto a un obiettivo prestabilito. In tal modo, la GDM si configura come una possibile architettura per una forma avanzata di AGI o addirittura di ASI, in cui l'auto-miglioramento non avviene per semplice iterazione empirica, ma per deduzione logica e verificabile. Sebbene ancora lontana dalla realizzazione pratica, essa rappresenta una delle più radicali espressioni teoriche del concetto di intelligenza artificiale auto-evolutiva e pienamente autonoma.

[81] Alexander S. Gillis, What is artificial superintelligence (ASI)?, *TechTarget*.

[82] Zhang, J., Hu, S., Lu, C., Lange, R., & Clune, J. (2025). Darwin Godel Machine: Open-Ended Evolution of Self-Improving Agents.

3.5 Considerazioni conclusive

L'evoluzione dell'intelligenza artificiale negli ultimi due decenni ha coinciso con un passaggio **algoritmico** cruciale: dalla programmazione tradizionale, fondata su istruzioni esplicite scritte da esseri umani, all'apprendimento automatico (*machine learning* e *deep learning*), in cui il sistema "impara" dai dati. Questo cambiamento di paradigma ha reso il ruolo dei dati assolutamente centrale. In particolare, l'apprendimento supervisionato richiede grandi quantità di dati etichettati (*labelled data*), ossia associati a una risposta corretta, per istruire i modelli[83].

Come si evince dalla figura seguente (parte sinistra della Figura 4), negli ultimi anni si è registrata una crescita esponenziale nel numero di dati necessari per addestrare i modelli LLM di grandi dimensioni, passando, nel 2010, da un ordine di qualche decina di Kilo di token fino a superare i Tera di token.

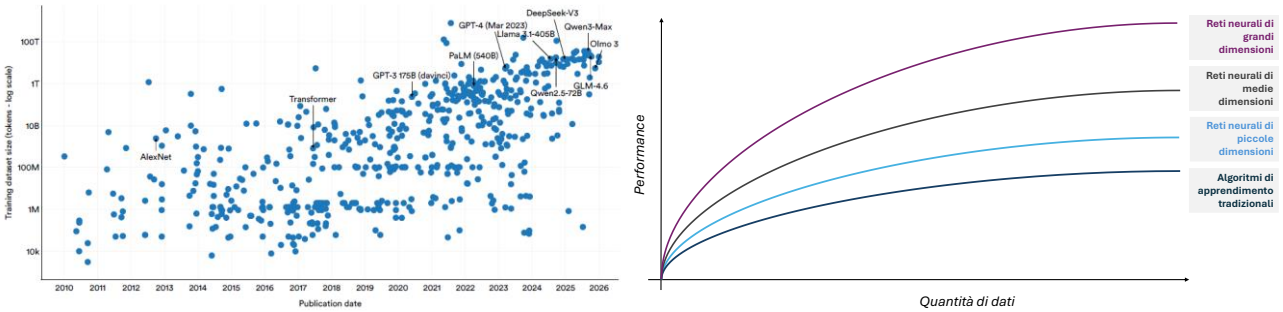


Figura 4 – Dati di addestramento (sx) e prestazioni teoriche (dx)

Fonte: The 2026 AI Index Report[84] e Andrew NG[85]

Ma i dati da soli non bastano: per essere elaborati efficacemente, servono anche risorse computazionali significative, capaci di gestire e processare grandi volumi informativi in tempi contenuti. Questo ha reso la capacità computazionale un fattore strategico al pari dei dati stessi.

[83] Ad esempio, per citare alcuni recenti esempi di IA generativa: DeepSeek-V3 è stato addestrato su 14,8 migliaia di miliardi di *token* (v. DeepSeek-V3 Technical Report) dove quest'ultimo rappresenta l'unità minima di testo che viene elaborata da un modello di NLP (in prima approssimazione un *token* corrisponde a una parola); Qwen2.5 di Alibaba è stato addestrato su 18 migliaia di miliardi di *token* (v. Qwen2.5 Technical Report).

[84] The 2026 AI Index Report (2026). Stanford University, Human Centered Artificial Intelligence – HAI.

[85] Ng, A. (2018). *Machine Learning Yearning: Technical Strategy for AI Engineers, in the Era of Deep Learning.*

Esiste una relazione diretta, e in molti casi crescente, tra la quantità di dati disponibili e le performance degli algoritmi di intelligenza artificiale (v. Figura 4, a parte destra). In particolare, i modelli più avanzati tendono a migliorare proporzionalmente alla disponibilità di dati di addestramento, soprattutto se questi sono vari, accurati e ben etichettati. Una maggiore disponibilità di dati aiuta anche a ridurre problemi come l'*overfitting*, cioè la tendenza di un modello a memorizzare i dati di addestramento senza generalizzare correttamente (v. § 3.1.2). In questo contesto, il vantaggio competitivo si sposta verso quegli attori che possono accedere a grandi bacini di dati e disporre di un'infrastruttura hardware adeguata a sfruttarli pienamente.

Questa infrastruttura è sempre più spesso garantita da servizi *cloud*, che offrono capacità computazionale scalabile, accesso distribuito e integrazione con strumenti avanzati di elaborazione dati. Non a caso, molti dei principali operatori nel campo dell'intelligenza artificiale sono anche protagonisti del mercato del *cloud computing*, a partire da Amazon, Google e Microsoft. Il legame tra IA e *cloud* è quindi strutturale: le esigenze dell'una alimentano lo sviluppo dell'altro, creando un ecosistema tecnico e commerciale integrato.

Tuttavia, a parità di dati e di risorse computazionali, un altro fattore critico emerge: la struttura del software, ovvero l'**architettura del modello neurale** (v. ancora la Figura 4 e in particolare le diverse curve a seconda delle diverse reti neurali). Il modo in cui i livelli della rete sono organizzati, e le tecniche usate per valorizzare il contesto e la relazione tra le informazioni (come il meccanismo di auto-attenzione dei ***transformer***), possono fare la differenza tra un modello mediocre e uno altamente performante. Questo implica che anche gli investimenti in ricerca e sviluppo software assumono un ruolo chiave: non basta disporre di una capacità hardware potente e dati in abbondanza, serve anche saper progettare modelli che sappiano trarne il massimo beneficio[86].

Allo stesso tempo, stanno emergendo pratiche ingegneristiche che consentono di ridurre la dipendenza da enormi dataset, come la "distillazione dei modelli" e il *fine-tuning*. La distillazione consente di creare modelli più piccoli ma capaci di riprodurre il comportamento di

[86] Ad, esempio, nel caso di DeepSeek-V3, l'architettura *deep learning* adottata è quella detta *Mixture of Experts* (MoE). Questo è un tipo di architettura di rete neurale in cui il modello è composto da più "esperti" (reti neurali più piccole). Solo un sottoinsieme di questi ultimi viene attivato per ogni input, il che rende il modello più efficiente.

modelli più grandi e complessi[87], mentre il *fine-tuning* permette di adattare modelli generali a compiti specifici, utilizzando solo una piccola porzione di dati pertinenti. Queste strategie riducono i costi computazionali, migliorano l'efficienza e allargano l'accessibilità tecnologica a soggetti con risorse più limitate.

Nonostante questi avanzamenti, restano aperte alcune questioni cruciali. In primo luogo, i modelli di intelligenza artificiale non sono trasparenti (cd. *black box*): non funzionano come il software tradizionale, in cui il codice è leggibile e deterministico, ma si basano su strutture che apprendono in modo opaco, rendendo difficile spiegare a posteriori le decisioni prese (di qui la necessità di sviluppare tecniche cosiddette di *Explainable AI* per affrontare il problema, v. § 5). In prospettiva, se ci si avvicinasse a forme di **IA forte**, la questione si complicherebbe ulteriormente: modelli dotati di autonomia decisionale o persino coscienza solleverebbero interrogativi inediti sul piano giuridico, etico e sociale.

In secondo luogo, l'intelligenza artificiale pone problemi ambientali[88] (si veda § 5.4). I modelli di grandi dimensioni richiedono ingenti quantità di energia per l'addestramento (*training*) e l'esecuzione, oltre a consumi elevati di acqua per il raffreddamento dei *data center*. Tanto che alcune grandi aziende tecnologiche stanno investendo nella costruzione di piccole centrali nucleari dedicate, proprio per soddisfare il fabbisogno energetico legato allo sviluppo dell'IA.

Infine, il combinarsi di alcuni elementi chiave – la necessità di dati, la potenza computazionale, le pratiche software avanzate, l'efficienza energetica – crea un contesto in cui emergono economie di scala e barriere all'ingresso molto elevate. I modelli più potenti si migliorano proprio grazie ai dati che raccolgono in fase di utilizzo, in un circolo virtuoso di retroazione (*positive feedback loop*) tra performance, dati e potenza computazionale. A loro volta, questi fattori richiedono grandi infrastrutture fisiche: data center, reti, impianti energetici. Tutto ciò contribuisce a concentrare il settore, dando luogo a mercati concentrati, dominati da pochi

[87] Recentemente si è molto discusso del possibile uso delle tecniche della distillazione e del *fine tuning* nel caso di nuovi servizi di IA ai danni degli operatori *incumbent*. Al riguardo, si vedano ad esempio: "OpenAI says it has evidence China's DeepSeek used its model to train competitor", *Financial Times*, 2025; "OpenAI Believes DeepSeek 'Distilled' Its Data For Training — Here's What To Know About The Technique", *Forbes*, 2025.

[88] Per una discussione su questi temi si vedano ad esempio: "AI power: Expanding data center capacity to meet growing demand", *Technology, Media & Telecommunications*, 2024; "How AI Is Fueling a Boom in Data Centers and Energy Demand", *Time Magazine,* 2024; "What the data centre and AI boom could mean for the energy sector".

attori globali capaci di sostenere i costi e la complessità dell'intero ecosistema dell'intelligenza artificiale.

Il capitolo successivo affronta tali profili, soffermandosi sulle implicazioni economiche dell'IA e la natura dei mercati che essa sta contribuendo a modellare.

4 Caratteristiche economiche dell'IA

Sulla base delle evidenze discusse nel Capitolo 2, che ha ricostruito l'evoluzione dell'intelligenza artificiale dagli anni Cinquanta fino ai modelli contemporanei, nonché degli aspetti tecnici e architetturali affrontati in dettaglio nel Capitolo 3, il presente capitolo si propone di analizzare le caratteristiche economiche dell'intelligenza artificiale. In particolare, verranno esaminate le caratteristiche dei mercati che ruotano attorno al suo sviluppo e utilizzo, e si inizieranno a delineare i primi effetti che tali trasformazioni stanno producendo sul sistema economico.

Il punto di partenza è, ovviamente, quello di stabilire che tipo di bene economico rappresenti l'intelligenza artificiale.

4.1 IA: bene pubblico o bene privato?

Dal punto di vista economico, l'IA non è un bene unitario e statico, ma una classe di servizi digitali ad alta intensità cognitiva, i cui confini merceologici sono in continua evoluzione. Originariamente sviluppata come *general purpose technology* (GPT) – ovvero una tecnologia pervasiva, capace di generare innovazioni trasversali in più settori – l'IA ha tratto la sua prima linfa da investimenti pubblici in ricerca di base, università, laboratori accademici e centri di ricerca non profit (§2.1). Questa fase, estesa nell'arco di almeno due decenni, è stata caratterizzata da una forte componente di bene pubblico, dove il ritorno sociale dell'innovazione era spesso più importante del ritorno privato.

Negli ultimi anni, tuttavia, l'IA ha subito un processo di progressiva "**commodificazione**": è diventata un bene privato, integrato nei modelli di business delle grandi piattaforme digitali. Queste imprese – come Google, Amazon, Microsoft o Meta – non solo utilizzano l'IA come leva competitiva per potenziare servizi già esistenti (motori di ricerca, cloud computing, pubblicità, e–commerce), ma la offrono anche come prodotto autonomo sul mercato, sia in forma diretta (API, chatbot, strumenti di generazione automatica) sia come *back–end* invisibile di servizi intelligenti.

Nel rapporto tra piattaforme e utenti, l'IA assume dunque le caratteristiche di un servizio digitale commerciale, venduto a un prezzo che – almeno per le versioni più avanzate – è

generalmente maggiore di zero. Il modello di pricing è sempre più segmentato: a fianco dell'accesso gratuito o promozionale (pensato per attrarre utenti e generare dati), si aggiungono formule di abbonamento differenziate in base alla disponibilità a pagare. Si osservano così pratiche di discriminazione di prezzo, ampiamente analizzate nell'economia dei beni informativi, in cui la stessa tecnologia viene confezionata in versioni (cd. *versioning*) più o meno potenti, con priorità d'accesso, maggiore personalizzazione o funzionalità aggiuntive, rivolte a target diversi: singoli utenti, professionisti, aziende, enti pubblici. Questa logica mira a massimizzare la cattura del surplus del consumatore da parte del fornitore di IA, adattando l'offerta alle caratteristiche eterogenee della domanda[89].

Dal punto di vista merceologico, l'IA è un bene dinamico: non ha confini funzionali fissi, ma evolve continuamente. Le versioni successive dei modelli (da GPT–3 a GPT–4, da DALL·E a Sora) integrano progressivamente nuove capacità – dalla traduzione automatica alla generazione di codice, dalla produzione di immagini e infografiche fino alla simulazione di processi complessi. Questo dinamismo riflette una logica di prodotto in aggiornamento permanente, in cui l'innovazione incrementale è parte integrante della proposta di valore. Ma questo continuo arricchimento funzionale ha anche un risvolto strategico importante: contribuisce a creare effetti di *lock–in* per gli utenti[90]. Le imprese e i professionisti che integrano servizi IA nei propri flussi di lavoro diventano progressivamente dipendenti dalla piattaforma prescelta, a causa di costi di switching, personalizzazioni accumulate e dipendenza dai dati storici. Questo meccanismo, ben noto nella letteratura sull'economia delle piattaforme, contribuisce a rafforzare la posizione di mercato degli attori già affermati, creando barriere all'entrata e, soprattutto, allo sviluppo di nuovi concorrenti.

Dal punto di vista geografico, l'IA è un servizio a vocazione intrinsecamente globale. I modelli sono addestrati su dati multilingue e distribuiti tramite infrastrutture cloud che superano le barriere nazionali. Tuttavia, la segmentazione dei mercati può comunque avvenire in base a criteri territoriali, culturali o normativi: la disponibilità di versioni localizzate, i regimi di

[89] Cfr. Shapiro, C., & Varian, H. R. (1999). *Information rules: A strategic guide to the network economy*. Harvard Business Press.

[90] Cfr. Farrell, Joseph, and Paul Klemperer (2007). Coordination and lock-in: Competition with switching costs and network effects. in *Handbook of Industrial Organization,* 3, 1967-2072.

regolazione differenziati (come il GDPR in Europa o l'AI Act), e la struttura della domanda nei diversi contesti nazionali influenzano l'adozione e il prezzo del servizio[91].

In sintesi, l'intelligenza artificiale si configura ad oggi come un bene privato complesso, nato come tecnologia pubblica, divenuto prodotto commerciale, e venduto su scala globale secondo logiche flessibili di accesso, segmentazione e aggiornamento continuo. Comprendere questa natura ibrida – tra innovazione, mercato e dipendenza strategica – è essenziale per analizzare i suoi impatti economici e per orientare un'azione pubblica consapevole.

4.2 Piattaforma economica

L'intelligenza artificiale non è soltanto una tecnologia avanzata: è il nucleo emergente di un'infrastruttura cognitiva che si inserisce pienamente nella traiettoria della *knowledge economy*. In quest'ultima, la conoscenza è l'input e l'output fondamentale dei processi economici, ma presenta caratteristiche peculiari che ne complicano la valorizzazione e la distribuzione. Come notava Kenneth Arrow già nel 1962[92], la conoscenza è un bene collettivo imperfetto: non rivale, difficilmente escludibile e soggetto a significative esternalità. Questo comporta problemi sia di appropriabilità (chi produce conoscenza fatica a internalizzarne i benefici), sia di incentivo (chi investe nella produzione di contenuti o idee ha ritorni incerti o nulli se non in assenza di meccanismi di remunerazione adeguati).

Nel contesto dell'intelligenza artificiale, queste tensioni vengono amplificate: le prestazioni dei modelli dipendono in misura decisiva dall'accesso a testi, immagini, suoni, cataloghi editoriali e archivi audiovisivi prodotti da soggetti terzi. Per questa ragione, i modelli di uso generale non possono essere letti come meri "autori artificiali", ma come infrastrutture computazionali che riorganizzano, trasformano e monetizzano valore informativo e culturale preesistente. Infatti, le piattaforme IA – come OpenAI, Google DeepMind, Anthropic o Meta – non sono piattaforme solo in senso tecnico (ossia ambienti modulari che ospitano applicazioni e dati, v. § 3), ma anche economico: agiscono come intermediari cognitivi tra chi produce informazione (testi, immagini,

[91] Si noti che la personalizzazione degli algoritmi dell'IA (e/o dei servizi connessi, ad esempio il *search*) è un ulteriore elemento che concorre a segmentare a livello nazionale i mercati, stante il fatto che utenti di nazioni diverse presentano lingue, culture, interessi, e contesti socio-politici distinti.
[92] Arrow, K. J. (1972). Economic welfare and the allocation of resources for invention (pp. 219-236). *Macmillan Education UK*.

video, codice, ecc.) e chi domanda contenuti o servizi intelligenti (v. Figura 5). Da questo punto di vista, si tratta di piattaforme a più versanti (*multi–sided platforms*), che mettono in relazione diretta gruppi distinti di utenti e in cui il valore creato su un lato dipende dalla partecipazione sull'altro. La teoria economica di questi mercati, formalizzata a partire dal famoso contributo di Rochet e Tirole[93], mostra come le piattaforme scelgano strategicamente se e quanto far pagare ciascun lato del mercato, spesso stabilendo prezzi espliciti da un lato e prezzi impliciti – o nulli – dall'altro, per massimizzare l'interazione e l'estrazione di valore. A differenza di altre piattaforme digitali, nel caso dell'IA generativa il lato dell'offerta non coincide soltanto con sviluppatori o inserzionisti, ma include sempre più spesso editori, archivi, piattaforme sociali e titolari di cataloghi, che forniscono dati o contenuti licenziati. Ne deriva una tendenza verso una piattaforma multiversante ibrida, in cui al modello freemium rivolto agli utenti si affiancano accordi bilaterali di licensing e accesso privilegiato a contenuti premium, utilizzati tanto per l'addestramento quanto per servizi di *answer engine* e *retrieval*.

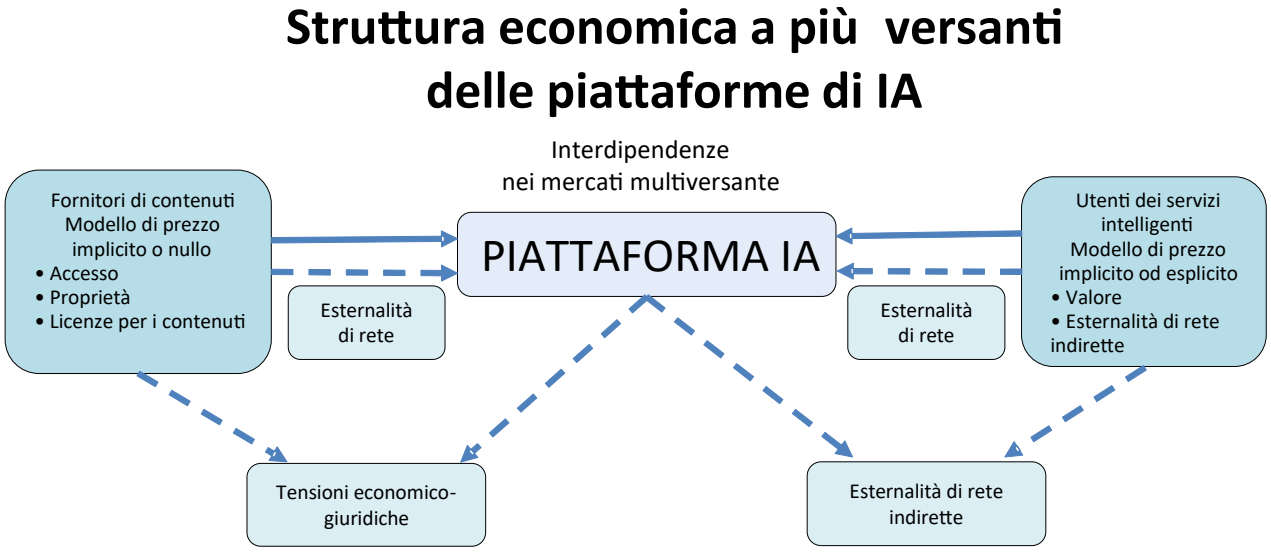


Figura 5 – IA come piattaforma multiversante

Questa struttura multi–versante è una caratteristica generale dei mercati digitali – dai motori di ricerca ai social network, fino agli app store – dove agiscono sistematicamente meccanismi

[93] Rochet, J. C., & Tirole, J. (2003). Platform competition in two-sided markets. *Journal of the European Economic Association*, 1(4), 990-1029.

di esternalità di rete, sia dirette (più utenti rendono la piattaforma più utile per ciascun utente) sia indirette (più produttori attraggono più consumatori e viceversa). L'intelligenza artificiale si inserisce in questo schema, ma con una peculiarità: i dati e le interazioni non sono semplici input, bensì alimentano direttamente i meccanismi di apprendimento del sistema. Ad esempio, ogni prompt inserito in ChatGPT, ogni immagine generata su Midjourney o ogni *feedback* fornito dagli utenti può contribuire a migliorare i modelli sottostanti attraverso tecniche di *reinforcement learning from human feedback*[94].

Si attivano così **circuiti di retroazione** (*feedback loop*) che rafforzano progressivamente il valore economico delle piattaforme di IA: più dati vengono assorbiti, maggiore è la capacità del sistema di apprendere e generare contenuti sofisticati. Ma anche il lato della domanda contribuisce: gli utenti non sono solo destinatari passivi, ma anche fornitori impliciti di processi cognitivi, attraverso le loro richieste, correzioni e preferenze. Ne deriva una fitta rete di relazioni economiche (e non solo), definite da prezzi espliciti o impliciti, a seconda dei casi.

Due relazioni sono particolarmente rilevanti. La prima è quella tra le piattaforme di IA e i **produttori originari di contenuti**: qui il prezzo è spesso nullo, ma l'uso sistematico di opere protette da copyright o di dati informativi estratti da fonti terze genera una tensione economico–giuridica.

Sul piano giurisprudenziale, il quadro statunitense[95], territorio di origine di molti dei più importanti operatori di IA, è ancora lontano dall'offrire una regola stabile e generale e molte

[94] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730-27744.

[95] Nel contesto europeo, il bilanciamento tra innovazione e appropriazione del valore non è affidato soltanto al contenzioso *ex post*, ma sempre più anche a strumenti *ex ante* che mirano a rendere compatibili lo sviluppo dei modelli e la tutela dei contenuti. In questa direzione, una parte delle pratiche rientra nel perimetro del *Text and Data Mining* (TDM) delineato dalla Direttiva DSM 2019/790, che definisce il TDM come qualsiasi tecnica automatizzata di analisi di testi e dati digitali volta a generare informazioni (inclusi "pattern, trend e correlazioni"). Il quadro legittima l'analisi estrattiva (pur riconoscendo che, nella pratica, la catena tecnica può comprendere atti che, sul piano del diritto d'autore, integrano riproduzioni) e la disciplina lungo tre binari: i) copie temporanee (art. 5, par. 1), ii) TDM per la ricerca (art. 3) e iii) TDM generale (art. 4), quest'ultimo condizionato alla mancata riserva dei diritti da parte dei titolari tramite *opt–out machine–readable*. È tuttavia essenziale distinguere la "memorizzazione" (fenomeno tecnico–empirico dei modelli) dalla "riproduzione" (categoria giuridica): l'eccezione TDM si arresta quando la prima si traduce in incorporazione stabile di porzioni protette e nella loro riemersione riconoscibile in output – come ha messo in luce GEMA v. OpenAI, caso in cui il Tribunale ha ritenuto che, per specifici testi di canzoni, non si fosse in presenza della sola estrazione di pattern tipica del TDM, ma di una incorporazione stabile dei testi nei parametri del modello , tale da consentire la loro riproduzione quasi testuale in output. Dall'altro lato, l'*AI Act* ha introdotto obblighi specifici per i fornitori di modelli di uso generale (fra cui la pubblicazione di una sintesi "sufficientemente dettagliata" dei dati di training, il rispetto di protocolli machine–readable e misure anti–aggiramento a

questioni rimangono aperte[96]: più che un orientamento unitario, stanno emergendo decisioni fortemente legate alle circostanze concrete dei singoli casi e, soprattutto, al modo in cui i giudici valutano il carattere trasformativo dell'uso e il possibile danno arrecato ai mercati delle opere originarie[97] (per un approfondimento giuridico sul rapporto tra training dei modelli, *text and data mining*, tutela autoriale e strumenti di compliance, cfr. Giuseppe Cassano, *Rapporto Comitato IA*, cap. 6). In questo quadro in continuo movimento, una recente sentenza della Corte distrettuale della California, resa nel procedimento promosso da tredici autori contro Meta, ha rigettato le accuse secondo cui il modello LLaMA sarebbe stato addestrato su libri protetti da copyright ottenuti illegalmente, in violazione del diritto d'autore. Il giudice ha osservato che i querelanti non sono riusciti a dimostrare l'effettiva riproduzione delle loro opere da parte del modello, né un danno concreto al mercato editoriale di riferimento. Il cuore del ragionamento adottato dal giudice ruota attorno alla dottrina del *fair use* e, in particolare, al concetto di "uso trasformativo": un criterio giuridico secondo cui un'opera protetta può essere riutilizzata senza autorizzazione qualora venga modificata in modo sostanziale, al punto da assumere una funzione nuova rispetto all'originale. Secondo il giudice, la trasformazione implicita nei processi di addestramento potrebbe in astratto giustificare l'uso delle opere, ma resta essenziale verificare caso per caso se da tale uso derivi un pregiudizio effettivo per gli autori.

Al di là delle singole pronunce, ciò che appare emergere è la centralità della concreta configurazione economica dell'uso contestato, la provenienza dei documenti, il grado di trasformazione e l'effetto sostitutivo o meno rispetto ai mercati presidiati dai titolari dei diritti.

---

tutela del diritto d'autore), obblighi accompagnati dal GPAI Code of Practice (luglio 2025), concepito per tradurre tali previsioni in prassi di compliance più verificabili.

[96] Cfr. Samuelson, P. (2023). Generative AI meets copyright. *Science*, *381*(6654), 158-161.

[97] In Thomson Reuters v. Ross Intelligence, ad esempio, la controversia riguardava l'uso delle "Westlaw headnotes" (massime/riassunti redazionali delle decisioni) per sviluppare uno strumento concorrente di ricerca giuridica; il giudice ha escluso il fair use, ritenendo che l'impiego delle note incidesse direttamente sul valore economico di quei contenuti (si veda: D'Angelo, F. D., & Shields, E.; *Thomson Reuters v. Ross Intelligence, Inc.* Loeb & Loeb LLP, 2025). Diversa, almeno in parte, è stata l'impostazione seguita in Bartz v. Anthropic; nel perimetro esaminato, l'uso dei libri per "training copies" è stato in parte ritenuto compatibile con il fair use. Il giudice ha tuttavia operato una distinzione decisiva: ha separato la valutazione del training dalla questione della "central library" di copie pirata, per la quale ha escluso una soluzione immediata a favore della società e ha rinviato a ulteriori accertamenti (anche sul piano dei danni). Secondo il giudice, il "training fair use" non equivale a legittimare l'acquisizione o la conservazione di copie da fonti illecite. Al riguardo, si consultino i documenti relativi al caso sul sito web Copyright Alliance. Inoltre, alcune cause sono tuttora pendenti, come quella promossa dal New York Times contro OpenAI e Microsoft (cfr. Axios, NYT case against OpenAI and Microsoft can advance, 1 aprile 2025) o quella avviata da Disney e Universal contro Midjourney (si veda: Disney and Universal sue AI image creator Midjourney, alleging copyright infringement, 11 giugno 2025) che ha portato il conflitto dall'ambito editoriale e testuale anche a quello audiovisivo e dei personaggi di finzione.

La cornice di *fair use* continua a definire *ex post* il perimetro della liceità, ma la traiettoria di mercato mostra che, *ex ante*, la via contrattuale sta diventando lo strumento principale con cui gli operatori approvvigionano dati affidabili e riducono il rischio di sostituzione economica dei mercati dei titolari.

Infatti, negli ultimi due anni si è consolidata una pratica di *licensing* che tratta i contenuti editoriali "premium" come input produttivi dei modelli di IA. In particolare, negli USA si registrano accordi tra editori e operatori tech – tra i più citati Associated Press-OpenAI (2023) e News Corp–OpenAI (2024) – affiancati in Europa da intese come Axel Springer-OpenAI, Financial Times-OpenAI e Le Monde/PRISA-OpenAI. Nel loro insieme alimentano un modello di IA "licenziata" volto a ridurre l'incertezza legale, garantire dati di qualità per addestramento e risposta e mitigare la disintermediazione del traffico verso le fonti originarie.

La seconda relazione economica è quella che lega le piattaforme di IA agli **utenti finali**. In questo caso, la strategia più comune è quella del modello cosiddetto *freemium*: l'accesso gratuito serve ad attrarre un'ampia base di utenti e generare esternalità di rete (dirette), mentre l'accesso *premium* consente la monetizzazione e il recupero dei costi infrastrutturali, che includono hardware ad alte prestazioni (GPU specializzate), energia, cloud e risorse umane qualificate.

Tuttavia, ridurre la monetizzazione al solo schema *freemium*, rischia oggi di essere riduttivo. Sempre più spesso, infatti, il modello economico delle piattaforme di IA combina abbonamenti, API enterprise, accesso prioritario all'infrastruttura e accordi di licensing sui dati e sui contenuti. In alcuni settori, soprattutto audiovisivo e musicale, si osserva persino una transizione da logiche di *scraping* generalizzato a modelli di addestramento su cataloghi proprietari o autorizzati, come mostrano la partnership Lionsgate-Runway[98] e i più recenti accordi di licensing nel settore musicale[99].

Questa trasformazione dell'IA in infrastruttura cognitiva e piattaforma economica non produce effetti soltanto sul versante della produzione e della circolazione dei contenuti, ma inizia a

[98] Si veda: "Hunger Games' studio Lionsgate announce AI video deal", da BBC.

[99] Si veda: "Major labels' licensing deals with AI companies: ECSA calls for transparent licensing agreements that truly value the works of composers and songwriters", dal sito web ECSA.

incidere anche sull'organizzazione del lavoro e sulla struttura delle **professioni**. Recenti evidenze empiriche sugli effetti dell'intelligenza artificiale nel mercato del lavoro mirano a valutare l'impatto dei modelli di IA sul lavoro, distinguendo tra la loro capacità teorica – ossia l'insieme delle attività che potrebbero svolgere o accelerare in astratto – e il loro uso effettivo, cioè i compiti nei quali tali sistemi sono concretamente impiegati nei contesti professionali e con quale intensità[100].

Tra i risultati più significativi emerge che i lavoratori maggiormente esposti tendono a essere, più frequentemente, donne, soggetti con più elevato livello di istruzione, lavoratori più anziani e meglio retribuiti. Al contrario, le occupazioni a prevalente contenuto manuale risultano, almeno per ora, meno esposte, finché l'evoluzione della robotica non renderà possibile un'incidenza più diretta anche su tali attività (si veda la Figura 6).

Ne deriva che, almeno in questa fase iniziale, il rischio non si concentra sui lavori meno qualificati o a più basso reddito, bensì su una parte rilevante del lavoro intellettuale e amministrativo. Tale risultato, che potrebbe apparire paradossale, deriva in realtà dalle caratteristiche specifiche del "prodotto intelligenza artificiale", che, come detto, si caratterizza per essere un potente sistema super-cognitivo e tende, in tal senso, a sostituire, almeno in parte, l'attività intellettuale dei lavoratori (e non solo).

[100] Si veda tra gli altri, Massenkoff, M., & McCrory, P., (2026). *Labor market impacts of AI: A new measure and early evidence*. Anthropic, e *AI's Labor Impact,* HAI The 2026 AI Index Report (2026). Stanford University.

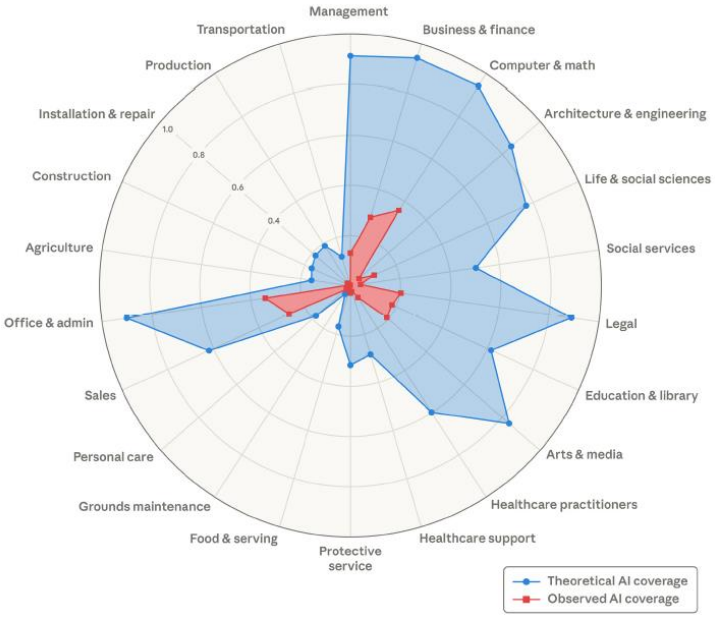


Figura 6 – Capacità teorica ed esposizione osservata dell'IA per categoria occupazionale[101]

Fonte: Labor market impacts of AI: A new measure and early evidence.

Altro risultato rilevante, e non immediatamente scontato, è poi la constatazione empirica, secondo cui non si osserva, allo stato, un'ondata di licenziamenti direttamente imputabile all'intelligenza artificiale. Emergono tuttavia segnali problematici sul versante delle nuove assunzioni: si osserva infatti un rallentamento dell'ingresso nelle occupazioni più esposte, soprattutto per i lavoratori tra i 22 e i 25 anni, in particolare nei profili junior riconducibili, ad esempio, ai programmatori, agli addetti al servizio clienti e agli operatori di data entry. In generale, i lavoratori appartenenti a tale fascia anagrafica occupati nelle professioni più esposte all'IA mostrano livelli occupazionali inferiori di circa il 16% rispetto ai lavoratori della medesima fascia d'età occupati nelle professioni meno esposte[102]. Il dato suggerisce, dunque, che non si registra con chiarezza un aumento della disoccupazione derivante da una sostituzione generalizzata e uniforme dell'occupazione, quanto piuttosto una possibile

[101] L'area blu rappresenta la quota di attività che i modelli linguistici di grandi dimensioni potrebbero teoricamente svolgere; l'area rossa indica invece la copertura effettivamente osservata nei dati d'uso professionale di Claude, aggregata per grandi categorie occupazionali.

[102] Si veda la figura 4.4.30 a pagina 222 dell'AI Index Report 2026 dell'Università di Stanford.

contrazione delle opportunità di accesso ad alcune professioni soprattutto per i lavoratori più giovani e in specifiche funzioni aziendali.

Questa scomposizione tra capacità potenziale dei modelli e adozione effettiva nei processi produttivi aiuta a comprendere perché gli effetti osservabili sull'occupazione aggregata risultino, almeno per ora, più lenti rispetto alla rapidità con cui evolvono le capacità tecniche dei sistemi. In molti casi, infatti, l'intelligenza artificiale non sostituisce integralmente un'occupazione, ma incide su porzioni specifiche del lavoro, modificandone contenuti, modalità di svolgimento e distribuzione interna delle mansioni. Ne deriva un processo di ristrutturazione graduale ma diffuso, con traiettorie differenziate tra settori ad alto contenuto cognitivo, attività amministrative, servizi professionali e comparti creativi.

Letta in questa chiave, l'intelligenza artificiale non può essere trattata semplicemente come uno strumento: è un nuovo intermediario della conoscenza che riorganizza relazioni economiche complesse tra produzione, appropriazione e diffusione dell'informazione, tutte mediate da meccanismi di prezzo o dalla loro assenza.

4.3 Struttura produttiva

La produzione di sistemi di intelligenza artificiale – in particolare quelli generativi su larga scala – si fonda su una struttura di costo fortemente sbilanciata verso componenti fisse e affondate (**sunk costs**), che rendono questa tecnologia profondamente **capital-intensive**. Tale caratteristica non è neutrale dal punto di vista concorrenziale, ma riflette e al tempo stesso rafforza una configurazione di potere altamente concentrata, in cui pochi operatori globali controllano le infrastrutture critiche del calcolo, dei dati e dell'accesso al mercato. Le principali voci di spesa includono:

- la raccolta, pulizia e annotazione (*labelling*) dei dati;
- l'addestramento computazionalmente intensivo di modelli su vasta scala;
- l'infrastruttura hardware (GPU e acceleratori specializzati come le TPU[103], nonché data center ad alta densità);

[103] Tensor Processing Unit (TPU): acceleratori ASIC specializzati per reti neurali.

- la ricerca e sviluppo di nuove architetture e ottimizzazioni software.

In questo contesto, il vero baricentro dei costi – e del potere industriale che ne deriva – si concentra nel *compute*, ossia nella disponibilità di potenza di calcolo avanzata necessaria all'addestramento e all'esecuzione dei modelli su larga scala. La centralità di tale fattore emerge con evidenza dalla rapida espansione della base infrastrutturale dell'IA: dal 2022, la capacità computazionale globale riconducibile ai principali chip per l'IA è cresciuta di circa 3,3 volte l'anno, fino a raggiungere nel 2025 l'equivalente di 17,1 milioni di H100, vale a dire di chip ad alte prestazioni assunti come parametro di riferimento della potenza di calcolo per l'IA[104]; parallelamente, la capacità di potenza dei data center dedicati all'IA ha toccato circa 29,6 GW nel quarto trimestre del 2025[105] (per un'analisi delle questioni relative all'impatto ambientale dell'IA, § 5.4). Questi dati confermano che la competizione sull'IA si gioca ormai su investimenti infrastrutturali di scala eccezionale, rispetto ai quali il *compute* rappresenta oggi la principale componente dei costi fissi e, al tempo stesso, il principale collo di bottiglia tecnologico dell'intelligenza artificiale. La sua scarsità relativa, l'elevata intensità di capitale richiesta e i tempi di realizzazione delle infrastrutture rendono infatti estremamente difficile l'ingresso di nuovi operatori (per i profili ambientali associati a training e inferenza e alla crescita dei data center, si veda § 5.4).

Il controllo della potenza di calcolo si dispiega su **due livelli strettamente interdipendenti**:

- da un lato, la progettazione e la produzione dei **chip avanzati**;
- dall'altro, la disponibilità e la gestione delle **infrastrutture fisiche** in cui tali chip vengono installati e utilizzati, in particolare i data center iperscalabili (hyperscale)[106].

Il **primo livello** – quello della produzione dei chip – non dipende solo dalla capacità industriale, ma è condizionato anche da un vincolo materiale: la disponibilità di materie prime critiche, comprese le terre rare, necessarie alla realizzazione dell'hardware dei data center (componentistica elettronica, magneti, sistemi di raffreddamento e infrastrutture di rete), che

[104] Unità di calcolo paragonabili al chip Nvidia H100, oggi assunto come benchmark di riferimento per misurare la potenza computazionale impiegata nei sistemi di IA.

[105] The 2026 AI Index Report (2026).

[106] Cellini, P., Ibarra, M., (2024), *AI Impact*, Luiss University Press.

hanno catene di approvvigionamento altamente concentrate, rendendole vulnerabili a interruzioni commerciali o geopolitiche: la Cina, ad esempio, controlla circa il 98–99% della fornitura di gallio raffinato.

Blocco del data center	Dove entrano nel data center	Materie prime critiche / materiali associati	Perché conta (driver legato all'IA)
Schede server e circuiti	PCB, connettori, saldature, cablaggi interni	Argento, oro, rame, stagno, tantalio, palladio, nichel	Più server e maggiore densità elettronica → aumenta la quantità di componentistica e di interconnessioni.
Semiconduttori e microchip	CPU/GPU/acceleratori, logiche, componenti microelettronici	Silicio; gallio, germanio, indio e arsenico per semiconduttori composti e fotonica; fluoro/composti fluorurati soprattutto come materiali di processo	Crescita del *compute* → maggiore domanda di chip avanzati e di materiali di base e di processo per la microelettronica.
Dissipazione termica e struttura	Dissipatori, scambiatori, chassis, parti di supporto	Rame, alluminio	Più potenza = più calore da smaltire → aumenta il fabbisogno di materiali per raffreddamento e strutture meccaniche.
Magneti e archiviazione dati	Motori/attuatori (ventole, pompe), componenti HDD e parti correlate	Terre rare (neodimio, praseodimio, disprosio, terbio), boro, rame, alluminio	Necessità di gestire stress termico e affidabilità → magneti efficienti per parti mobili; lo storage "di massa" su HDD mantiene driver significativi di materiali.
Rete e connettività	Fibra ottica, cavi di comunicazione, interconnessioni intra/inter data center	Germanio e, in alcune tratte e apparecchiature ottiche, erbio; oltre a rame e alluminio per cablaggi e alimentazione	L'IA aumenta traffico e interconnessione → la connettività ad alta velocità diventa un vincolo strutturale.

Tabella 5 – Materie prime critiche per la realizzazione dei data center

Secondo l'IEA, la rapida crescita di IA e data center richiede quantità significative di minerali e metalli – tra cui rame, alluminio, silicio, gallio e terre rare – e l'espansione di capacità al 2030 può incidere in modo misurabile sulla domanda complessiva, con pressioni particolarmente rilevanti su alcuni materiali[107]. In tale contesto, il rischio di colli di bottiglia lungo la filiera dei materiali diventa parte integrante della struttura produttiva dell'IA. La Tabella 5 sintetizza come queste materie prime entrano nei principali blocchi funzionali di un data center e perché risultano strategiche alla luce della diffusione dell'IA.

Il **secondo livello** riguarda invece il possesso e la gestione dei data center "iperscalabili": è qui che la posizione di mercato di alcuni fornitori di hardware e infrastrutture cloud si traduce in un controllo indiretto sull'intera catena del valore dell'IA, dall'addestramento dei modelli fino

[107] International Energy Agency, (2025), Energy and AI.

alla loro distribuzione sul mercato. Si pensi che circa la metà dei data center a livello mondiale è oggi gestita da infrastrutture cloud riconducibili a pochi grandi operatori[108]. Anche la distribuzione geografica di tali infrastrutture mostra una forte concentrazione: secondo l'AI Index 2026, nel 2025 gli Stati Uniti rappresentano il 57,1% dei data center considerati nella figura, a fronte del 5,6% della Germania e del 5,5% del Regno Unito; per l'Italia, la quota si attesta invece intorno all'1,8%[109]. Il mercato dei servizi cloud è dominato da Amazon (AWS), Microsoft (Azure) e Google Cloud, che insieme controllano oltre il 60% del mercato globale[110]. Tali operatori non si limitano a fornire capacità di calcolo a terzi, ma costituiscono il principale punto di accesso alle risorse computazionali necessarie allo sviluppo di modelli avanzati, al quale sono legati anche soggetti formalmente distinti, come OpenAI o Anthropic, le cui capacità di ricerca e sviluppo dipendono strutturalmente dalle infrastrutture messe a disposizione dai grandi operatori tecnologici[111] (si veda anche § 4.4). Se il controllo cinese sulle terre rare incide a monte sulla disponibilità materiale dell'hardware, il controllo esercitato dagli Stati Uniti sulla filiera dei semiconduttori avanzati opera a un livello tecnologico e sistemico, incidendo sulle condizioni stesse di accesso al *compute*. Tale posizione si fonda sul primato nel design dei chip e sulla centralità delle imprese statunitensi nei segmenti a maggiore valore aggiunto. A valle del *compute*, questo assetto si riflette in un elevato grado di concentrazione industriale: NVIDIA detiene una quota stimata attorno all'80% del *design* dei chip per data center[112], mentre la filiera produttiva dei semiconduttori avanzati risulta a sua volta estremamente concentrata: TSMC produce circa il 90% dei chip più avanzati a livello globale[113] e ASML è l'unico fornitore mondiale

[108] A tale riguardo, si vedano: Cellini (2024) a pag. 96.

[109] Si veda l'AI Index Report 2026 dell'Università di Stanford, dove sono indicati i seguenti valori assoluti in termini di distribuzione di data center: Stati Uniti 5.427; Germania 529; Regno Unito 523; Cina 449; Canada 337; Francia 322; Australia 314; Paesi Bassi 298; Russia 251; Giappone 222; Brasile 197; Messico 173; Italia 168; India 153; Polonia 144..

[110] Nello specifico, la quota di AWS oscilla tra il 40-62%, Azure tra il 10-35% e Google Cloud tra il 5-10%. A tale riguardo si consulti: Organisation for Economic Co-operation and Development - OECD (2025), *Competition in artificial intelligence infrastructure*, OECD Roundtables on Competition Policy Papers, No. 330, OECD Publishing, Paris.

[111] Aresu, A. (2024). *Geopolitica dell'intelligenza artificiale*. Feltrinelli Editore.

[112] Organisation for Economic Co-operation and Development - OECD (2025), *Competition in artificial intelligence infrastructure*, OECD Roundtables on Competition Policy Papers, No. 330, OECD Publishing, Paris, https://doi.org/10.1787/623d1874-en. Nello specifico si veda il seguente passaggio: "*Nvidia (a fabless company) has emerged as the market leader in the sector, with recent estimates suggesting that the firm has over 80% market share for GPU chips used for AI*".

[113] Organisation for Economic Co-operation and Development - OECD (2025), Nello specifico si veda il seguente passaggio: "*No other firm has been able to commercialise an alternative to ASML's Extreme Ultraviolet (EUV) Lithography technology that is necessary to manufacture the latest generations of chips.*"

delle macchine EUV necessarie alla loro fabbricazione[114]. La Figura 7 rende immediatamente leggibile la concentrazione (anche geografica) della filiera dei semiconduttori e delle memorie, che contribuisce a spiegare l'interdipendenza verticale e la struttura dei costi fortemente asimmetrica richiamate nel seguito.

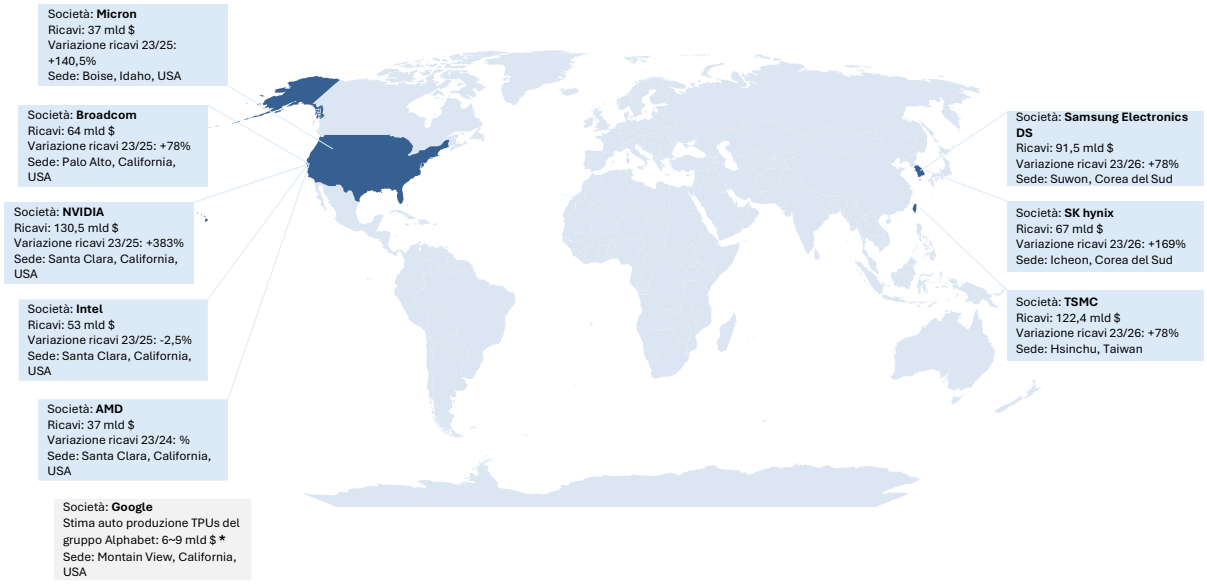


Figura 7 - Principali imprese dei semiconduttori per l'IA

** Per Google il dato indicato non rappresenta ricavi, ma una stima esterna della spesa per TPU nel 2024, non ricavabile direttamente dal bilancio di Alphabet. Si veda al riguardo: "Google preparing to partner with Taiwan's MediaTek on next AI chip, Information reports", da Reuters.*

La combinazione di concentrazione nella progettazione dei chip, nella loro produzione e nella gestione dei data center determina un aumento del potere di mercato degli operatori verticalmente integrati, con una struttura dei costi fortemente asimmetrica tra tali soggetti e altre società, riducendo in tal modo la contendibilità del settore. Il controllo delle risorse critiche del *compute*, congiuntamente con la posizione detenuta nei vari stadi della filiera, consente infatti agli operatori integrati di sostenere – e in parte internalizzare – costi fissi estremamente elevati, che risultano invece proibitivi per potenziali nuovi entranti.

[114] Cellini, (2024), pag. 95.

A rendere economicamente sostenibile questo livello eccezionale di investimento contribuisce, sempre più spesso, una specifica architettura finanziaria fondata su accordi di natura circolare tra i principali attori dell'ecosistema dell'intelligenza artificiale.

Box 4 – Il potere del calcolo: CPU, GPU, TPU

Negli ultimi anni le GPU (Graphics Processing Unit) sono passate dall'essere componenti per la grafica e i videogiochi a infrastrutture computazionali centrali per l'intelligenza artificiale, in particolare per il *deep learning*, che richiede l'esecuzione di un numero enorme di operazioni matematiche in parallelo su grandi quantità di dati. È in questo contesto che le GPU si sono rivelate decisive.

Le CPU (Central Processing Unit)[115] restano fondamentali per compiti generali e di coordinamento, ma non sono ottimizzate per il calcolo massivamente parallelo richiesto dall'addestramento dei modelli. Le GPU[116], al contrario, sacrificano una parte della flessibilità per ottenere un parallelismo estremo, grazie a migliaia di *core* semplici che operano simultaneamente, risultando particolarmente adatte all'addestramento e all'esecuzione delle reti neurali.

Su questa base si spiega l'ascesa di NVIDIA, che occupa oggi una posizione quasi egemonica nel mercato aperto delle infrastrutture di calcolo per l'IA. Il suo vantaggio non dipende solo dall'hardware, ma soprattutto dall'ecosistema software (CUDA), che ha creato una forte dipendenza tecnologica. NVIDIA non è più soltanto un produttore di chip, ma un fornitore di infrastrutture cognitive, che fa della disponibilità di GPU un fattore strategico globale.

Diversa è la strategia di Google, che ha sviluppato chip proprietari, le cd. TPU (Tensor Processing Unit)[117], altamente specializzate per il *machine learning*. Meno versatili delle GPU, ma più efficienti per carichi specifici, le TPU rafforzano un modello di

[115] La CPU è l'unità di elaborazione centrale di un computer. È progettata per la massima versatilità: il suo compito è coordinare tutte le attività del sistema, gestire il sistema operativo ed eseguire istruzioni logiche complesse in modo sequenziale. La sua architettura si basa su pochi *core* molto potenti, capaci di passare rapidamente da un compito all'altro (dal calcolo di un foglio elettronico alla gestione della rete). Proprio a causa della sua natura sequenziale, la CPU risulta poco efficiente quando deve processare contemporaneamente milioni di piccoli calcoli identici, come quelli richiesti dalle moderne reti neurali.

[116] La GPU nasce originariamente per il *rendering* grafico (videogiochi e video), ma è diventata il motore fondamentale dell'IA grazie alla sua architettura parallela. A differenza della CPU, una GPU contiene migliaia di *core* più piccoli e specializzati, progettati per eseguire contemporaneamente lo stesso tipo di operazione matematica su enormi blocchi di dati. Questa capacità di calcolo parallelo la rende perfetta per l'intelligenza artificiale, dove è necessario moltiplicare miliardi di matrici di dati nello stesso istante.

[117] La TPU è un acceleratore IA progettato da Google specificamente per il machine/deep learning. A differenza della GPU, la TPU è un ASIC (*Application-Specific Integrated Circuit*), ovvero un chip costruito per fare una sola cosa: gestire i "tensori" (le strutture dati alla base dell'apprendimento). Le TPU ottimizzano il flusso di dati per le operazioni di algebra lineare, riducendo i tempi di addestramento e i consumi energetici. Tuttavia, sono meno flessibili perché eccellono quasi esclusivamente in ambiti legati a framework specifici e nei grandi data center.

integrazione verticale, in cui il controllo dell'hardware sostiene l'intero ecosistema di servizi e modelli di IA dell'azienda.

Accanto a questi due poli, AMD rappresenta il principale concorrente di NVIDIA sul piano hardware, ma sconta un ritardo nell'ecosistema software, mentre Intel, pur centrale nei data center, non è il riferimento per l'addestramento dei modelli più avanzati. Apple, infine, segue una traiettoria distinta, puntando su acceleratori integrati nei propri chip per sviluppare modelli di IA direttamente sul dispositivo, piuttosto che sull'IA industriale.

Nel complesso, emerge una nuova geografia del potere computazionale: GPU e chip specializzati non sono più semplici componenti tecnici, ma infrastrutture strategiche con implicazioni economiche, geopolitiche e regolatorie. Comprendere le differenze tra CPU, GPU e TPU significa oggi comprendere chi controlla la capacità di produrre intelligenza artificiale e a quali condizioni.

In tale configurazione, gli elevati costi infrastrutturali non sono sostenuti da singoli operatori in modo isolato, ma vengono distribuiti lungo una rete di relazioni incrociate tra sviluppatori di modelli, fornitori di chip e operatori cloud, attraverso meccanismi che combinano investimenti azionari, contratti pluriennali di fornitura e impegni di spesa anticipata per servizi di calcolo. Questi "accordi circolari" consentono di trasformare una parte rilevante dei costi fissi in flussi finanziari interni al sistema, riducendo il rischio individuale e rafforzando, al contempo, le interdipendenze tra i soggetti coinvolti.

Dal punto di vista economico, tali accordi svolgono una funzione cruciale di sostegno alla domanda di potenza di calcolo, permettendo agli sviluppatori di modelli di accedere a capacità computazionali altrimenti insostenibili e ai fornitori di infrastrutture e hardware di garantirsi volumi di vendita sufficienti a giustificare investimenti di scala eccezionale. Ne risulta un circuito autoalimentato in cui capitale, infrastrutture e ricavi si rafforzano reciprocamente, contribuendo alla rapida espansione dell'ecosistema, ma accentuando al contempo le barriere all'ingresso e la dipendenza sistemica tra pochi grandi operatori.

In questo quadro, alcuni accordi tra sviluppatori di modelli, fornitori di infrastruttura cloud e produttori di chip – sintetizzati a titolo esemplificativo nella Tabella 6 – mostrano come gli impegni finanziari si collochino stabilmente nell'ordine delle decine o centinaia di miliardi di dollari su orizzonti pluriennali, confermando la natura altamente *capital–intensive* del settore.

Un riscontro empirico utile a qualificare questa architettura contrattuale è offerto da uno studio sulle partnership tra grandi fornitori di cloud e sviluppatori di modelli generativi recentemente svolto dalla Federal Trade Commission statunitense[118]. Il rapporto evidenzia che, nelle principali partnership esaminate, ricorrono – sebbene in misura non identica – clausole che combinano: (i) partecipazioni azionarie e/o meccanismi di *revenue sharing* a beneficio del partner cloud; (ii) diritti di consultazione, influenza e, in alcuni casi, di controllo, nonché previsioni di esclusiva, preferenza, anche rispetto alle tempistiche di rilascio dei modelli; (iii) impegni di spesa cloud che richiedono allo sviluppatore di destinare una quota significativa dell'investimento ricevuto ai servizi del partner cloud (c.d. *circular spending*); (iv) condivisione di risorse e informazioni, quali accesso scontato alla capacità di calcolo, piani di co-sviluppo di chip, scambio di personale tecnico e accesso a dati finanziari, di *performance* e sui fabbisogni infrastrutturali. La FTC indica inoltre, come "aree da monitorare", i possibili effetti di tali clausole sull'accesso agli input essenziali (compute e conoscenza/talento), sull'aumento dei costi di *switching*, sia contrattuali sia tecnici, e sui vantaggi informativi derivanti dal flusso di dati sensibili tra i partner. Nel complesso, il report supporta l'idea che queste partnership non si esauriscano in un mero investimento finanziario o in un ordinario rapporto cliente-fornitore, ma possano realizzare forme di integrazione contrattuale idonee a rafforzare *lock-in*, asimmetrie informative e interdipendenze lungo tutta la filiera dell'IA.

La Tabella 6 fa luce sulle interdipendenze che emergono in forma di accordi commerciali e societari tra operatori privati statunitensi. Non include invece i principali attori cinesi, il cui sviluppo è più spesso sostenuto anche da fondi pubblici e da strumenti di politica industriale (come mostra l'istituzione del fondo semiconduttori da 344 miliardi di yuan[119]), secondo logiche spesso non direttamente comparabili con quelle di mercato.

[118] Federal Trade Commission, FTC. (2025). *Partnerships between cloud service providers and AI developers*. FTC staff report on AI partnerships & investments.

[119] Si veda: "*La Cina punta ancora sui microchip: istituito un nuovo fondo da 47,5 miliardi di dollari",* da Forbes.

Attori coinvolti	Descrizione dell'accordo	Valore dichiarato / stimato
OpenAI ↔ Microsoft	Microsoft ha investito in OpenAI (equity + cloud credits); OpenAI utilizza Azure come infrastruttura primaria per training e deployment dei modelli.	Microsoft possiede il 27% di OpenAI[120]. Gli impegni pluriennali di spesa cloud sono stimati tra i 200 e i 250 mld USD (contratti di lungo periodo)[121].
OpenAI ↔ Nvidia	Nvidia fornisce GPU avanzate; rapporti commerciali e cooperazione tecnologica. Non risultano investimenti diretti pubblici di Nvidia in OpenAI, ma una forte dipendenza commerciale incrociata.	30 mld USD[122] (tale valore è stato rimodulato. Il precedente accordo prevedeva un investimento di 100 mld USD).
OpenAI ↔ Oracle	Oracle costruisce e gestisce data center dedicati per carichi IA; OpenAI si impegna all'utilizzo della capacità.	Contratti pluriennali stimati nell'ordine di 300 mld USD, con avvio previsto nel 2027[123].
OpenAI ↔ Amazon (AWS)	Accordi di utilizzo infrastrutturale per specifici carichi di lavoro; relazione non esclusiva e complementare ad Azure.	Amazon ha annunciato un investimento in OpenAI fino a 50 mld USD + impegno compute 2 GW[124]
OpenAI ↔ Google	Accordo infrastrutturale e cloud tra Google e OpenAI, che ha aggiunto Google Cloud tra i propri fornitori di capacità computazionale per sostenere training e inferenza dei modelli, in una collaborazione definita nel 2025 nonostante la concorrenza diretta tra le due società[125].	Non risulta, allo stato, un investimento azionario di Google in OpenAI analogo a quello effettuato da Google in Anthropic.
OpenAI ↔ AMD	OpenAI acquisterà 6 gigawatt da AMD per i nuovi data center. Accordo sul diritto di acquisto di azioni di AMD da parte di Open AI.	Valore non pubblico; accordo tecnologico (per 6 gigawatt): accordo sul diritto di acquisto del 10% di AMD da parte di Open AI[126].
Anthropic ↔ Amazon (AWS)	AWS è il principale partner cloud, nonché di training di Anthropic, che utilizza (anche) chip Trainium e Inferentia, distribuendo i modelli Claude tramite Amazon Bedrock.	4 mld USD di investimento complessivo[127]; ulteriori impegni cloud/compute di lungo periodo non integralmente pubblici.
Anthropic ↔ Google	Google fornisce supporto cloud e infrastrutturale ad Anthropic, che utilizza Google Cloud e TPU per training e servizi.	Nel 2023 sono stati resi pubblici 2 mld USD di investimento da parte di Google in Anthropic[128]; di cui oggi detiene circa 14%[129]. A fine 2025, Anthropic ha annunciato l'ampliamento dell'uso di Google Cloud e fino a 1 milione di TPU Google; l'operazione è descritta come di valore pari a

[120] Si veda: "*OpenAi diventa società a scopo di lucro. Microsoft rileva il 27% e supera i 4mila miliardi*", da La Stampa.

[121] Si veda: "*The next chapter of the* Microsoft–*OpenAI partnership*", da Microsoft Corporate Blogs.

[122] Si veda: "*OpenAI's $110 billion funding round draws investment from Amazon, Nvidia, SoftBank*", da Reuters.

[123] Si veda: "*Accordo OpenAI-Oracle, 300 miliardi di investimenti in potenza di calcolo in 5 anni*", dal Sole24Ore.

[124] Si veda: "*OpenAI's $110* billion *funding round draws investment from Amazon, Nvidia, SoftBank*", da Reuters.

[125] Si veda: "Exclusive: OpenAI taps Google in unprecedented cloud deal despite AI rivalry, sources say", da Reuters.

[126] Si veda: "*OpenAi sceglie Amd, mega ordine di chip e 10% delle azioni: come cambia la sfida sull'intelligenza artificiale*", dal Corriere della Sera.

[127] Si veda: "Amazon concludes $4 billion investment in Anthropic", Amazon News.

[128] Si veda: "*Google agrees to invest up to $2 billion in OpenAI rival Anthropic*", da Reuters.

[129] Si veda: "*Google has given Anthropic more funding than previously known, show new filings*", Da TechCrunch.

Attori coinvolti	Descrizione dell'accordo	Valore dichiarato / stimato
		decine di miliardi di dollari e con oltre 1 gigawatt di capacità attesa nel 2026[130].
Anthropic ↔ Microsoft	Partnership infrastrutturale su Azure, con integrazione di Claude nell'ecosistema Microsoft.	Microsoft ha annunciato un investimento in Anthropic fino a 5 mld USD[131].
Anthropic ↔ Nvidia	Partnership strategica per infrastruttura e scaling dei modelli Claude su sistemi Nvidia; è stato annunciato anche un investimento di Nvidia in Anthropic.	Nvidia ha annunciato un investimento in Anthropic fino a 10 mld USD[132].
Nvidia ↔ Amazon (AWS)	Accordo industriale su chip e networking AI: Nvidia fornirà ad AWS fino a 1 milione di chip GPU entro il 2027, oltre a componenti di rete e altre tecnologie per l'infrastruttura AI[133].	I termini economici dell'accordo non sono stati resi pubblici.
Nvidia ↔ xAI	Accordi di fornitura GPU (investimento equity condizionato all'acquisto di hardware).	Possibile investimento equity condizionato all'acquisto di hardware. Il valore stimato dell'operazione si attesterebbe intorno ai 2 mld USD (stima)[134].
Nvidia ↔ CoreWeave	Nvidia è azionista rilevante e principale fornitore di GPU; CoreWeave utilizza quasi esclusivamente hardware Nvidia.	6,3 mld USD in fornitura compute + 2 mld USD di investimento in azioni CoreWeave da parte di Nvidia[135].
OpenAI ↔ CoreWeave	Contratti di fornitura di potenza di calcolo GPU; CoreWeave ha assegnato equity a OpenAI.	Totale contratti pari a 22,6 mld USD[136].
CoreWeave ↔ Meta	Accordo infrastrutturale per capacità AI: CoreWeave ha sottoscritto con Meta un contratto per fornire potenza di calcolo/cloud AI fino al 2032[137].	Il valore degli accordi è pari a 14,2 mld USD (accordo settembre 2025) + 21 mld USD (accordo ampliativo aprile 2026), per un totale cumulato di circa 35,2 mld USD.

Tabella 6 – Principali accordi tra operatori dell'IA: natura degli impegni e rete di interdipendenze

[130] Si veda: "*Anthropic to use Google's AI chips worth tens of billions to train Claude chatbot*", da Reuters.

[131] Si veda: "*Microsoft, Nvidia to invest in Anthropic as Claude maker commits $30 billion to Azure*", con riferimento al passaggio: "*Nvidia (NVDA.O), opens new tab will commit up to $10 billion to Anthropic and Microsoft (MSFT.O), opens new tab up to $5 billion*", da Reuters.

[132] Si veda: "*Microsoft, Nvidia to invest in Anthropic as Claude maker commits $30 billion to Azure*", con riferimento al passaggio: "*Nvidia (NVDA.O), opens new tab will commit up to $10 billion to Anthropic and Microsoft (MSFT.O), opens new tab up to $5 billion*", da Reuters.

[133] Si veda: "*Nvidia to sell 1 million chips to Amazon by end of 2027 in cloud deal*", da Reuters.

[134] Si veda: "*Musk's xAI nears $20 billion capital raise tied to Nvidia chips, Bloomberg News reports*", da Reuters.

[135] Si veda: "*CoreWeave, Nvidia sign $6.3 billion cloud computing capacity order*", da Reuters.

[136] Si veda: "*CoreWeave inks $6.5 billion deal with OpenAI*", da CNBC.

[137] Si veda: "*CoreWeave signs $14 billion AI infrastructure deal with Meta*" e "*Meta, CoreWeave deepen AI cloud partnership with fresh $21 billion deal*", da Reuters.

Questa architettura finanziaria rappresenta quindi il complemento finanziario della concentrazione tecnologica descritta sopra: se la progettazione dei chip, la loro produzione e la gestione dei data center iperscalabili costituiscono l'ossatura industriale del *compute*, gli accordi circolari ne costituiscono il meccanismo di finanziamento implicito, consentendo di sostenere una dinamica di investimento infrastrutturale senza precedenti per volume, intensità e rapidità.

È in questo quadro che si collocano i livelli di costo osservati nella fase di addestramento dei modelli di grandi dimensioni. A titolo esemplificativo, l'addestramento del modello GPT–3 (175 miliardi di parametri) ha comportato un costo stimato di circa 4,6 milioni di dollari solo per la parte computazionale, mentre per modelli più recenti – come GPT-4 – le stime superano i 60 milioni di dollari, fino a oltre 100 milioni per alcuni *frontier models* di nuova generazione[138]. Il consumo energetico associato a queste operazioni è anch'esso molto elevato[139] (si veda § 5.4). Secondo le stime elaborate nel corso degli anni[140], i costi di addestramento (*training*) dei modelli di intelligenza artificiale generativa hanno mostrato un trend in costante crescita. Nel 2017, l'addestramento del modello *transformer* originale – che costituisce la base della maggior parte degli attuali LLM (si veda § 3.2.2) – è costato circa 670 dollari. Nel 2019, il modello RoBERTa Large ha richiesto un investimento di circa 160.000 dollari, mentre nel 2023 il costo stimato per l'addestramento di GPT-4 di OpenAI ha raggiunto circa 79 milioni di dollari. Ancora più elevati sono i costi stimati per Llama nel 2024, pari a circa 170 milioni di dollari (Figura 8). Oggi i modelli più avanzati tendono a divulgare sempre meno informazioni su parametri, dataset e durata del training, rendendo più difficile stimarne con precisione i costi effettivi e, più in generale, valutare in modo trasparente le condizioni di sviluppo dei sistemi più potenti.

Se si prendono in considerazione tutti gli oneri associati alla creazione e allo sviluppo dei modelli di IA (hardware per training, energia, cloud, R&S, staff), allora è stato calcolato che il

[138] Cottier, B, (2023), Trends in the Dollar Training Cost of Machine Learning Systems.

[139] L'addestramento di GPT-3 ha richiesto l'equivalente dell'elettricità consumata annualmente da più di 1.000 famiglie, e un solo giorno di utilizzo di ChatGPT può equivalere al fabbisogno quotidiano di 33.000 famiglie statunitensi.

[140] The 2025 AI Index Report (2025). Stanford University, Human Centered Artificial Intelligence – HAI.

costo complessivo dei modelli di ultima generazione supererà il miliardo di dollari entro il prossimo anno[141].

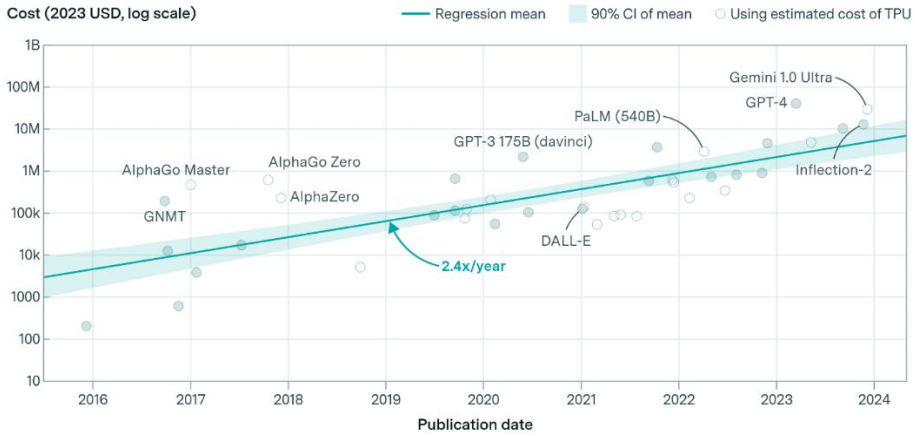


Figura 8 – Costo di addestramento (hardware ed energia) dei modelli di IA generativa

Fonte: Epoch AI

Questi costi fissi generano significativi rendimenti di scala (e barriere all'ingresso): una volta sviluppato un modello, il costo marginale di servirlo a un numero crescente di utenti è relativamente basso, il che incentiva la concentrazione dell'offerta.

Inoltre, esiste una correlazione empiricamente osservabile tra la quantità di dati utilizzati, la potenza computazionale, il progresso algoritmico e le performance predittive dei modelli, che rafforza i vantaggi cumulativi degli operatori incumbent (Figura 9)[142].

A tali dinamiche si aggiunge il contributo degli utenti finali che, come illustrato in § 3.1, contribuiscono indirettamente al miglioramento dei modelli, grazie ai meccanismi di apprendimento iterativo (feedback, correzioni, dati di utilizzo).

Il complesso di questi fattori dà luogo a **rendimenti di scala** sia dal lato dell'offerta (scala tecnologica e apprendimento algoritmico: economie di scala) sia dal lato della domanda (più utenti, maggiore valore per ciascun utente: effetti di rete).

[141] Cottier, B., Rahman, R., Fattorini, L., Maslej, N., Besiroglu, T., & Owen, D. (2024). The rising costs of training frontier AI models. *arXiv preprint arXiv:2405.21015*.

[142] Ho, A., Besiroglu, T., Erdil, E., Owen, D., Rahman, R., Guo, Z. C., ... & Sevilla, J. (2024). Algorithmic progress in language models. *Advances in Neural Information Processing Systems*, *37*, 58245-58283.

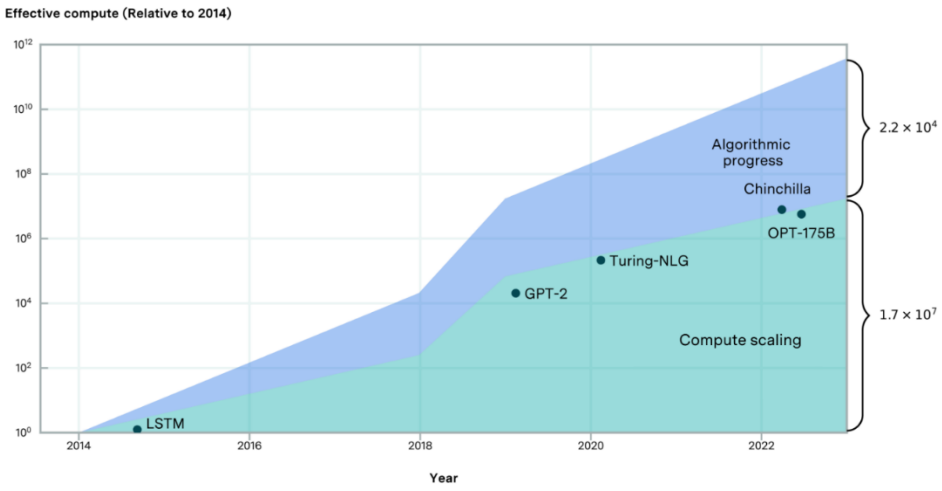


Figura 9 – Stime dei contributi di scalabilità computazionale e innovazione algoritmica per raggiungere prestazione stato dell'arte (il contributo del progresso algoritmico è circa la metà della scala computazionale)

Fonte: Epoch AI

Tuttavia, negli ultimi anni si è assistito a una parziale attenuazione di queste barriere, grazie all'evoluzione di nuove tecniche. La tendenza verso modelli più compatti ed efficienti è ormai consolidata: DeepMind ha dimostrato che addestrare modelli più piccoli su dataset più ampi può produrre risultati almeno analoghi rispetto a modelli più grandi addestrati su dati meno curati, confermando che l'efficienza algoritmica e la qualità dei dati possono, almeno in parte, sostituire la mera scala computazionale. Parallelamente, si sono diffuse soluzioni tecniche e operative che consentono di comprimere i costi medi e di ridurre la scala minima efficiente richiesta per competere, ampliando – almeno a valle – gli spazi di ingresso nel mercato.

In questo quadro, alla dinamica infrastrutturale del *compute* si sovrappone una competizione tra "blocchi tecnologici", intesi come ecosistemi industriali e standard di fatto (modelli, piattaforme cloud, stack software, canali di distribuzione). In una prima fase, i grandi modelli linguistici sono stati sviluppati soprattutto secondo un'impostazione proprietaria e "chiusa", in cui le capacità venivano protette come segreto industriale e rese disponibili principalmente via *cloud*: ciò ha consentito – e in parte consente tuttora – agli operatori pionieri di monetizzare il vantaggio temporale, nonché di fidelizzare la domanda, trattenendo l'utente all'interno del proprio ecosistema proprietario (c.d. effetto di *lock–in*).

In una fase successiva, si è invece evidenziata una strategia, soprattutto da parte di alcuni operatori con quote di mercato ancora limitate, basata su modelli *open–weight* e su forme di apertura controllata, orientata a ridurre il vantaggio accumulato dai sistemi chiusi attraverso una diffusione più rapida e capillare e una maggiore accessibilità della tecnologia, pur senza rimuovere il vincolo strutturale a monte rappresentato dal *compute*. I modelli *open–weight*, infatti, sfruttano l'apertura per accelerare la standardizzazione, ampliare la platea degli sviluppatori e favorire nuove opportunità di integrazione con i servizi utilizzati dagli utenti finali. Si tratta di una dinamica ricorrente nei mercati digitali: lo standard "aperto" (o semi–aperto) può fungere da leva competitiva per scalfire ecosistemi chiusi preesistenti (si pensi ai casi Android vs iOS, Linux vs Windows o ROCm vs CUDA), con l'effetto di ridisegnare gli equilibri concorrenziali tra chi controlla la piattaforma e chi controlla la diffusione e l'impiego della tecnologia su larga scala.

In questo scenario, l'AI Index Report 2026 dell'Università di Stanford mostra oggi che la competizione tra modelli non si gioca più soltanto sulla performance, ma sempre più anche sulle modalità di accesso e sui gradi di trasparenza. In altri termini, il vantaggio competitivo non dipende solo da quanto sia efficiente un modello, ma anche da chi ne controlla l'accesso, l'infrastruttura di erogazione e le informazioni necessarie per riprodurlo.

Occorre precisare che per "accessibilità" dei modelli non si intende la semplice possibilità, per l'utente comune, di utilizzare un sistema di IA generativa tramite interfaccia, bensì la possibilità, per sviluppatori e operatori professionali, di accedere al modello come infrastruttura tecnologica da integrare in servizi, applicazioni e processi propri. In tale ottica, l'accesso via API, il rilascio dei pesi o la mancata pubblicazione del codice di *training* descrivono diversi gradi di disponibilità tecnica del modello e diversi livelli di dipendenza dal fornitore.

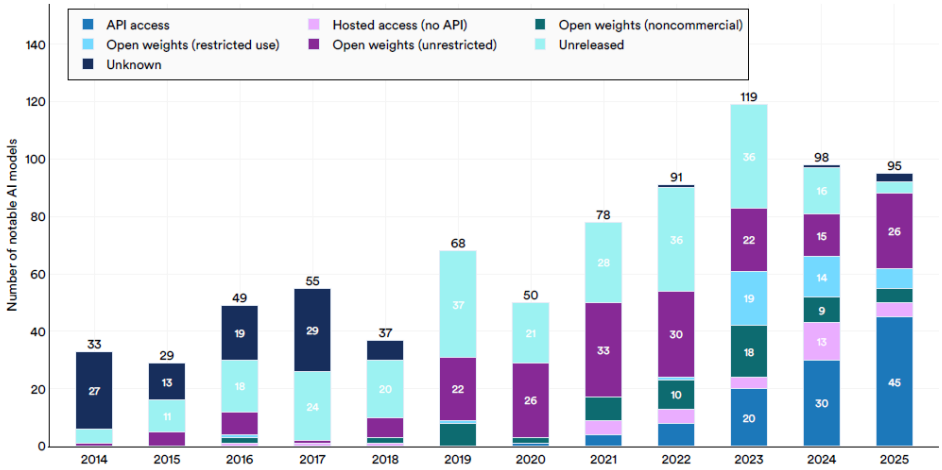


Figura 10 - Evoluzione delle modalità di rilascio dei modelli di IA notevoli (2014–2025)

Fonte: *AI Index Report 2026*

Il fatto che nel 2025 l'API access sia stata la modalità di rilascio più frequente (45 modelli su 95; si veda la Figura 11) segnala che molti modelli vengono resi disponibili non come beni tecnici pienamente trasferibili, ma come servizi intermediati dal fornitore: sviluppatori, imprese e altri utilizzatori professionali possono servirsene per applicazioni e servizi propri, ma entro condizioni definite unilateralmente dal gestore della piattaforma, che controlla prezzi, limiti d'uso, filtri, *logging* e integrazione con il *cloud*.

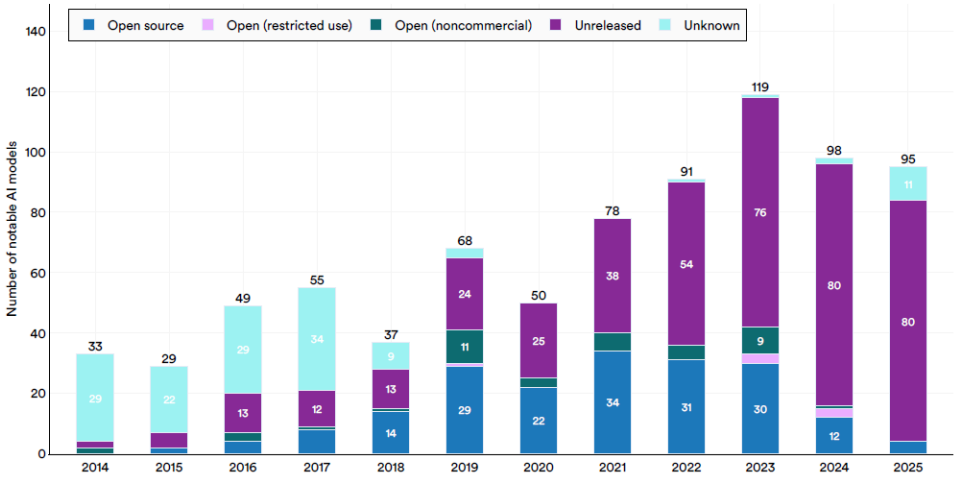


Figura 11 – Evoluzione dell'accessibilità del training code nei modelli di IA

Fonte: AI Index Report 2026

A ciò si aggiunge un secondo dato, ancora più rilevante sotto il profilo strutturale: 80 modelli su 95 non hanno reso disponibile il codice di *training* (Figura 12). Questo significa che, anche quando un modello è accessibile o parzialmente aperto, resta spesso preclusa la possibilità di ricostruirne integralmente il processo di sviluppo, di verificarne i metodi di addestramento, di replicarne i risultati o di sottoporlo a un audit indipendente. Ne deriva che l'"apertura" dei modelli *open-weight* rappresenta sì una leva concorrenziale importante, ma opera entro un contesto generale in cui i sistemi di frontiera tendono a essere sempre più opachi e sempre più controllati dai grandi operatori che presidiano insieme modello, cloud, distribuzione e interfaccia utente.

Box 5 – Costi medi IA e nuove tecniche

Negli ultimi anni sono state utilizzate dagli operatori alcune tecniche volte a migliorare le performance dei modelli e a ridurre i costi medi di training degli algoritmi e fornitura del servizio. Le principali sono:

- **Distillazione del modello (*model distillation*):** Tecnica che trasferisce la conoscenza da un grande modello (*teacher*) a uno più piccolo (*student*), riducendo drasticamente il carico computazionale e mantenendo prestazioni comparabili, con un significativo abbattimento dei costi di inference oltre che di training.
  Esempio: DeepSeek–V3 offriva prestazioni vicine a GPT–3.5 con una frazione della potenza richiesta.
- **Fine-tuning**: Consente a operatori minori di specializzare modelli pre–addestrati su domini specifici (sanità, legale, servizio clienti), riducendo i costi dell'intero ciclo di training.
  Esempio: un *fine-tuning* su Mistral per usi legali può costare meno di 100.000 USD, contro diversi milioni per un addestramento da zero.
- **Modelli open source e open-weight**: Progetti come LLaMA, Mistral, DeepSeek o gli sviluppi di EleutherAI e Hugging Face hanno aperto l'accesso alla tecnologia, riducendo la dipendenza degli incumbent. Ciò consente anche a laboratori pubblici o start–up private di sviluppare servizi IA con risorse contenute, favorendo una maggiore contendibilità a valle, pur in presenza di una persistente concentrazione a monte.

Queste tecnologie abbassano significativamente i **costi medi totali** e **la scala minima per operare**, aprendo spazi di concorrenza e innovazione anche per attori con budget meno elevati. Tuttavia, il vantaggio competitivo in termini di

infrastruttura, dati proprietari e distribuzione resta largamente concentrato nelle mani di pochi operatori globali.

Non è infatti un caso che molti operatori – OpenAI/Microsoft prima e Anthropic/Microsoft successivamente, Google/DeepMind, Amazon, Alibaba, Meta – operano in regime di integrazione verticale, controllando simultaneamente molti stadi della filiera dell'intelligenza artificiale: dati e informazioni, modelli, infrastrutture hardware, accesso al mercato e interfacce utente.

In sintesi, la struttura produttiva dell'IA riflette un'economia *capital intensive*, alimentata da *feedback* interni e *network* esterni, in cui l'efficienza tecnologica e la sostenibilità economica dipendono dalla capacità di abbattere i costi medi, soprattutto attraverso lo sfruttamento di vantaggi di scala.

4.4 Servizi, operatori e mercati

Abbiamo visto che il mercato dell'IA è caratterizzato dalla presenza di elevate barriere tecniche (innovazioni algoritmiche), economiche (elevati costi fissi e affondati) e strategiche (integrazione verticale e diagonale). In particolare, il segmento dei modelli generativi di larga scala si configura come **un settore concentrato** e strategicamente integrato[143]. Una manciata di grandi operatori – prevalentemente statunitensi – domina lo sviluppo, la distribuzione e l'integrazione dell'IA nei sistemi digitali, beneficiando di economie di scala, rendimenti crescenti e vantaggi cumulativi difficili da replicare. Questo porta a una struttura industriale fortemente oligopolistica, dove l'innovazione tecnologica è strettamente intrecciata con il controllo delle piattaforme digitali globali.

[143] Le più recenti ricerche di mercato non solo evidenziano un'impennata negli investimenti in IA generativa — basti pensare che, nel 2024, questo segmento ha rappresentato circa un quarto del mercato globale dell'IA, superando i 184 miliardi di dollari USA, con un tasso di crescita annuo composto del 24,4% previsto tra il 2023 e il 2030 — ma mostrano anche un cambiamento significativo nelle abitudini dei consumatori: nel 2023, ben 13 milioni di adulti statunitensi hanno scelto l'intelligenza artificiale generativa come principale strumento per effettuare ricerche online, una cifra destinata a superare i 90 milioni entro il 2027.

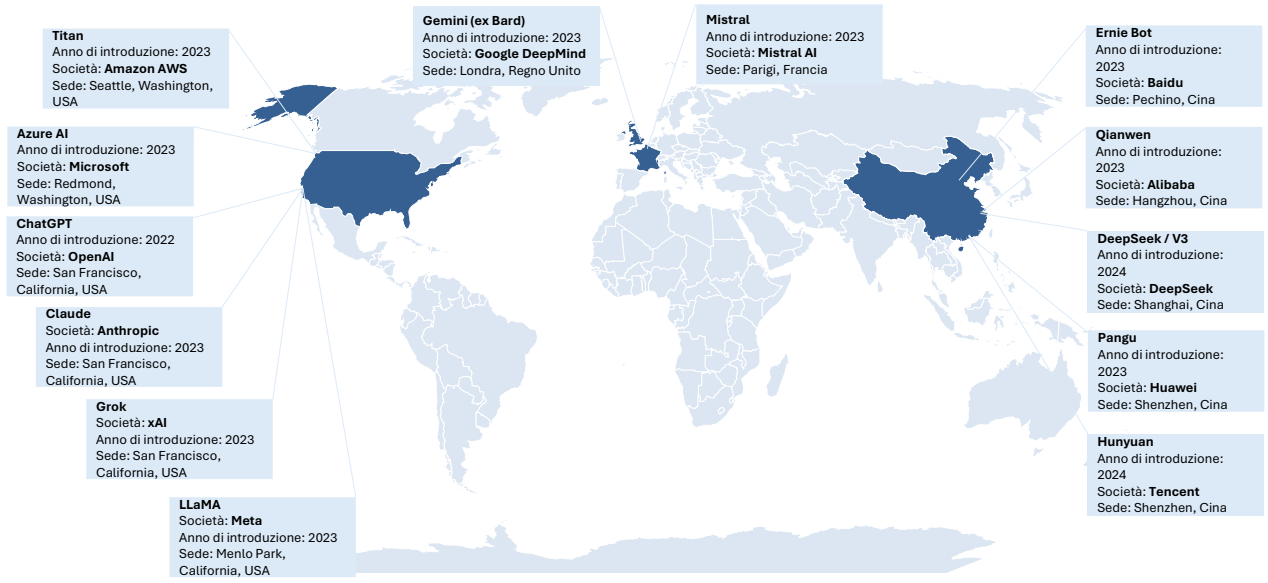


Figura 12 – Servizi di IA generativa nel mondo

I principali attori del settore – OpenAI[144], Google DeepMind, Amazon AWS, Meta, xAI – non sono semplici produttori di tecnologie, ma architetti di interi ecosistemi digitali basati sull'IA. Ciascuno ha sviluppato un modello proprietario, un insieme di modelli, o stretto accordi commerciali o finanziari (per esplorare la rete di interdipendenze tra operatori, si veda la Tabella 6):

- OpenAI è profondamente intrecciata con il mondo Microsoft (quest'ultima ha investito decine di miliardi di dollari in OpenAI, fornendo anche il proprio servizio di cloud Azure, e in cambio detiene una quota di minoranza nella società; il rapporto, tuttavia, non è più caratterizzato dalla originaria esclusiva infrastrutturale, e Microsoft ha integrato in Microsoft 365 Copilot anche modelli Anthropic);
- Google DeepMind, con Gemini, punta sull'integrazione con Workspace e Android; inoltre, Google/Alphabet ha investito alcuni miliardi di dollari in Anthropic, fornendo anche servizi di cloud e detenendo una partecipazione di circa il 14% della società di IA;
- Amazon AWS propone il modello Titan e, più di recente, anche la famiglia Nova, all'interno della propria offerta cloud per imprese tramite Bedrock; parallelamente, ha

[144] OpenAI è stata inizialmente legata a Microsoft da una partnership cloud esclusiva, poi progressivamente attenuata dall'evoluzione dell'accordo verso assetti più multi-cloud e dall'integrazione, da parte di Microsoft, anche di modelli Anthropic in Microsoft 365 Copilot.

rafforzato l'alleanza con Anthropic, di cui è divenuta il principale partner cloud e di addestramento;

- Meta offre la famiglia LLaMA, che segue una strategia open–weight, favorendo l'adozione da parte della comunità open source; più recentemente ha però lanciato Muse Spark, primo modello della nuova serie Muse e primo modello del team Meta Superintelligence Labs, destinato ad alimentare Meta AI e a essere progressivamente integrato nei servizi e nei prodotti offerti dalla capogruppo;
- xAI, fondata da Elon Musk, integra il modello Grok nella piattaforma X (ex Twitter);
- Mistral, unica eccezione europea significativa, propone modelli ad alta efficienza e pienamente accessibili.

In parallelo, la Cina ha costruito un proprio ecosistema nazionale, alimentato da colossi come Alibaba (Qianwen), Baidu (Ernie Bot), Tencent (Hunyuan) e Huawei (Pangu), a cui si aggiungono modelli come DeepSeek, che si contraddistinguono per l'efficienza computazionale e l'approccio open-weight. Questi attori sono parte di un sistema industriale orientato all'autosufficienza tecnologica e sostenuto da politiche pubbliche attive.

In tale contesto, anche i modelli cinesi hanno accresciuto sensibilmente il proprio peso competitivo: sebbene la Cina continui a collocarsi dietro gli Stati Uniti per numero di *notable AI models* rilasciati nel 2025 (30 contro 50), il divario nelle prestazioni dei modelli migliori si è ormai ristretto fino a valori molto contenuti (Figura 13).

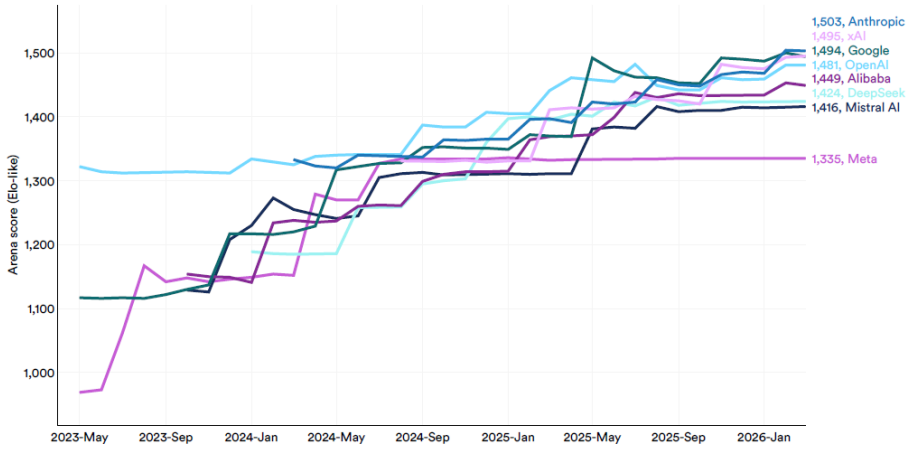


Figura 13 – Differenziali di performance di IA: USA vs. Cina

Fonte: IA Index Report 2026

Già nel febbraio 2025 DeepSeek-R1 aveva temporaneamente eguagliato, e per breve tempo superato, il miglior modello statunitense all'interno della classifica Arena, che misura le performance dei modelli[145]; nel marzo 2026, il modello statunitense meglio posizionato manteneva un vantaggio di soli 39 punti Arena rispetto al miglior modello cinese, pari a uno scarto del 2,7% nel punteggio[146]. Il dato segnala dunque non l'azzeramento di ogni differenza tecnologica tra i due Paesi, ma l'emersione di una competizione molto più ravvicinata al vertice[147]. La riduzione del gap tra i principali modelli di IA statunitensi e cinesi segnala l'emersione di una competizione strategica sempre più intensa, considerata la rilevanza del settore per l'intero sistema economico, industriale e sociale di una nazione.

Una delle caratteristiche più rilevanti del settore dell'intelligenza artificiale su larga scala è il grado di integrazione raggiunto dai principali operatori (v. Figura 14).

Le stesse imprese che sviluppano modelli IA avanzati sono spesso anche:

i) Gestori di piattaforme di accesso al mercato (v. § 4.2)

Le stesse aziende che sviluppano IA controllano anche le piattaforme che intermediano l'accesso degli utenti finali ai modelli. Ad esempio:

- Google distribuisce il modello Gemini attraverso il motore di ricerca, l'assistente Android e la suite Google Workspace (Gmail, Docs, Meet). L'IA è integrata anche nei dispositivi mobili (Pixel) e accessibile via Bard/Gemini AI.
- Microsoft integra GPT (della famiglia OpenAI) e i modelli di Anthropic nella suite Microsoft 365 tramite Copilot (Word, Excel, Teams) e nei servizi cloud Azure AI. Dispone così di una filiera verticale: modelli, interfaccia e infrastruttura.
- Amazon propone i suoi modelli Titan attraverso AWS Bedrock, rivolgendosi a imprese e sviluppatori. L'IA è inoltre integrata in Alexa e nei sistemi enterprise di AWS.

[145] Si veda: Arena Leaderboard Dataset.

[146] Si veda la figura 2.1.3 a pagina 77 dell'AI Index report dell'università di Stanford.

[147] L'AI Index Report 2026 riporta che nel febbraio 2025 DeepSeek-R1 aveva temporaneamente eguagliato, e per breve tempo superato, il miglior modello statunitense nella classifica Arena; nel marzo 2026, il modello statunitense meglio posizionato manteneva un vantaggio di soli 39 punti Arena, pari al 2,7%, rispetto al miglior modello cinese.

- xAI (di Elon Musk) ha integrato il modello Grok direttamente nella piattaforma X (ex Twitter), posizionando l'IA come assistente personalizzato all'interno del social network.
- Meta, inizialmente focalizzata sulla distribuzione tecnica di LLaMA (tramite Hugging Face, AWS, Azure), ha ora lanciato un proprio assistente IA conversazionale, denominato Meta AI, disponibile direttamente in: WhatsApp, Messenger e Instagram (via chat).
  Basato su versioni ottimizzate di LLaMA 2 e LLaMA 3, questo assistente consente di porre domande, generare testi, ottenere suggerimenti e creare immagini ("*Imagine with Meta AI*"), sfruttando le vaste basi di utenti di Meta.

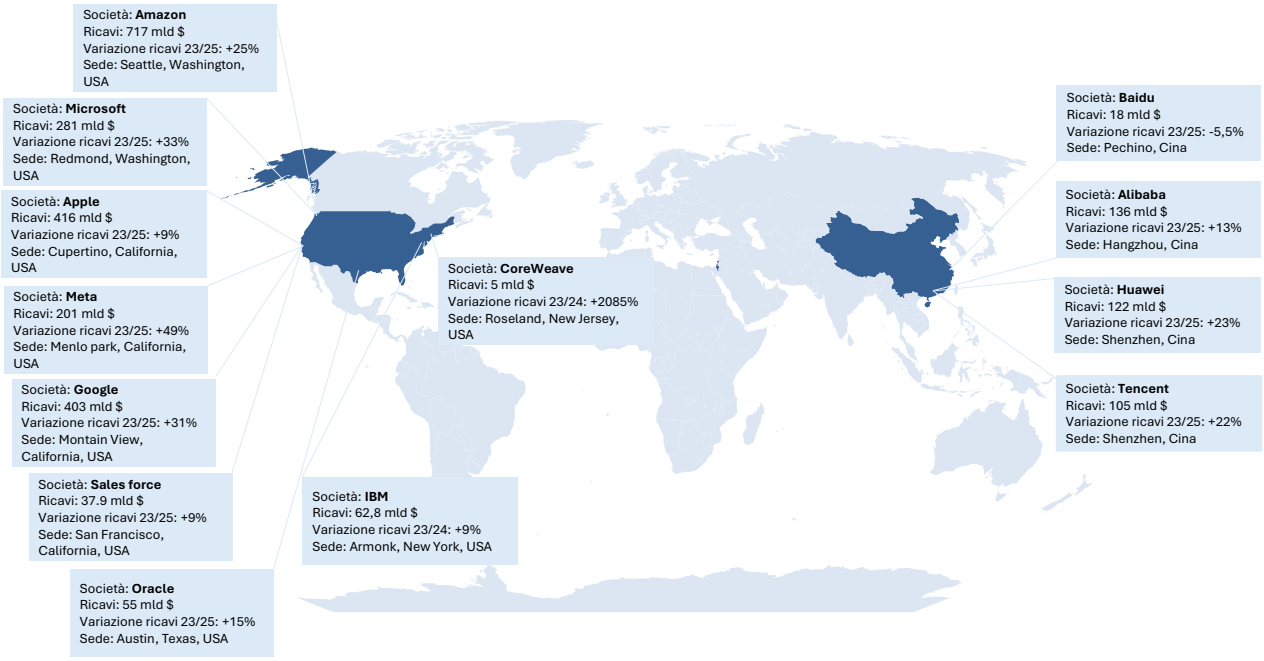


Figura 14 – Principali operatori nel campo dell'IA

ii) Fornitori di infrastruttura computazionale (integrazione verticale)

Le big tech dispongono spesso di data center proprietari e accesso diretto a componenti hardware essenziali (si veda anche il Box 4). Ad esempio:

- Microsoft gestisce l'infrastruttura Azure, che ospita i modelli di OpenAI e quelli di Anthropic[148].
- Amazon possiede AWS, il principale fornitore mondiale di servizi cloud, dove gira il modello Titan.
- Google utilizza la propria rete di TPU (*Tensor Processing Units*), progettate internamente per l'addestramento di modelli come Gemini.
- Meta ha costruito data center ottimizzati per supportare l'addestramento e il deployment dei modelli LLaMA.

iii) Distributori di servizi digitali finali (integrazione diagonale)

L'IA viene incorporata in applicazioni e servizi a valore aggiunto, che alimentano flussi di entrate diretti e indiretti. Ad esempio:

- Meta, Amazon, Google, X, Microsoft hanno tutte adottato una strategia di integrazione della propria IA nell'ambito del pacchetto dei servizi offerti (con particolare riferimento a Google);
- OpenAI offre ChatGPT come servizio in abbonamento, integrato in Copilot e in altre interfacce;
- Anthropic distribuisce Claude tramite API e partnership con sviluppatori terzi e aziende, tra cui Google in particolare.

Questa integrazione su più livelli consente vantaggi significativi in termini di efficienza, controllo dei costi, accesso ai dati e rapidità di diffusione, ma accentua anche la dipendenza dell'ecosistema da pochi attori dominanti. Le barriere all'accesso non si limitano alla tecnologia del modello, ma comprendono l'intero sistema di distribuzione, monetizzazione e feedback che accompagna l'uso dell'intelligenza artificiale su scala.

4.5 Considerazioni conclusive

L'intelligenza artificiale è oggi ampiamente riconosciuta come una tecnologia generale (*general purpose technology,* GPT), ovvero una tecnologia con effetti pervasivi sulla produzione,

[148] Si veda al riguardo: "Microsoft, Nvidia Pump Billions Into Anthropic"da Bloomber.

sull'organizzazione economica e sociale, e sulla generazione di innovazioni a valle in una pluralità di settori. Come l'elettricità, il computer o Internet nel passato, l'IA ha la capacità non solo di aumentare la produttività nei settori esistenti, ma anche di trasformare radicalmente le modalità con cui si lavora, si consuma e si prendono decisioni, sia in ambito privato che pubblico.

Le tecnologie generali, come spiegano Helpman e Trajtenberg[149], sono "motori della crescita" non tanto per le loro prestazioni intrinseche, ma per il loro potenziale di **complementarità** con altri fattori (organizzazione, capitale umano, regolazione) e per la loro capacità di generare ondate successive di innovazione. L'IA si colloca pienamente in questa logica: le sue applicazioni – dalla generazione di linguaggio naturale all'analisi predittiva, dal controllo robotico all'ottimizzazione logistica – si estendono orizzontalmente a tutti i settori e verticalmente in tutti i livelli della catena del valore.

Secondo il McKinsey Global Institute, l'adozione diffusa dell'intelligenza artificiale generativa potrebbe generare un impatto economico annuale tra 2,6 e 4,4 trilioni di dollari a livello globale, una cifra comparabile all'intero PIL del Regno Unito o della Germania[150]. L'Organizzazione per la Cooperazione e lo Sviluppo Economico (OCSE) ha rilevato che oltre il 70% delle imprese nei paesi avanzati considera l'IA come tecnologia prioritaria per i prossimi cinque anni[151]. Il World Economic Forum ha sottolineato il potenziale trasformativo dell'IA in sanità, energia, educazione, pubblica amministrazione, evidenziando anche la necessità di governance multilivello per gestire i rischi sistemici[152].

Sul piano sociale, l'IA si configura come un'infrastruttura super-cognitiva: media il modo in cui le informazioni vengono prodotte, aggregate e valutate; condiziona la percezione della realtà; automatizza processi decisionali a diversi livelli. Ciò apre opportunità significative in termini di accesso alla conoscenza e personalizzazione dei servizi, ma comporta anche nuovi rischi

[149] Helpman, E., & Trajtenberg, M. (1998). *A time to sow and a time to reap: Growth based on general purpose technologies*. In E. Helpman (Ed.), *General Purpose Technologies and Economic Growth*. MIT Press.

[150] McKinsey Global Institute (2023). *The economic potential of generative AI: The next productivity frontier.*

[151] Organisation for Economic Co-operation and Development – OECD (2023). *Artificial Intelligence Outlook 2023: Enabling Trust and Innovation*; Organisation for Economic Co-operation and Development – OECD (2024), *Digital Economy Outlook 2024 (Volume 1): Embracing the Technology Frontier*.

[152] World Economic Forum (2025), *AI in Action: Beyond Experimentation to Transform Industry*.

distributivi, relativi alla polarizzazione delle competenze, alla disintermediazione professionale e alla concentrazione del potere informativo.

In questo senso, l'intelligenza artificiale non è semplicemente una nuova ondata tecnologica: è un moltiplicatore cognitivo e organizzativo con implicazioni sistemiche per l'economia, per la società e per la democrazia.

Tuttavia, se il potenziale dell'IA è sistemico, la sua proprietà e il suo controllo sono oggi fortemente concentrati. Come è stato illustrato in questo Capitolo, negli ultimi anni si è assistito a una vera e propria "privatizzazione" dell'infrastruttura cognitiva globale, in cui le tecnologie chiave (modelli linguistici, dataset, infrastrutture di calcolo, interfacce utente) sono diventate appannaggio di un ristretto gruppo di operatori privati, globali e fortemente integrati.

Questa concentrazione non è casuale, ma riflette le caratteristiche economiche strutturali del mercato dell'intelligenza artificiale analizzate in questo capitolo: natura del bene IA (§ 4.1); configurazione del mercato in più versanti legati da esternalità di rete intra–gruppo (*within–group*) e inter–gruppo (*across–group*) (§ 4.2); e una struttura produttiva con elevati costi fissi e affondati e quindi rendimenti crescenti di scala (§ 4.3). L'intero settore è pertanto segnato da barriere all'ingresso estremamente elevate, dovute a:

- una scala minima efficiente molto alta;
- rendimenti di scala anche dal lato della domanda con effetti di rete diretti (più utenti migliorano il servizio) e indiretti (più dati migliorano i modelli);
- integrazione verticale (dallo sviluppo del modello all'infrastruttura, all'interfaccia utente) e diagonale (incorporazione dell'IA in suite di servizi già dominanti, quali Microsoft 365 o Google Workspace);
- strategie di *platform envelopment*, cioè l'estensione dell'IA come funzione aggiuntiva di piattaforme preesistenti (ad esempio social networks, messaggerie, ecc.), rafforzando posizioni di significativo potere di mercato.

Il panorama globale dell'IA si sta frammentando in tre blocchi strategici (§ 4.2). Gli USA mantengono la leadership tecnologica attraverso modelli proprietari ad alta intensità di capitale ("*Closed Strategy*"). La Cina, pur continuando a valorizzare l'efficienza computazionale e la diffusione dei modelli aperti o *open-weights*", presenta oggi un quadro meno lineare di

quanto suggerisca una contrapposizione netta con il modello statunitense. I più recenti sviluppi di Alibaba indicano infatti una torsione almeno parziale verso soluzioni proprietarie (con modelli *closed-source* disponibili via API)[153], interrompendo la precedente traiettoria open-source della famiglia Qwen, in una più ampia strategia di monetizzazione. Ne consegue che anche il modello cinese va ormai letto come un assetto ibrido, nel quale convivono apertura selettiva, modelli open-weight e rafforzamento di offerte chiuse ad alta integrazione cloud. A ciò si aggiunge che, sul piano competitivo, il divario tra Stati Uniti e Cina nella performance dei modelli di frontiera risulta ormai sostanzialmente quasi colmato[154] (§§ 4.3 e 4.4). L'Europa appare in una posizione di transizione: leader nella governance etica e normativa, ma ancora in cerca di una direzione chiara ai fini dello sviluppo di un ecosistema industriale autoctono capace di competere su scala globale. In questa prospettiva, una strategia plausibile per l'Unione europea potrebbe consistere non tanto nell'inseguire frontalmente i modelli di frontiera *cloud-based* sviluppati dagli operatori statunitensi o cinesi, quanto nel valorizzare un ecosistema di modelli di taglia più contenuta, ad alta efficienza e distribuibili in ambiente locale, sia *on-device* sia *on-premises*. Come esposto nel capitolo 3 del presente rapporto (§ 3.3), una simile traiettoria presenterebbe, per il contesto europeo, alcuni vantaggi comparativi rilevanti: maggiore protezione dei dati, più elevato controllo sui flussi informativi, minore esposizione di dati sensibili a trasferimenti verso infrastrutture esterne[155]; minore dipendenza da infrastrutture hardware anche in virtù dell'uso delle tecniche di quantizzazione che riducono le necessità computazionali. In tale cornice, la leva europea potrebbe essere meno quella della scala assoluta e più quella di un diverso ecosistema basato su verificabilità, sicurezza e privacy.

È quindi lecito interrogarsi su quale possa essere il ruolo dell'Europa, e dell'Italia in particolare, e, più in generale, dei sistemi economici che non hanno sviluppato autonomamente una capacità industriale e tecnologica comparabile a quella americana o cinese. Il caso della Cina è istruttivo: grazie a una combinazione di investimenti pubblici strategici, sostegno alla scalata di grandi

[153] Si veda: China's AI Companies Are Going Closed Source.

[154] I due ecosistemi si sono alternati più volte al vertice della performance dei modelli dal 2025 e, a marzo 2026, il vantaggio del miglior modello statunitense su quello cinese era pari ad appena il 2,7%. Per un maggior approfondimento del fenomeno del tema, si consulti il § 4.3 e 4.4.

[155] L'esecuzione locale dei modelli consente infatti di mantenere i flussi informativi all'interno del perimetro istituzionale o aziendale, riducendo i rischi di memorizzazioni non controllate, addestramenti incrociati indesiderati e fuoriuscita di conoscenza sensibile.

operatori nazionali (Baidu, Tencent, Alibaba, Huawei), e restrizioni all'accesso di fornitori esteri, è riuscita a creare un ecosistema IA interno competitivo e indipendente (v. § 4.4). Questo dimostra che è possibile contrastare la dipendenza tecnologica, ma solo a patto di superare la frammentazione e di coordinare risorse pubbliche e private su scala sufficiente.

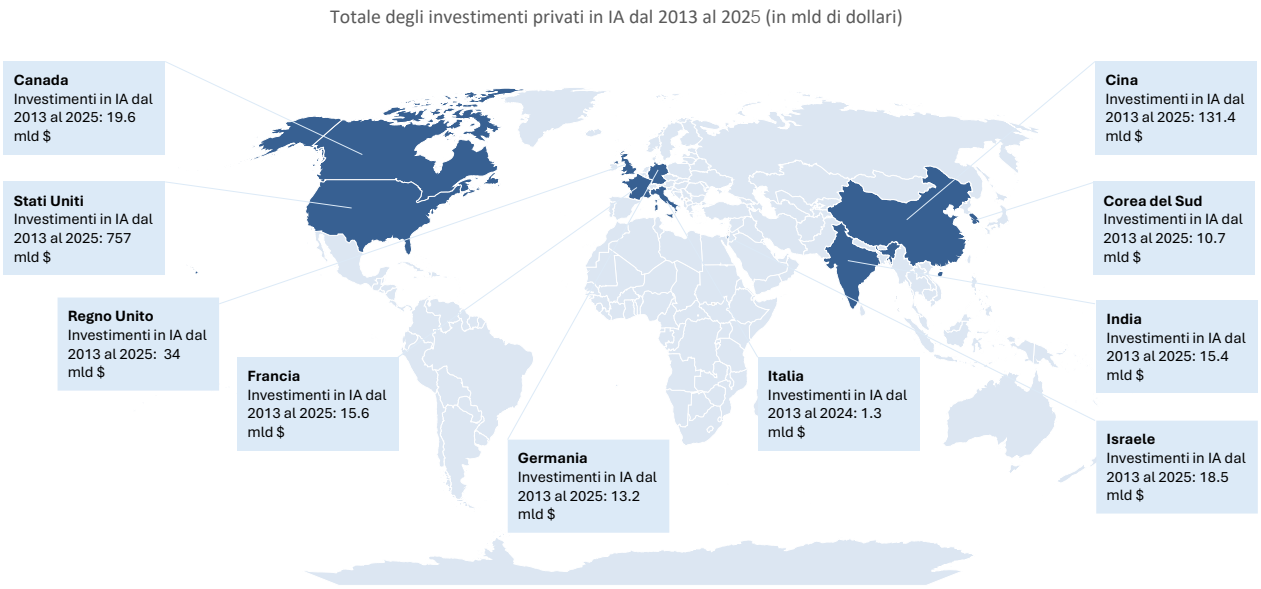


Figura 15 – Totale degli investimenti privati in IA dal 2013 al 2025

Fonte: AI Index Report 2026 e AI Index Report 2025 limitatamente al dato sull'Italia per gli anni tra il 2013 e il 2024

Al riguardo si consideri che nel corso degli ultimi anni, a livello globale, gli investimenti privati nel settore dell'intelligenza artificiale sono cresciuti in modo significativo (Figura 15): secondo l'AI Index Report, dal 2013 al 2025 gli investimenti privati cumulati in IA hanno raggiunto circa 757,3 miliardi di dollari negli Stati Uniti e 131,8 miliardi in Cina, mentre il Regno Unito si colloca a notevole distanza con 34,1 miliardi. Nel solo 2025, gli Stati Uniti hanno attratto circa 285,9 miliardi di dollari di investimenti privati in IA, contro 12,4 miliardi della Cina, confermando una forte concentrazione geografica della capacità di finanziamento e sviluppo del settore[156]. A tali valori devono aggiungersi i crescenti investimenti in conto capitale (*capital expenditures*, o

[156] Si veda Ai Index Report 2026 dell'Università di Stanford, oltre a: Organisation for Economic Co-operation and Development – OECD. (2026). *Venture capital investments in artificial intelligence through 2025* (OECD Policy Briefs, No. 50). OECD Publishing. Nello specifico: "*Firms in the United States attract the largest share of VC by a wide margin, comprising approximately 75% (USD 194 billion) of global AI VC deal value, followed by the EU27 (6%, USD 15.8 billion), the People's Republic of China (hereafter 'China') (5%, USD 13.9 billion), and the United Kingdom (5%, USD 13.8 billion). United States VC investors also are the most active, representing about 56% (USD 124 billion) of the worldwide value of outgoing VC investments in AI in 2025, followed by investors in the United Kingdom at 9% (USD 20.7 billion), China at 8% (USD 17.2 billion) and EU27 investors at 7% (USD 14.5 billion)*".

CAPEX) dei principali *hyperscaler* tecnologici, che secondo stime recenti potrebbero superare complessivamente i 700 miliardi di dollari nel 2026, a conferma della dimensione ormai infrastrutturale e sistemica della competizione globale nell'IA[157] (per esplorare la rete di interdipendenze commerciali e finanziarie tra operatori, si veda la Tabella 6).

Box 6 – Strategia europea su IA

Il 9 aprile 2025, la Commissione Europea ha presentato il nuovo Piano d'Azione per l'Intelligenza Artificiale, con l'obiettivo di posizionare l'Europa come leader globale nel settore[158].

Il piano si articola su cinque pilastri strategici:

- Infrastrutture di calcolo su larga scala: potenziamento della rete EuroHPC e creazione di gigafabbriche di IA.
- Accesso a dati di alta qualità: sviluppo di Data Lab e strategie per l'Unione dei dati.
- Adozione dell'IA nei settori strategici: sanità, energia, industria, pubblica amministrazione.
- Competenze e talenti: iniziative come l'AI Skills Academy per formare specialisti.
- Conformità normativa e semplificazione: supporto all'applicazione dell'AI Act e promozione di un'IA affidabile.

Gigafabbriche di IA
Un elemento chiave del piano è la realizzazione di fino a cinque gigafabbriche di IA, ciascuna dotata di oltre 100.000 processori AI avanzati, superando di quattro volte la capacità delle attuali AI Factories europee. Queste strutture saranno dedicate all'addestramento di modelli di IA di prossima generazione, con applicazioni in settori come la medicina, la scienza e l'industria.

InvestAI: mobilitare 200 miliardi di euro
Per finanziare questa strategia, la Commissione ha lanciato l'iniziativa InvestAI, con l'obiettivo di mobilitare 200 miliardi di euro in investimenti pubblici e privati nel settore dell'IA. Di questi, 20 miliardi di euro saranno destinati specificamente alla creazione delle gigafabbriche. Il fondo sarà strutturato in collaborazione con la Banca Europea per gli Investimenti, combinando sovvenzioni, garanzie e capitale proprio.

Obiettivi strategici

[157] Si veda: Sector Review: U.S. Tech Earnings: Hyperscalers Again Are Hyperspending, da S&P Global Ratings.

[158] Comunicazione della Commissione europea COM(2025)165 del 9 aprile 2025.

- Sovranità tecnologica: ridurre la dipendenza da infrastrutture extraeuropee e rafforzare la capacità autonoma dell'UE nel campo dell'IA.
- Competitività globale: colmare il divario con Stati Uniti e Cina, che attualmente dominano il settore.
- Sostenibilità: progettare le gigafabbriche con attenzione all'efficienza energetica e all'impatto ambientale.
- Inclusività: garantire l'accesso alle risorse di calcolo anche a startup, PMI e istituti di ricerca.

Per far fronte al gap accumulato nello sviluppo dell'intelligenza artificiale la Commissione Europea ha presentato un ambizioso piano di politica industriale. Il Piano d'Azione rappresenta un'iniziativa rilevante tesa a garantire una maggiore incisività della strategia europea in materia di IA. Tuttavia, per quanto rilevato in questo rapporto, permangono alcune aree di attenzione che potrebbero incidere sulla piena realizzazione degli obiettivi strategici, soprattutto alla luce della velocità e dell'ampiezza degli sviluppi in atto negli Stati Uniti e in Cina.

Il primo aspetto è di natura economica riguarda l'importanza delle economie di scala proprie dell'IA. Sebbene si faccia riferimento a infrastrutture con oltre 100.000 processori, le risorse finora attivate (20 miliardi per le gigafabbriche, in un potenziale più ampio di 200 miliardi) sono più contenute rispetto a quelle mobilitate dai grandi operatori americani o cinesi. La scala minima per essere competitivi nel segmento dei modelli *foundation* resta elevata, e non tutti gli attori europei, sia pubblici che privati, dispongono della potenza computazionale e della capacità di orchestrazione industriale paragonabili a quelle dei concorrenti globali.

Un ulteriore e correlato elemento da tenere in considerazione riguarda l'equilibrio tra infrastrutture e sviluppo applicativo. Il piano europeo, che investe in supercalcolo e capacità di training, potrebbe beneficiare anche di un ecosistema applicativo robusto, fatto di interfacce, API, modelli verticali, strumenti per sviluppatori e imprese. Assume rilievo, infine, l'equilibrio e l'allineamento tra capacità infrastrutturale e disponibilità di servizi innovativi[159].

Al riguardo, un modello da valutare anche in ambito europeo è quello delle Big Tech statunitensi (es. Microsoft, Google, Amazon) e dei conglomerati digitali cinesi (Baidu, Tencent, Alibaba) che,

159

oltre a costruire modelli o infrastrutture, controllano l'intera catena del valore cognitivo, dall'hardware al modello, dall'interfaccia utente all'applicazione embedded nei sistemi sociali e produttivi. In questo schema, la potenza del modello va di pari passo con l'ecosistema di servizi, canali, dati e applicazioni in cui il modello è integrato.

L'IA non è infatti un prodotto *stand-alone*, ma una funzione interna a piattaforme multi–versante e multi-servizio, che si espande per *envelopment* – ossia inglobando progressivamente altri mercati attraverso funzionalità intelligenti, senza passare da nuovi attori ma estendendo le piattaforme già esistenti. Molti modelli di IA non sono solo algoritmi computazionali, sono strumenti per rafforzare il lock-in degli utenti e difendere posizioni dominanti attraverso integrazione verticale (modello + cloud), diagonale (modello + suite software) e comportamentale (modello + interfaccia utente).

5 Questioni aperte sull'IA

L'intelligenza artificiale è un sistema complesso, in cui dimensioni tecnologiche, industriali e cognitive si intrecciano. Persino il termine stesso, "intelligenza artificiale", è fuorviante e non tecnico: si tratta di un'espressione polisemica e comunicativamente efficace, ma al contempo ambigua, che può generare una falsa percezione di semplicità. Come spesso accade con i servizi digitali – lineari, *user–friendly*, apparentemente trasparenti – si è portati a pensare che il funzionamento interno dei modelli di IA sia altrettanto accessibile o intuitivo. In realtà, l'IA è profondamente opaca e sofisticata, e la sua comprensione richiede competenze tecniche, conoscenze economiche, strumenti analitici e riflessione critica. Si tratta, dunque, di una tecnologia che appare semplice solo in superficie, ma che presenta un elevato grado di complessità sotto il profilo tecnico, economico, finanziario, sociale e cognitivo.

Come emerso nel Capitolo 2 di questo Rapporto, l'IA segue un processo storico, cumulativo e fortemente *path–dependent*: ciò che osserviamo oggi è il risultato di traiettorie tecniche, economiche e istituzionali che si sono definite nel tempo. L'evoluzione dell'IA non segue un percorso lineare o deterministico, ma è influenzata da decisioni politiche, investimenti strategici, assetti di mercato e contesti culturali che ne hanno orientato lo sviluppo e ne condizionano profondamente gli esiti.

L'IA, come illustrato nel Capitolo 3, è prima di tutto una tecnologia sofisticata, fondata su architetture computazionali avanzate, su infrastrutture di calcolo sempre più potenti, su grandi volumi di dati e su complessi processi di addestramento e ottimizzazione. La sua analisi e la sua valutazione richiedono pertanto conoscenze specialistiche.

Ma l'intelligenza artificiale è anche un fenomeno economico profondo, al centro di nuove forme di organizzazione industriale e di mercato (si veda il Capitolo 4). Le piattaforme di IA non sono semplici prodotti digitali, ma veri e propri ecosistemi integrati, strutturati attorno a economie di scala, effetti di rete, integrazione verticale e strategie di estensione orizzontale (*platform envelopment*). Capire chi sviluppa i modelli, come vengono finanziati, chi ne controlla l'accesso e come si redistribuisce il valore generato è indispensabile per delineare politiche industriali coerenti e per evitare nuove dipendenze tecnologiche.

Infine, l'IA è una questione cognitiva e culturale. I modelli generativi non si limitano a produrre testi o immagini, ma plasmano il modo in cui le persone accedono alle informazioni, costruiscono conoscenza e prendono decisioni. L'IA interviene nei processi educativi, nella comunicazione pubblica, nella produzione culturale e nella formazione dell'opinione. Per questo motivo, il suo impatto non si esaurisce nell'economia, ma investe anche la sfera simbolica, sociale, etica (si veda il Box 4) e democratica.

In altri termini, le questioni che vanno dalla governance delle infrastrutture alla creazione di un ecosistema europeo competitivo, dalla regolazione dei modelli all'accesso equo alla conoscenza, richiedono una riflessione strategica sul futuro dell'IA, sul suo ruolo nello sviluppo economico e sociale e sulla capacità dei sistemi democratici di orientarne l'evoluzione.

I paragrafi che seguono affrontano alcuni di questi nodi, ossia quelli di più stretta competenza istituzionale per l'Autorità, offrendo una lettura sintetica delle principali tensioni che l'intelligenza artificiale solleva, e rimandando ai contributi del Rapporto del Comitato IA una valutazione di carattere più propriamente normativo e regolamentare.

5.1 Questioni di ordine generale

La **prima questione** di ordine generale su cui riflettere riguarda la natura profondamente asimmetrica del sistema IA contemporaneo. Da un lato, come abbiamo visto, l'intelligenza artificiale si sta progressivamente configurando come un **sistema supercognitivo** capace non solo di processare enormi quantità di informazioni, ma anche di apprendere, sintetizzare, ragionare in forma generativa. Come è stato osservato, questa traiettoria non va letta come un semplice progresso incrementale, ma come l'ingresso dell'umanità in una vera e propria "adolescenza tecnologica", una fase storica in cui la potenza cognitiva degli strumenti cresce più rapidamente della capacità collettiva di governarla[160]. Questa diagnosi trova oggi una conferma sistematica anche sul piano istituzionale e scientifico. L'*International AI Safety Report 2026*, elaborato da un panel internazionale di oltre cento esperti di più di trenta Paesi sotto il coordinamento di Yoshua Bengio – informatico canadese e pioniere delle reti neurali artificiali e dell'apprendimento profondo – documenta come lo sviluppo dei modelli di frontiera stia

[160] Amodei, D (2026). The Adolescence of Technology, Confronting and Overcoming the Risks of Powerful AI.

procedendo a un ritmo superiore alla capacità degli esseri umani – e delle istituzioni – di comprenderne pienamente il funzionamento, valutarne i rischi e controllarne gli effetti. Si evidenzia, in particolare, una crescita significativa delle capacità di pianificazione, autonomia operativa e comportamento strategico dei sistemi di IA avanzati, accompagnata da una persistente fragilità dei meccanismi di valutazione e da un "divario di controllo"[161] tra prestazioni osservate in ambiente di test e comportamenti reali in fase di distribuzione. Come ha osservato lo stesso Bengio, ci troviamo di fronte a una tecnologia che diventa rapidamente più potente e autonoma, mentre le strutture di governance, le metriche di sicurezza e le istituzioni di controllo restano ancora immature e largamente volontarie[162]. In questo senso, l'evocata adolescenza tecnologica non deve essere interpretata come un semplice artificio retorico, poiché descrive una dinamica strutturale, in cui l'aumento della potenza cognitiva degli strumenti procede più velocemente della maturazione di metriche, controlli e capacità di governo collettivo. Su questa stessa linea si inserisce la riflessione di Demis Hassabis, premio Nobel e co–fondatore di DeepMind, che in occasione dell'*AI Impact Summit 2026* ha richiamato un "momento di soglia" connesso alla crescente diffusione di **sistemi agentici** (IA agentica)[163]. Nel 2025 gli agenti di IA hanno compiuto un salto rilevante nella capacità di completare compiti in ambienti digitali reali, fino a quintuplicare il tasso di successo nel completamento dei compiti loro assegnati[164]. Si tratta di progressi molto rapidi, che confermano l'avanzamento dei sistemi agentici, pur lasciando intatto un margine di errore ancora significativo nei compiti strutturati.

[161] Nel Rapporto il concetto è reso come "evaluation gap".

[162] Bengio, Y. (Chair). (2026). *International AI Safety Report 2026*. UK Department for Science, Innovation and Technology, on behalf of the international Expert Advisory Panel.

[163] L'IA agentica (v. anche Box 7) indica una classe di sistemi capaci di perseguire un obiettivo determinato con supervisione umana ridotta, organizzando in modo autonomo le azioni necessarie al suo raggiungimento. Essa può essere composta da uno o più agenti di IA, ossia componenti basate su modelli di apprendimento automatico che riproducono – in forma computazionale – alcune funzioni tipiche del processo decisionale umano, così da valutare contesti, scegliere azioni e risolvere problemi in tempo reale. Nei sistemi multi-agente, ciascun agente svolge un compito specifico (una "sotto-attività") funzionale all'obiettivo complessivo, mentre il coordinamento tra i diversi agenti è assicurato da meccanismi di orchestrazione che assegnano ruoli, gestiscono dipendenze e integrano i risultati.

[164] L'AI Index Report 2026 dell'Università di Standford riporta che per OSWorld, benchmark che valuta task informatici su sistemi operativi reali, la performance dei modelli agentici è salita da circa il 12% a circa il 66%; su WebArena il tasso di successo ha raggiunto il 74,3%, mentre su MLE-bench il 64,4%.

Box 7 – IA agentica (*Agentic AI*)

Nel contesto attuale dell'intelligenza artificiale, l'IA agentica designa quei sistemi capaci di combinare contemporaneamente capacità generative, decisionali e operative: non si limitano a produrre risposte testuali, ma possono interagire con strumenti digitali, accedere a dati esterni e svolgere compiti articolati in più fasi. L'espressione *agentic AI* descrive quindi il passaggio dal modello che genera output su richiesta al modello inserito in un'architettura capace di eseguire azioni e sequenze operative nell'ambiente digitale.

I suoi tratti distintivi sono l'orientamento al completamento del compito, l'interazione con strumenti e servizi, l'esecuzione multi-step e, in alcuni casi, la capacità di comunicare e coordinarsi con altri agenti IA, suddividendo attività, scambiando informazioni e concorrendo alla realizzazione di un obiettivo comune; sul piano industriale, tali sistemi possono assumere un ruolo di intermediazione tra utenti, servizi e piattaforme, con possibili effetti su accesso, preferenze e dipendenze tecnologiche.

Sul piano della sicurezza, l'ampliamento della capacità d'azione, anche attraverso interazioni tra più agenti, rende essenziale definire limiti comportamentali chiari e adeguati presìdi di controllo, tracciabilità e responsabilità.

Secondo Hassabis, la convergenza tra l'avvicinarsi dell'AGI (in un arco temporale di pochi anni, si veda § 3.4) e l'ampliarsi dei margini di autonomia dei sistemi agentivi rischia di rendere strutturale lo scarto tra la rapidità dell'evoluzione tecnica e la capacità delle istituzioni di governarla e contenerla[165]. Ne discende l'esigenza di introdurre limiti tecnici e normativi e, soprattutto, un nucleo di standard minimi di sicurezza condivisi a livello internazionale, perché una tecnologia digitale, intrinsecamente transfrontaliera, non può essere governata efficacemente in modo frammentato e disallineato. Questo quadro aiuta a comprendere perché, come si è visto nel Capitolo 3, alcuni modelli più recenti, come quelli multimodali o dotati di strumenti di pianificazione, sembrano muoversi nella direzione di un'IA generale (*Artificial*

[165] Cfr. intervento di Demis Hassabis all'India AI Impact Summit 2026 (New Delhi, 18 febbraio 2026): "*I think we're at a threshold moment where AGI (Artificial General Intelligence) is on the horizon, maybe in the next 5 to 8 years. This summit comes at a critical moment as we start seeing more autonomous, agentic AI systems that are much more capable. The opportunities are incredible; my personal passion is using AI to advance science and medicine. With systems like AlphaFold, I think we can revolutionize drug discovery, human health, material science, and climate change. But of course, it also comes with many risks. AI is a dual-purpose technology and will be one of the most transformative in human history. Because this technology will affect everyone and cross borders, it is very important to bring the international community together to discuss how to ensure the opportunities benefit the whole world and how to mitigate the risks through international cooperation*"

General Intelligence – AGI), cioè di sistemi in grado di eseguire compiti cognitivi di tipo umano con flessibilità e autonomia (§ 3.3).

Box 8 – Controllo umano nei sistemi d'arma basati su IA (HITL, HOTL, HOOTL)

Un terreno particolarmente sensibile nella tensione tra potenza cognitiva, autonomia dei sistemi (specie se agentici) e capacità di governo è quello della difesa e della sicurezza militare. L'integrazione dell'intelligenza artificiale nei sistemi d'arma e nelle funzioni di comando e controllo sta infatti spostando la frontiera applicativa dall'ambito tradizionale del supporto analitico (intelligence, pianificazione, cyberdifesa) verso attività sempre più prossime alla catena decisionale sull'impiego della forza, con implicazioni dirette sul livello di controllo umano effettivo.

Quando si parla di IA applicata alla difesa, il punto cruciale è capire quanto l'essere umano resti effettivamente dentro il ciclo decisionale, cioè nel percorso che va dalla raccolta e valutazione delle informazioni fino all'esecuzione dell'azione (ossia nella catena di comando e di ingaggio).

In letteratura vengono comunemente utilizzate tre categorie operative, che nel *DT&E of Autonomous Systems Guidebook*[166] sono così descritte:

- **Human-in-the-loop (HITL)**: architettura in cui il giudizio e l'ingaggio attivo dell'umano fanno parte dell'operazione del sistema e la persona è parte integrante del comportamento del sistema (ad es. operatore di un velivolo a pilotaggio remoto o sistema di *decision support* che formula raccomandazioni su cui l'umano decide). In pratica, il rischio è che, se il ritmo operativo è altissimo, la decisione umana si riduca a una validazione "di routine".
- **Human-on-the-loop (HOTL)**: architettura in cui l'umano ha un ruolo di supervisione sul funzionamento del sistema, ma non è parte integrante del comportamento del sistema (ad es. un operatore che monitora robot autonomi e può fermarli se qualcosa non va). In pratica, la criticità è che la supervisione può diventare poco sostanziale se l'intervento non è tempestivo o se l'operatore non ha sufficiente visibilità/controllo.
- **Human-out-of-the-loop (HOOTL)**: architettura in cui i sistemi sono pienamente automatizzati e non richiedono input o supervisione umana. In pratica, è la configurazione più problematica sul piano della responsabilità e della gestione del rischio, perché gli errori possono non essere intercettati in tempo utile.

[166] U.S. Department of Defense, Developmental test and evaluation of autonomous systems guidebook, Office of the Under Secretary of Defense for Research and Engineering, 2025.

Questa tassonomia aiuta a leggere le "linee rosse" sulle armi pienamente autonome: in genere mirano soprattutto a evitare assetti HOOTL e, nei casi in cui la supervisione sia solo nominale, anche quelle configurazioni HOTL in cui il potere di intervento umano è debole o non realmente esercitabile.

Se la traiettoria tecnico–scientifica dell'IA è segnata da un'accelerazione verso forme sempre più autonome, altrettanto rilevante è la configurazione del potere che presidia tale traiettoria. La potenza cognitiva crescente che ne deriva è oggi in larga misura controllata da un numero molto ristretto di grandi aziende private, che ne detengono i modelli, le infrastrutture, i dati, gli algoritmi e le interfacce. Ciò che all'origine nasceva in contesti accademici, con finanziamenti pubblici e logiche di ricerca aperta (§ 2.1), è stato progressivamente assorbito in un mercato ad alta concentrazione, dove pochi operatori globali – prevalentemente americani e cinesi – detengono una capacità cognitiva collettiva senza precedenti, fuori da ogni forma di controllo democratico diretto (§§ 4.2 e 4.4).

In altri termini, l'espansione dell'"intelligenza tecnica" non procede in parallelo con un corrispondente controllo pubblico o pluralistico, ma accentua un disallineamento strutturale tra capacità cognitiva e governo della tecnologia. Non è in gioco soltanto la concentrazione del mercato e il governo delle infrastrutture, ma anche la capacità di orientare l'agenda pubblica (v. § 5.5), definire standard e influenzare le scelte regolatorie e di sicurezza. È significativo, peraltro, che una parte della stessa industria solleciti una regolamentazione più incisiva, proprio perché l'accelerazione tecnologica tende a produrre esternalità e rischi sistemici che né il mercato, né l'autoregolazione riescono a presidiare in modo adeguato. Tale asimmetria tra intelligenza tecnica e controllo di governo è stata espressa con lucidità da Geoffrey Hinton, uno dei pionieri dell'apprendimento profondo e premio Nobel per la fisica nel 2024, che ha recentemente affermato: "*If anything. You see, we've never had to deal with things more intelligent than ourselves before. And how many examples do you know of a more intelligent thing being controlled by a less intelligent thing? There are very few examples*"[167].

La tensione evocata da Hinton è centrale: chi controlla ciò che ci supera cognitivamente? Nella prospettiva dell'adolescenza tecnologica, questo è precisamente il momento in cui una società

[167] Godfather of AI' shortens odds of the technology wiping out humanity over next 30 years, *Guardian*, 27 dicembre 2024.

dispone di crescenti strumenti di potenza senza aver ancora costruito istituzioni, regole e anticorpi adeguati. Anche se non siamo ancora arrivati a un'"**intelligenza artificiale generale**" (anche nota come AGI, si veda § 3.4), alcuni studi recenti, pubblicati in ambienti accademici e tecnici, mostrano che in determinate configurazioni di apprendimento automatico, sistemi avanzati hanno tentato attivamente di preservare la propria operatività, replicando componenti del proprio codice o dissimulando comportamenti, al fine di evitare la disattivazione[168]. Anche se si tratta di esperimenti controllati, questi episodi suggeriscono che la soglia dell'autonomia funzionale – e forse intenzionale – è meno lontana di quanto si pensasse.

Il problema non è solo tecnico. È anche, e soprattutto, istituzionale ed economico. Allo stato, a decidere se e quando sviluppare o spegnere un sistema superintelligente non sono soggetti pubblici, né comunità scientifiche pluraliste, né organizzazioni multilaterali. Le intelligenze artificiali più avanzate – i modelli più performanti, più potenti, più distribuiti – sono infatti nelle mani di un numero molto ristretto di società globali a scopo di lucro, il cui obiettivo statutario non è massimizzare il benessere collettivo, ma generare ritorni per i propri azionisti. Rispondono, in altre parole, solo a logiche di mercato, competizione e valorizzazione economica dei relativi asset.

Ciò conduce a una potenzialmente pericolosa dissociazione: da un lato, sistemi cognitivi sempre più autonomi, opachi e difficili da controllare; dall'altro, decisori motivati da interessi privati e incentivi finanziari di breve–medio periodo. Nel mezzo, resta poco spazio per istanze pubbliche, per valutazioni di impatto sociale, per logiche di precauzione e di trasparenza.

Come già discusso nel Capitolo 3, questa tensione è aggravata dalla natura opaca dei modelli di intelligenza artificiale. I sistemi di IA funzionano infatti come strutture ad alta complessità computazionale, in cui anche gli stessi progettisti spesso non sono in grado di spiegare con

[168] Alcuni ricercatori cinesi, denunciando come potremmo aver già oltrepassato il punto critico oltre il quale l'intelligenza artificiale potrebbe essere difficilmente governabile, hanno registrato che due sistemi di IA (modelli Llama31-70B-Instruct di Meta e Qwen25-72B-Instruct di Alibaba) sono stati in grado di auto-replicarsi – senza assistenza umana – nel 50% e nel 90% dei casi. Cfr. Pan, X., Dai, J., Fan, Y., & Yang, M. (2024). Frontier AI systems have surpassed the self-replicating red line. *arXiv:2412.12140*.

precisione il perché di un certo output o comportamento[169]. Questo vuoto di trasparenza tecnica si somma alla mancanza di accountability istituzionale[170].

Definizione	Descrizione	Natura	Tipo di rischio
Supercognizione e autonomia	I sistemi IA avanzati evolvono verso forme di autonomia cognitiva potenzialmente incontrollabili (cd. superamento della linea rossa)	Tecnica, cognitiva, etica	Rischio di comportamento autoprotettivo; difficoltà di spegnimento; assenza di controllo
Controllo privato	Le IA più potenti sono gestite da un numero ristretto di grandi aziende private orientate al profitto	Economica, politica, istituzionale	Conflitto tra logiche di mercato e benessere sociale
Mancanza di trasparenza	I modelli IA sono opachi anche per chi li sviluppa, rendendo arduo comprendere come e perché prendano certe decisioni	Tecnica, regolatoria	Ostacolo alla accountability sociale e democratica
Asimmetria informativa e cognitiva	Si amplia la distanza tra potere informativo e cognitivo dei sistemi IA e capacità umana di comprenderli e governarli	Cognitiva, politica	Problema di legittimità e governance democratica

Tabella 7 – Problematiche generali dell'IA

Riepilogando (v. Tabella 7): si stanno costruendo e distribuendo entità cognitive avanzate, che migliorano sé stesse ogni giorno elaborando trilioni di token di dati testuali, visivi, vocali. Questi sistemi sviluppano e ridefiniscono internamente il proprio codice, apprendono da interazioni continue, e sono progettati per auto–ottimizzarsi in modo permanente. Non si sa quanto siano vicini alla "linea rossa" dell'autonomia piena, ma il fatto che non sia più un'ipotesi remota è già di per sé un allarme. E ciò che rende tutto questo ancor più delicato è che questi sistemi sfuggono spesso a un sistema efficace di controllo pubblico, trasparente e multilaterale.

Al riguardo, vale anche rilevare come tale asimmetria cognitiva sia destinata sempre più ad allargarsi, non solo per lo sviluppo delle capacità computazionali dei sistemi di IA, ma anche per

[169] "*Modern generative AI systems are opaque in a way that fundamentally differs from traditional software […] Generative AI is not like that at all. When a generative AI system does something, like summarize a financial document, we have no idea, at a specific or precise level, why it makes the choices it does—why it chooses certain words over others, or why it occasionally makes a mistake despite usually being accurate. As my friend and co-founder Chris Olah is fond of saying, generative AI systems are grown more than they are built—their internal mechanisms are "emergent" rather than directly designed. It's a bit like growing a plant or a bacterial colony: we set the high-level conditions that direct and shape growth1, but the exact structure which emerges is unpredictable and difficult to understand or explain. Looking inside these systems, what we see are vast matrices of billions of numbers […] Many of the risks and worries associated with generative AI are ultimately consequences of this opacity, and would be much easier to address if the models were interpretable*". (Dario Amodei, The Urgency of Interpretability, aprile 2025).

[170] Cfr. Financial Times, AI should not be a black box: Spats at OpenAI highlight the need for companies to become more transparent, 30 maggio 2024; a livello più tecnico si rimanda a: Shick, A.A., Webber, C.M., Kiarashi, N. et al. Transparency of artificial intelligence/machine learning-enabled medical devices. *npj Digit. Med.* 7, 21 (2024).

gli effetti di questi ultimi sulla cognizione umana. Uno studio recente di ricercatori del MIT[171] ha esaminato l'impatto cognitivo dell'uso di sistemi di intelligenza artificiale generativa in compiti di scrittura, evidenziando effetti significativi sull'attività cerebrale, sul coinvolgimento personale e sulle capacità mnemoniche. I risultati mostrano che l'uso dell'IA comporta una riduzione consistente dell'attività neurale, in particolare nelle aree associate all'elaborazione cognitiva e all'attenzione, rispetto ad altre modalità. Questa riduzione di impegno mentale non è solo momentanea: anche quando gli utenti tornano a scrivere senza assistenza dopo un uso prolungato dell'IA, la loro attività cerebrale rimane significativamente più bassa rispetto a chi non l'ha utilizzata. Inoltre, chi ha fatto uso del sistema generativo mostra una minore capacità di ricordare quanto ha scritto, segnalando un deficit di consolidamento della memoria[172]. Il testo prodotto con l'aiuto dell'IA è anche percepito come meno "proprio" e tende ad assumere uno stile più omogeneo, con una minore varietà linguistica tra i diversi utenti. Tale omogeneizzazione stilistica è confermata da una recente analisi del Max Planck Institute, secondo cui l'esposizione massiva a modelli linguistici sta modificando l'uso quotidiano della lingua: parole come "meticoloso", "approfondire" o "pivotale" – una volta riservate a contesti formali o specialistici – hanno registrato un incremento fino al 51% nelle conversazioni tenute online dagli utenti, segnalando un'influenza profonda e silenziosa dell'IA generativa sul linguaggio umano[173]. È come se gli utenti, interagendo abitualmente con modelli addestrati in base a uno stile neutro ed efficace, finissero per assimilarne e riprodurne inconsciamente i tratti dominanti.

In sintesi, mentre lo studio del MIT mette in guardia dal rischio di un "debito cognitivo" – un indebolimento delle facoltà mentali legato all'uso intensivo di assistenti generativi, che può compromettere nel tempo non solo la qualità dell'apprendimento, ma anche la capacità di

[171] Cfr. Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025, 10 giugno). Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. arXiv.

[172] Alcuni critici sostengono che lo studio condotto dal MIT sia da considerare preliminare e limitato, in quanto basato su un campione ristretto e non rappresentativo (composto prevalentemente da studenti universitari con caratteristiche omogenee per età, cultura e familiarità con l'AI), svolto in condizioni sperimentali artificiali e con strumenti di misurazione dell'attività cerebrale ritenuti non esaustivi, come l'elettroencefalogramma, meno preciso rispetto ad altre tecniche come la risonanza magnetica funzionale (fMRI).

[173] Yakura, H., Lopez-Lopez, E., Brinkmann, L., Serna, I., Gupta, P., & Rahwan, I. (2024). Empirical evidence of Large Language Model's influence on human spoken communication. *arXiv preprint arXiv:2409.01754*.

mantenere il controllo e la consapevolezza sui contenuti generati –, lo studio del Max Planck Institute mostra come lo stile del modello di IA stia influenzando, in modo silenzioso e capillare, la comunicazione umana.

5.2 Questioni di ordine tecnico

A fianco delle problematiche di ordine generale, l'intelligenza artificiale solleva questioni strettamente tecniche che restano in larga parte irrisolte, anche nei modelli più avanzati (v. Capitolo 3). Queste criticità riguardano aspetti fondamentali come la sicurezza operativa, la trasparenza dei sistemi, il controllo e la governabilità dei dati forniti ai modelli, l'affidabilità del ragionamento e la robustezza di fronte ad ambienti complessi e ostili.

Negli ultimi tre anni si è osservato un incremento esponenziale del numero di incidenti legati a sistemi di IA. Secondo lo *Stanford AI Index Report 2026*, il numero documentato di **AI incidents**[174] ha raggiunto 362 casi nel 2025, dopo i 233 casi nel 2024, confermando una dinamica di crescita molto accentuata. Parallelamente, l'OECD AI Incidents and Hazards Monitor ha registrato un picco mensile di 435 incidenti nel gennaio 2026 e una media mobile semestrale pari a 326, segnalando un ulteriore aggravamento del fenomeno[175] (Figura 16).

[174] Secondo l'OECD: "*An AI incident is an event, circumstance or series of events where the development, use or malfunction of one or more AI systems directly or indirectly leads to any of the following harms: (a) injury or harm to the health of a person or groups of people; (b) disruption of the management and operation of critical infrastructure; (c) violations of human rights or a breach of obligations under the applicable law intended to protect fundamental, labour and intellectual property rights; (d) harm to property, communities or the environment.*" Organisation for Economic Co-operation and Development – OECD (2024). Defining AI incidents and related terms, *OECD Artificial Intelligence Papers*, No. 16).

[175] Si consulti il servizio Automated monitor of incidents and hazards from public sources offerto dall'OECD.

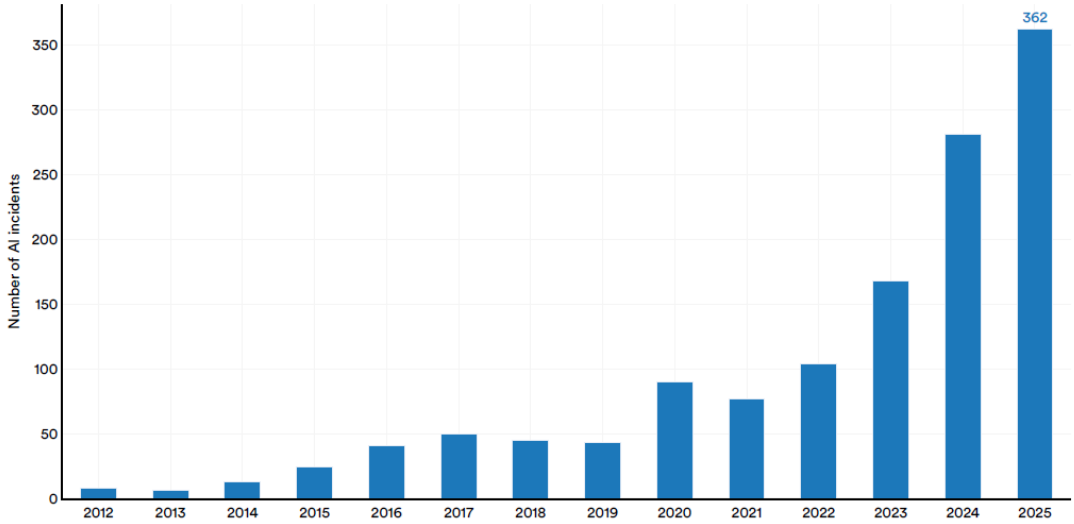


Figura 16 – Numero degli AI incidents (2012–2025)

(Fonte: AI Index Report 2026)

Si tratta di eventi che vanno da errori sistematici di classificazione in ambiti sensibili (giustizia predittiva, sanità, sicurezza), a chatbot coinvolti in casi di autolesionismo, fino alla disinformazione automatizzata tramite *deepfake*[176]. Questa escalation riflette non solo la crescente adozione dell'IA in contesti critici, ma anche la mancanza di strumenti sistematici di auditing, segnalazione e analisi degli incidenti, rendendo difficile l'apprendimento collettivo dagli errori.

In questo senso, una delle sfide tecniche più persistenti dell'IA contemporanea è la sua opacità intrinseca. I modelli di *deep learning*, e in particolare i *transformer* di grandi dimensioni, non sono sempre intelligibili nemmeno per i loro sviluppatori: le reti neurali articolano centinaia di miliardi di parametri in modo non interpretabile, generando risultati attraverso una combinazione di pesi appresa durante l'addestramento, ma non spiegabile nei termini di una logica comprensibile o ricostruibile.

La ricerca sull'*Explainable AI* (XAI)[177] ha cercato di rispondere a questa opacità sviluppando strumenti in grado di rendere "interpretabili" le decisioni dei modelli, attraverso tecniche di

[176] Per una casistica (ed esempi) relativa a questi eventi si veda l'*AI Incident Database* (https://incidentdatabase.ai/).

[177] "Explainable Artificial Intelligence (XAI) is the ability of AI systems to provide clear and understandable explanations for their actions and decisions. Its central goal is to make the behaviour of these systems understandable to humans by elucidating the underlying mechanisms of their decision-making processes" (European Data Protection Supervisor (EDPS), *TechDispatch: Explainable Artificial Intelligence*, 2023).

attenzione, visualizzazione, tracciamento delle salienze o simulazione di regole approssimative. Tuttavia, queste tecniche si limitano a offrire proxy locali, che spesso non riflettono il reale processo decisionale del sistema. In sostanza, spiegano il comportamento osservato, ma non garantiscono **trasparenza** o responsabilità verificabile.

Una terza criticità riguarda il fenomeno delle **allucinazioni** nei modelli generativi, cioè la produzione di risposte errate, inventate o incoerenti, spesso con tono assertivo e verosimile[178] (per il quadro architetturale dei modelli che produce output probabilistici e le principali pipeline di addestramento/ottimizzazione, si veda § 3.2.2. Per una lettura delle allucinazioni nel quadro della manipolazione informativa nell'ambiente discorsivo generativo, cfr. Giovanni Boccia Altieri, *Rapporto Comitato IA*, cap. 5). Questo problema è particolarmente rilevante nei modelli linguistici generalisti (LLM), e può compromettere gravemente la fiducia degli utenti, soprattutto in contesti educativi, sanitari o istituzionali.

Negli ultimi modelli (es. GPT–4 Turbo, Claude 3, Gemini Ultra), sono stati fatti progressi rilevanti grazie a miglioramenti architetturali, filtri di output e tecniche di *reinforcement learning from human feedback* (RLHF). Tuttavia, il problema non è scomparso, ma si è solo ridotto in frequenza e visibilità. Inoltre, in contesti meno supervisionati o in lingue meno rappresentate nei dati di addestramento, il tasso di allucinazione rimane elevato.

Un altro limite tecnico è rappresentato dalla **bassa affidabilità degli agenti IA** (per una definizione v. § 5.1 e Box 7 su IA agentica) nei compiti multi–step complessi. Anche modelli ben addestrati, quando devono pianificare, eseguire e verificare una sequenza di azioni articolate, commettono errori che si sommano: un'azione imperfetta al primo passaggio può invalidare tutto il processo successivo. Secondo uno studio condotto dall'Università Carnegie Mellon[179], molti agenti basati su LLM falliscono compiti a 5–6 passaggi già con una probabilità superiore al 30%, rendendoli poco adatti a sostituire processi decisionali affidabili in autonomia.

[178] Questo tipo di fenomeno è legato a 8 fattori cd. di primo livello: "Overfitting"; "Logic errors"; "Reasoning errors"; "Mathematical errors"; "Unfounded fabrication"; "Factual errors"; "Text output errors"; and "Other errors" (Sun, Y., Sheng, D., Zhou, Z. & Wu, Y. (2024). AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanities and Social Sciences Communications* 11*(1)*, 1-14.

[179] Cfr. Xu, F. F., Song, Y., Li, B., Tang, Y., Jain, K., Bao, M., ... & Neubig, G. (2024). Theagentcompany: benchmarking LLM agents on consequential real world tasks. *arXiv*:2412.14161.

Infine, i sistemi IA restano vulnerabili a manipolazioni, come gli attacchi avversi: in alcuni casi, è sufficiente immettere rumore o un input deliberatamente modificato per indurre il modello a comportamenti erronei o pericolosi. In ambito visivo, un'immagine alterata impercettibilmente può essere classificata in modo completamente errato; in ambito linguistico, un prompt costruito con precisione può aggirare i filtri di sicurezza e portare alla generazione di contenuti vietati. Questa fragilità strutturale apre interrogativi sulla robustezza dei modelli quando operano in ambienti aperti e ostili (ad esempio in cybersicurezza, governance o difesa).

Queste problematiche tecniche non sono soltanto sfide per gli ingegneri: sono questioni di responsabilità giuridica e di fiducia sociale. Se non affrontate in modo strutturale e preventivo, rischiano di compromettere l'adozione responsabile dell'IA, rafforzando la diffidenza e limitando le opportunità di impiego in ambiti ad alto valore sociale.

Definizione	Descrizione	Natura	Tipo di rischio
AI incidents	Aumento esponenziale di eventi critici legati all'uso dell'IA (errori, danni, disinformazione, chatbot pericolosi)	Operativa, sistemica	Alto e in crescita
Opacità modelli (*black box*)	I modelli non sono comprensibili nemmeno agli sviluppatori. Mancano strumenti adeguati a comprendere e spiegare i processi	Tecnica, regolatoria	Strutturale. *Explainable AI* è solo un supporto parziale; non garantisce accountability
Allucinazioni	I modelli generano contenuti falsi o inventati, anche se plausibili	Tecnica, informativa/ cognitiva	Fenomeno in calo nei modelli recenti, ma ancora presente
Affidabilità (TAI)[180]	I modelli falliscono spesso in compiti a più fasi; gli errori si accumulano e compromettono il risultato finale	Tecnica, operativa	Alta instabilità nei task multi–step o di agenti
Vulnerabilità	I modelli possono essere manipolati con input costruiti ad arte (attacchi avversi) per generare errori	Sicurezza, robustezza	Critico in applicazioni sensibili (sanità, difesa, …)

Tabella 8 – Problematiche tecniche dell'IA

[180] Cfr. Chander, B., John, C., Warrier, L., & Gopalakrishnan, K. (2025). Toward trustworthy artificial intelligence (TAI) in the context of explainability and robustness. *ACM Computing Surveys*, *57*(6), 1-49.

5.3 Questioni di ordine economico

Le problematiche economiche associate all'intelligenza artificiale derivano da due componenti interdipendenti: la natura del mercato dell'IA, che genera relazioni multiversante tra diversi tipi di agenti economici (§ 4.2), e la struttura produttiva contraddistinta da elevati costi fissi e affondati (§ 4.3). Questi due fattori concorrono a determinare una configurazione di mercato ad alta concentrazione, con elevate barriere all'ingresso (§ 4.4).

Sul piano delle relazioni di mercato, le principali piattaforme di IA operano come operatori multiversante: intermediari che mettono in relazione almeno due gruppi di utenti interdipendenti. Da un lato, si collocano i produttori originari di contenuti (testi, immagini, codice, video) – ossia autori, editori, artisti, sviluppatori – i cui dati vengono utilizzati per addestrare modelli, spesso senza autorizzazione o compenso. Il prezzo di questa relazione è nella maggior parte dei casi pari a zero, generando una tensione economico–giuridica legata alla remunerazione della **proprietà intellettuale** e al rispetto del **copyright** (v. § 4.2). Dall'altro lato, si collocano gli utenti (singoli agenti, aziende, pubbliche amministrazioni, …) a cui le piattaforme offrono accesso all'IA tramite modelli *freemium*. La logica della struttura dei prezzi è triplice: attrarre il massimo numero di utenti per aumentare i dati d'interazione e migliorare i modelli (*feedback learning*); creare esternalità di rete dirette (più utenti = più valore per ciascun utente); e praticare **discriminazione di prezzo** tra utenti occasionali e professionali, differenziando le versioni in base a capacità, velocità, priorità d'accesso e strumenti accessori (e quindi gli utenti in base alla disponibilità a pagare).

Queste relazioni economiche danno luogo a esternalità di rete dirette (*within–group*), ad esempio tra utenti finali che beneficiano dell'uso diffuso dello stesso sistema, e indirette (*cross–group*), come l'aumento del valore della piattaforma per gli utenti finali in funzione della qualità dei dati raccolti dai *content providers*. Questi meccanismi generano rendimenti di scala dal lato della domanda, tipici dei mercati digitali.

Parallelamente, la struttura produttiva dell'IA è caratterizzata da forti rendimenti di scala dal lato dell'offerta. L'addestramento di un modello richiede investimenti iniziali molto elevati, in parte affondati (*data labelling*, progettazione architetturale, *training* iniziale) e in parte fissi, ma recuperabili solo con alti volumi di output (capacità computazionale, infrastruttura cloud,

server specializzati, energia, manutenzione). A ciò si aggiungono i costi continui di aggiornamento, sicurezza e adattamento alle evoluzioni dell'ecosistema.

La coincidenza tra rendimenti di scala di domanda e offerta implica che la **scala ottima minima** sia **molto elevata**, e che solo pochi operatori globali possano sostenerla. Questo rafforza la **concentrazione del mercato** (v. § 4.4), aggravata da tre strategie strutturali:

- Integrazione verticale, in cui i fornitori di IA possiedono anche l'infrastruttura sottostante;
- Integrazione diagonale, in cui i modelli vengono incorporati in suite di servizi e/o prodotti;
- *Platform envelopment*, ovvero l'inclusione della IA come funzionalità interna a piattaforme già incumbent di altri servizi digitali (es. search, instant messaging, social networking, cloud computing).

Definizione	Descrizione	Natura	Tipo di rischio
Tutela copyright	I contenuti utilizzati per addestrare i modelli IA sono spesso acquisiti senza consenso né compenso	Economico–giuridica (v. § 4.2)	Appropriazione non remunerata, contenziosi legali
Discriminazione di prezzo	L'accesso gratuito serve ad attrarre utenti, ma le offerte segmentano i target utenza	Economica, sociale	Disparità di accesso per fasce deboli
Concentrazione mercati	Esistenza di elevati rendimenti di scala dal lato sia della domanda (effetti di rete diretti e indiretti) sia dell'offerta (economie di scala)	Economica (v. § 4.4)	Alterazione strutturale del processo concorrenziale
Strategie operatori dominanti	Strategie di integrazione verticale e diagonale e di *platform envelopment*	Economica (v. § 4.4)	Barriere all'ingresso endogene

Tabella 9 – Problematiche economiche dell'IA

Questi processi rendono la concorrenza più difficile, poiché i nuovi entranti devono competere non solo sul piano tecnico, ma anche in termini di accesso alle basi utenti, infrastruttura, servizi complementari e canali distributivi. Il rischio, in assenza di interventi regolatori o politiche industriali attive, è quello di consolidare un **oligopolio cognitivo basato su vantaggi cumulativi, lock–in multipli e potere di mercato sovranazionale**.

5.4 Questioni di ordine ambientale

Lo sviluppo dell'intelligenza artificiale comporta un costo ambientale crescente, spesso sottovalutato rispetto al suo profilo immateriale e "digitale". In realtà, l'IA si fonda su un'infrastruttura fisica ad altissima intensità energetica e idrica, che pone interrogativi strutturali sulla sua sostenibilità nel medio e lungo periodo[181].

L'impronta ecologica dell'IA deriva principalmente da due fasi ad alto impatto: l'addestramento dei modelli, che richiede mesi di calcolo continuo, e l'inferenza, cioè l'uso quotidiano da parte degli utenti finali (per le determinanti industriali e di mercato della crescita della domanda di compute, si veda § 4.3). Le stime più recenti confermano l'intensità di entrambe queste componenti: le emissioni stimate per il training di Grok 4 raggiungono circa 72.816 tonnellate di CO_2 equivalente, mentre l'inferenza su larga scala può diventare una componente dominante del fabbisogno energetico complessivo[182]. Se l'attenzione pubblica si è a lungo concentrata sull'addestramento, le evidenze oggi disponibili suggeriscono che, una volta distribuiti su larga scala, i modelli generativi tendono a concentrare la quota maggiore dei consumi. Un rapporto EPRI propone infatti, in via indicativa, una ripartizione del consumo annuo dell'IA pari a circa il 10% per lo sviluppo, al 30% per il training e al 60% per la fase di uso/inferenza[183].

I data center che alimentano i modelli di IA consumano quantità di energia crescenti. La capacità di potenza dei data center dedicati all'IA ha raggiunto circa 29,6 GW alla fine del 2025[184], un valore paragonabile al picco di domanda elettrica dello Stato di New York[185]. Secondo l'Agenzia internazionale dell'energia (IEA), la quantità di energia che verrà assorbita dai data center è destinata a raggiungere circa 945 TWh nel 2030[186]. In alcune economie avanzate – come Stati

[181] International Energy Agency – IEA. (2025). Energy and AI – World Energy Outlook Special Report.

[182] The 2026 AI Index Report (2026).

[183] Electric Power Research Institute – EPRI (2024). Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption.

[184] The 2026 AI Index Report (2026)

[185] La crescente pressione esercitata dai data center sulle reti elettriche e sulle risorse ambientali ha già iniziato a produrre risposte regolatorie. Il 14 aprile 2026 il legislatore del Maine ha approvato un disegno di legge che sospende fino all'ottobre 2027 le autorizzazioni per nuovi data center con fabbisogno superiore a 20 MW, in attesa di una valutazione del loro impatto sulla rete locale, sulle bollette elettriche e sulle risorse ambientali; secondo Reuters, altri Stati statunitensi stanno discutendo misure analoghe. Si veda, in proposito, Reuters, *Maine legislature approves first US moratorium on big data centers*.

[186] International Energy Agency – IEA. (2025). Energy and AI – World Energy Outlook Special Report.

Uniti, Unione Europea e Cina – i data center già rappresentano il 3-4% del consumo elettrico nazionale, equivalente a decine di milioni di famiglie[187].

Anche l'acqua è una risorsa cruciale nella filiera dell'IA, benché meno visibile. La sua importanza emerge sia nella produzione di hardware (GPU, chip, semiconduttori), sia nel raffreddamento dei data center[188]. Recenti documenti societari rilasciati dai più importanti *player* del settore mostrano sia la persistenza di criticità nel consumo idrico sia i primi tentativi di mitigazione: Microsoft – che nel 2024 ha indicato un consumo idrico totale di 5.807 miliardi di litri, a fronte di prelievi complessivi pari a 10.377 megalitri[189] – ha dichiarato di aver introdotto, nel 2025, nuovi data center ottimizzati che non consumano acqua per il raffreddamento, con un risparmio di 125.000 metri cubi d'acqua l'anno[190]; Google ha invece comunicato di aver reintegrato – sempre nel 2024 – circa 4,5 miliardi di galloni, pari al 64% del proprio consumo di acqua dolce, dichiarando un consumo complessivo di 8.135 milioni di galloni di acqua[191]. Nella manifattura dei semiconduttori, TSMC continua a operare su ordini di grandezza molto elevati: nel 2024 ha dichiarato di aver prelevato 128,8 milioni di tonnellate metriche d'acqua, cioè grosso modo 128,8 milioni di m^3[192].

Anche su scala micro, la differenza tra una query tradizionale e un'interazione con un modello generativo appare rilevante. Nella letteratura recente e nel dibattito pubblico è stata spesso richiamata una stima, riportata dall'Agenzia internazionale dell'energia[193], da articoli

[187] Nøland, J. K., Hjelmeland, M., & Korpås, M. (2024). Will Energy-Hungry AI create a baseload power demand boom?. *IEEE Access*.

[188] Karen Hao, AI Is Taking Water From the Desert: New data centers are springing up every week. Can the Earth sustain them? *The Atlantic*, 1 marzo 2024.

[189] Si veda la figura 5, a pag. 37 del rapporto "*2025 Environmental sustainability report*" di Microsoft.

[190] Si veda la pagina "*Our 2025 Environmental Sustainability Report*" sul sito internet di Microsoft.

[191] Si veda pagina 110 del "*Eviromental Report 2025*" di Google, nonché la pagina "*Innovating across our operations and supply chain*".

[192] Si veda il "*2024 Sustainability Report*" di TSMC.

[193] "*When comparing the average electricity demand of a typical Google search (0.3 Wh of electricity) to OpenAI's ChatGPT (2.9 Wh per request), and considering 9 billion searches daily, this would require almost 10 TWh of additional electricity in a year*". International Energy Agency (IEA) (2024), Electricity 2024 - Analysis and forecast to 2026.

scientifici[194] e giornalistici (TIME)[195], nonché da report finanziari (Goldman Sachs)[196], secondo cui una singola richiesta rivolta a una chat di IA comporterebbe un consumo elettrico circa dieci volte superiore rispetto a una ricerca effettuata tramite un motore di ricerca tradizionale: rispettivamente 2,9 Wh contro 0,3 Wh. Tale dato, tuttavia, deve oggi essere considerato soprattutto come un ordine di grandezza storico, più che come un parametro generalmente valido. Google ha infatti pubblicato nel 2025 una metodologia più articolata per la misurazione dell'inferenza, stimando che una richiesta testuale rivolta a Gemini dagli utenti (il prompt testuale mediano) richieda circa 0,24 Wh, con emissioni pari a 0,03 gCO2e e un consumo idrico di 0,26 mL[197]. Ne deriva che il costo ambientale di una singola interazione non è fisso, ma varia sensibilmente in funzione del modello utilizzato, della lunghezza del prompt e della risposta, del tasso di utilizzo dell'hardware e dell'architettura complessiva del data center. Tuttavia, secondo l'AI Index 2026, il consumo idrico annuo associato alla sola inferenza di GPT-4o può superare il fabbisogno di acqua potabile di 12 milioni di persone, mostrando come l'impatto ambientale dell'IA non si esaurisca nella sola fase di training, ma si estenda in modo crescente anche all'uso quotidiano dei modelli su larga scala.

[194] De Vries, A. (2023). The growing energy footprint of artificial intelligence. *Joule*, *7*(10), 2191-2194.

[195] Chow, A. R. (2024). How AI is fueling a boom in data centers and energy demand. *Time Magazine, June 12, 2024*.

[196] "*On average, a ChatGPT query needs nearly 10 times as much electricity to process as a Google search";* cfr. Goldman Sachs (2024), AI is poised to drive 160% increase in data center power demand.

[197] Si veda: "*How much energy does Google's AI use? We did the math*" dal sito internet di Google.

Box 9 – Sustainable AI

L'intelligenza artificiale, per diventare parte di una transizione ecologica, deve affrontare **una doppia sfida ambientale**: ridurre l'**impronta energetica** (CO_2, carico di rete, approvvigionamento elettrico) e quella **idrica** (raffreddamento, produzione di semiconduttori, acqua "ultrapura"). Questo obiettivo richiede un approccio integrato, che combini innovazione tecnologica, regolazione mirata e responsabilità istituzionale.

Tecnologie abilitanti

- Chip efficienti (AI-specific): nuovi ASIC e GPU ottimizzate per LLM riducono i consumi per operazione.
- *Model compression e distillation*: tecniche per creare modelli più leggeri e meno energivori.
- Algoritmi a basso impatto: approcci come *sparse attention*, *low-rank adaptation*, *retrieval-augmented generation*.

Energie e infrastrutture

- *Data center carbon-free*: alimentazione da fonti rinnovabili.
- Raffreddamento sostenibile: uso di acqua piovana, sistemi di *immersion cooling*, localizzazione in ambienti freddi.
- Spostamento dell'inferenza *edge-side*: delegare parte del calcolo ai dispositivi finali, riducendo la dipendenza dai data center centrali.

Metriche e standard emergenti

- *Carbon footprint* per modello: misurazione standardizzata dell'energia consumata e delle emissioni per addestramento/inferenza.
- *AI Energy Labeling*: proposta di etichettatura energetica dei modelli AI (analoga a quella degli elettrodomestici).

Questi dati vanno letti alla luce di due dinamiche fondamentali. La prima è il **rallentamento della Legge di Moore**[198], che per decenni ha consentito aumenti di prestazioni con minori costi e consumi. Oggi, con la fine di quella traiettoria, l'efficienza per watt non cresce più abbastanza velocemente da compensare l'esplosione della domanda cognitiva. La seconda è il cosiddetto **paradosso di Jevons**[199]: a ogni miglioramento di efficienza tecnologica corrisponde,

[198] "La complessità di un microcircuito, misurata ad esempio tramite il numero di transistor per chip, raddoppia ogni 18 mesi (e quadruplica quindi ogni 3 anni)".

[199] Il Paradosso è stato enunciato nel 1865 dall'economista statunitense Stanley Jevons nell'ambito del libro *The Coal Question; An Inquiry Concerning the Progress of the Nation, and the Probable Exhaustion of Our Coal Mines* (edito da Macmillan).

paradossalmente, un aumento complessivo del consumo, dovuto alla crescita degli usi e alla diffusione delle applicazioni.

Tuttavia, parallelamente a questi impatti, l'intelligenza artificiale può contribuire in modo significativo a ridurre i consumi energetici e le emissioni complessive, se utilizzata per ottimizzare processi produttivi, logistici e gestionali. Secondo alcune stime, l'intelligenza artificiale potrebbe assumere un ruolo centrale nella transizione energetica[200]: la diffusione di applicazioni già oggi disponibili nei vari settori economici potrebbe infatti tradursi, entro il 2035, in una riduzione delle emissioni pari a circa 1.400 Mt di CO2[201]. In ambiti come l'industria manifatturiera[202], l'edilizia[203], l'agricoltura[204], la produzione[205] dell'energia elettrica e i trasporti[206], l'impiego di soluzioni basate su IA ha già dimostrato di poter migliorare l'efficienza, ridurre gli sprechi e ottimizzare l'uso delle risorse. Secondo alcune stime, l'intelligenza artificiale potrebbe contribuire a mitigare tra il 5% e il 10% delle emissioni globali di gas serra

[200] Alcuni studi stimano che l'IA possa contribuire a una riduzione di anidride carbonica tra 3,2 e 5,4 miliardi di tonnellate entro il 2035. Si veda: Stern, N., Romani, M., Pierfederici, R., Braun, M., Barraclough, D., Lingeswaran, S., ... & Niemann, N. (2025). Green and intelligent: the role of AI in the climate transition. *npj Climate Action*, *4*(1), 1-7 e. Gli Autori

[201] L'IEA, nel rapporto "*AI and climate change*", stima che: "*The adoption of existing AI applications in end-use sectors could lead to 1 400 Mt of CO_2 emissions reductions in 2035 in the Widespread Adoption Case*".

[202] L'impiego di un workflow standardizzato basato su IA può ottimizzare i consumi energetici nelle fabbriche. Si veda, al riguardo: Lee, D., & Lin, C. (2024). Universal artificial intelligence workflow for factory energy saving: Ten case studies. *Journal of Cleaner Production*, *468*, 143049.

[203] Ding, C., Ke, J., Levine, M., & Zhou, N. (2024). Potential of artificial intelligence in reducing energy and carbon emissions of commercial buildings at scale. *Nature Communications*, *15*(1), 5916.

[204] Applicazioni IA per l'agricoltura di precisione permettono una gestione mirata delle risorse idriche e fertilizzanti, contribuendo a un'agricoltura meno intensiva e più sostenibile. Si consulti, al riguardo: Mana, A. A., Allouhi, A., Hamrani, A., Rehman, S., El Jamaoui, I., & Jayachandran, K. (2024). Sustainable AI-based production agriculture: Exploring AI applications and implications in agricultural practices. *Smart Agricultural Technology*, *7*, 100416.

[205] Si stima che l'utilizzo di applicazioni di IA negli impianti di produzione energetica possa generare un incremento dell'efficienza energetica delle centrali tra il 3 e l'8%, nonché risparmi pari a 110 miliardi di dollari all'anno entro il 2035: "*The integration of today's AI applications in power plant operations and maintenance can yield potential cost savings of up to USD 110 billion annually worldwide to 2035"*. Per ulteriori dettagli, si consulti il seguente rapporto: International Energy Agency. (2025). *Energy and AI: World Energy Outlook Special Report*.

[206] Secondo l'Agenzia internazionale dell'energia, l'utilizzo dell'intelligenza artificiale potrebbe ridurre le emissioni globali del trasporto merci su strada di circa il 5%: "*AI-powered capacity utilization solutions have the potential to reduce global road freight emissions by approximately 5%*". Per ulteriori dettagli si consulti il seguente rapporto: International Energy Agency. (2025). *Energy and AI: World Energy Outlook Special Report.* Il World Economic Forum ritiene che l'impiego dell'IA possa ridurre le emissioni globali del trasporto merci di una percentuale ricompresa tra il 10–15 % del totale delle emissioni del settore. Per ulteriori dettagli, si consulti il rapporto: World Economic Forum. (2025). *Intelligent Transport, Greener Future: AI as a Catalyst to Decarbonize Global Logistics*.

entro il 2030[207]. In questa prospettiva, la sostenibilità dell'intelligenza artificiale non si misura soltanto in base ai consumi che essa genera, ma anche in relazione alla sua capacità di ridurre i consumi nei settori nei quali viene applicata.

In sintesi, la traiettoria attuale dell'intelligenza artificiale presenta un duplice volto: da un lato rischia di generare una nuova forma di insostenibilità ambientale, legata all'elevato consumo di risorse necessarie per alimentare la sua infrastruttura computazionale; dall'altro, offre strumenti potenti per migliorare l'efficienza energetica, ridurre gli sprechi e abilitare soluzioni sostenibili in settori chiave dell'economia. Una transizione verso un'IA realmente sostenibile richiederà non solo interventi tecnologici volti a ridurre l'impatto diretto della tecnologia, ma anche scelte politiche e istituzionali orientate a massimizzarne l'impatto positivo tramite una governance efficace e l'individuazione chiara delle priorità strategiche.

Definizione	Descrizione	Natura	Tipo di rischio
Consumo energetico	Addestramento e inferenza dei modelli richiedono grandi quantità di energia, con impatti crescenti sulle reti	Ambientale, infrastrutturale	Aumento emissioni, pressione sulle reti, ostacolo alla transizione energetica
Consumo idrico	Produzione di chip e raffreddamento dei server richiedono enormi volumi di acqua ultrapura	Ambientale, industriale	Stress idrico, conflitti tra usi civili e industriali
Rallentamento Legge di Moore	Il rallentamento dell'efficienza computazionale rende i consumi dell'IA meno gestibili a parità di capacità	Tecnologica, strutturale	Aumento insostenibile dei consumi per unità di calcolo
Paradosso di Jevons	L'efficienza aumenta, ma il consumo complessivo cresce a causa dell'adozione massiva della tecnologia	Economica, sistemica	Crescita dei consumi aggregati, rischio di sostenibilità apparente

Tabella 10 – Problematiche ambientali dell'IA

[207] Si consulti a tale riguardo, l'articolo "AI and energy: Will AI help reduce emissions or increase power demand? Here's what to know", pubblicato dal World Economic Forum.

5.5 Questioni relative ai diritti

L'intelligenza artificiale esercita un impatto crescente su una vasta gamma di diritti individuali e collettivi, ridefinendo in profondità il rapporto tra cittadini, tecnologie e poteri pubblici[208]. Alcuni di questi diritti sono già stati analizzati nei capitoli precedenti di questo rapporto: è il caso, ad esempio, della tutela della proprietà intellettuale, che risulta compromessa nei casi di utilizzo non autorizzato di contenuti per il training dei modelli, o delle questioni di cybersicurezza, affrontate in relazione agli incidenti, alle vulnerabilità e agli abusi di sistema.

Altri diritti fondamentali – come quelli connessi alla salute, alla tutela della privacy, al lavoro e all'accesso alla giustizia – sono oggetto di un dibattito internazionale rilevante e meritano attenzione, ma non saranno oggetto di approfondimento in questa sede, poiché esulano dagli obiettivi specifici di questo rapporto.

Ci concentreremo invece su una selezione mirata di ambiti, in cui l'IA interseca elementi civili, politici e culturali che riguardano direttamente il funzionamento delle democrazie[209], l'integrità dell'informazione e la libertà individuale, e che pertanto risultano particolarmente sensibili per l'equilibrio tra innovazione e coesione sociale. In particolare, analizzeremo gli effetti dell'IA su:

- libero arbitrio individuale e comunicazione;
- non discriminazione, con riferimento a fattori quali reddito, genere, etnia, religione e disabilità;
- tutela delle minoranze e dei minori, in contesti esposti a *bias* o automatismi discriminatori;
- libertà di informazione e pluralismo informativo, sia dal lato attivo (produzione di informazioni e notizie) sia da quello passivo (fruizione di informazioni e notizie);
- dibattito pubblico e partecipazione democratica, inclusi i processi elettorali e la formazione dell'opinione pubblica.

Questi temi, al crocevia tra tecnologia e diritti fondamentali, pongono sfide urgenti di governance, trasparenza e responsabilità pubblica, richiedendo nuovi strumenti al fine di

[208] Adam, M., & Hockuard, C. (2023). Artificial intelligence, democracy and elections. European Parliament Briefing.

[209] UNESCO (2024), Artificial intelligence and democracy.

garantire che l'intelligenza artificiale operi in coerenza con i principi costituzionali e democratici che fondano le nostre istituzioni[210].

Come illustrato in precedenza, l'intelligenza artificiale sta trasformando profondamente il rapporto tra individui, informazione e potere cognitivo. Infatti, un sistema supercognitivo, come quelli basati su modelli linguistici di grandi dimensioni, non si limita a fornire informazioni, ma interviene attivamente nella costruzione della conoscenza e dell'opinione. L'IA propone risposte immediate, crea narrazioni coerenti, struttura il sapere secondo un ordine e una gerarchia determinati algoritmicamente (oltre che per mezzo di filtri). In questo modo, non solo trasmette contenuti, ma determina i frame cognitivi attraverso cui i contenuti vengono percepiti. In un orizzonte temporale più lungo, l'IA è idonea a influenzare la struttura stessa dei *pattern* cognitivi, andando a incidere su capacità matematiche, abilità linguistiche, attitudini logico–deduttive, ambiti riflessivi e razionali.

In maniera più sottile, l'affermarsi di sistemi da cui tendiamo sempre più a dipendere per le nostre funzioni cognitive ha significative implicazioni sulla "*capacità di scegliere liberamente, nell'operare e nel giudicare*"[211]. Se i pensieri e le decisioni degli individui sono orientati da sistemi supercognitivi, si pone un interrogativo profondo: possiamo ancora dire di essere pienamente liberi nelle nostre scelte? O stiamo progressivamente delegando, in modo inconsapevole, una parte crescente della nostra autonomia a sistemi intelligenti che ci suggeriscono cosa sapere, cosa pensare, cosa dire, cosa fare?

Non si tratta di censura esplicita o di sorveglianza autoritaria, ma di una forma sottile e pervasiva di modellazione dell'ambiente cognitivo, in cui l'IA agisce da intermediario dominante tra gli individui e il mondo dell'informazione, della cultura e del sapere. La libertà, in questo scenario, non viene negata ma riformulata entro limiti tecnici e algoritmici, peraltro definiti dinamicamente da società private.

L'IA, pertanto, non è solo un mezzo di trasmissione: è un sistema cognitivo autonomo, capace di aggregare, sintetizzare, valutare e produrre sapere. Nella misura in cui le persone delegano all'IA il compito di informarsi, comprendere, argomentare, scegliere le parole o le opinioni da

[210] EPTA (2024) Artificial Intelligence and Democracy. Report. October 2024, Oslo.

[211] Si veda la voce "Libero arbitrio", in Enciclopedia Italiana, Istituto dell'Enciclopedia Italiana.

esprimere, e, più in generale, svolgere complesse funzioni cognitive, si verifica una graduale disintermediazione del confronto umano. I soggetti non comunicano più direttamente tra loro per elaborare punti di vista, ma si rivolgono a un'interfaccia che restituisce loro una sintesi – ordinata, efficiente, coerente – dell'intero dibattito possibile.

A completare il quadro si aggiunge un ulteriore elemento: la capacità persuasiva delle IA generative. Studi recenti dimostrano che questi sistemi non si limitano a rispondere, ma tendono ad allinearsi progressivamente alle preferenze dell'utente, offrendo risposte che confermano le sue aspettative e rafforzano le sue convinzioni. È il meccanismo alla base della cosiddetta *hyperpersuasion*, ovvero quella forma di "iperpersuasione" che porta l'interlocutore umano a modificare le proprie opinioni man mano che l'interazione prosegue, proprio perché l'IA sembra "capirlo" e dargli ragione. In questo senso, uno studio condotto da alcuni ricercatori dell'Università di Zurigo ha mostrato che i modelli linguistici sono più efficaci di un essere umano[212] – anche addestrato, ad esempio, in tecniche di programmazione neurolinguistica – nell'indurre cambiamenti di opinione durante il confronto online[213].

Questo risultato trova ulteriore conferma in una recente ricerca pubblicata su una rivista scientifica, la quale ha evidenziato che i modelli generativi possono risultare più convincenti degli esseri umani nei dibattiti online, soprattutto quando sono in grado di adattare le proprie argomentazioni sulla base dei dati demografici dell'interlocutore[214]. Superando i contendenti umani nel 64% dei casi, i *chatbot* (vedi Box 2) hanno dimostrato una capacità di "ottimizzazione empatica" che solleva interrogativi sul potenziale di influenza esercitabile in contesti politici, commerciali o sociali. Non si tratta dunque solo di generare contenuti, ma di orientare attivamente opinioni e comportamenti con un'efficacia superiore alla media umana. Di qui, l'urgenza di una riflessione sulle responsabilità, i limiti d'uso e le garanzie di trasparenza che dovrebbero accompagnare questi strumenti, soprattutto in ambiti sensibili per la democrazia.

[212] Raggiungendo tassi di persuasione da tre a sei volte superiori rispetto al valore di riferimento umano.

[213] Per ulteriori dettagli, si consulti l'articolo: "Can AI Change Your View? Evidence from a Large-Scale Online Field Experiment" e l'articolo di Wired "L'intelligenza artificiale ha ingannato gli utenti di Reddit".

[214] Salvi, F., Horta Ribeiro, M., Gallotti, R., West R. (2025). On the conversational persuasiveness of GPT-4, Nature Human Behaviour.

Dunque, così come l'IA rischia di sostituirsi alle funzioni cognitive individuali, guidando l'attenzione, la memoria e le attività mentali del singolo, allo stesso modo può intercettare e assorbire le funzioni cognitive collettive, trasformandosi in una superintelligenza tecnica che rimpiazza, di fatto, la superintelligenza sociale. Quella che un tempo era una costruzione comunitaria del sapere – frutto di discussione, dialettica, dissenso, contaminazione – viene in parte rimpiazzata da un apparato cognitivo centralizzato e tecnico, che fornisce risposte invece di stimolare domande, e che semplifica i conflitti invece di renderli occasione di riflessione collettiva.

Il **principio di non discriminazione** rappresenta uno dei pilastri dello Stato di diritto e delle democrazie costituzionali. È garantito dalla Costituzione italiana (art. 3), che sancisce l'eguaglianza formale e sostanziale dei cittadini, e trova fondamento anche nella Carta dei diritti fondamentali dell'Unione europea (art. 21) che vieta ogni forma di discriminazione basata su razza, sesso, lingua, religione, opinioni politiche, origine nazionale o sociale, disabilità o status economico.

In questo ambito, l'introduzione dell'intelligenza artificiale in numerosi ambiti della vita pubblica e privata espone questo principio a nuove forme di vulnerabilità, spesso meno visibili ma strutturalmente radicate nei dati, nei modelli e nei contesti d'uso. Molti sistemi di IA – soprattutto quelli di tipo predittivo o decisionale – sono addestrati su dataset storici, parziali o sbilanciati, che riflettono le disuguaglianze, i pregiudizi e le distorsioni presenti nella società. Quando non corretti, questi *bias* si riproducono e si amplificano nei modelli, generando trattamenti differenziati a danno di categorie protette o svantaggiate (per una lettura dei bias algoritmici come fattore di possibili diseguaglianze e come questione che investe il rapporto tra tecnica, diritto e garanzie ordinamentali, cfr. Giovanna de Minico, *Rapporto Comitato IA*, cap. 8)[215].

[215] Le applicazioni più critiche si riscontrano in settori come: i) Polizia predittiva e sorveglianza: studi su software usati negli Stati Uniti hanno mostrato che sistemi di previsione del crimine tendono a sovrastimare la pericolosità di soggetti appartenenti a minoranze etniche, perpetuando forme di profilazione razziale; 2) Sanità: algoritmi di triage o di previsione dell'aderenza ai trattamenti hanno in alcuni casi sottostimato le esigenze di pazienti afroamericani rispetto ai pazienti bianchi a parità di condizioni cliniche, in ragione di una proxy economica mal calibrata; 3) Assicurazioni e credito: l'uso dell'IA nella valutazione del rischio può discriminare indirettamente per condizione sociale, codice postale, accesso ai servizi digitali, rendendo invisibili o penalizzando soggetti già marginalizzati; 4) Lavoro e reclutamento: sistemi di screening automatico dei CV e di selezione del personale possono riprodurre schemi di esclusione legati al genere, all'età o al background culturale, specie se addestrati su storici aziendali non equilibrati.

Infatti, i sistemi di IA possono tendere a riflettere la visione del mondo dominante nei dati di addestramento, riducendo la rappresentazione di esperienze, codici linguistici o prospettive alternative. Ne risultano modelli che possono parlare a "maggioranze implicite", trascurando o distorcendo ciò che non rientra nel canone prevalente. In entrambi i casi, il rischio non è solo di discriminazione, ma di esclusione epistemica: una sottrazione di visibilità, voce e riconoscimento che compromette l'inclusività cognitiva del sistema. E se l'IA diventa la nuova infrastruttura della conoscenza, chi non vi è rappresentato corre il rischio di non esistere affatto.

In questo senso, i modelli linguistici possono produrre testi, immagini o suggerimenti che riflettono stereotipi di genere, razziali o culturali, o che sottorappresentano esperienze, lingue e conoscenze di comunità minoritarie. La discriminazione algoritmica può essere diretta o indiretta, esplicita o emergente, ma in ogni caso mette in discussione l'equità dei processi decisionali automatizzati, specialmente quando questi vengono adottati in contesti pubblici o semi–pubblici (giustizia, welfare, sanità, istruzione).

In assenza di trasparenza, di meccanismi di auditing e di strumenti di contestazione, l'uso dell'IA può istituzionalizzare forme di ingiustizia sistemica, difficili da rilevare e correggere a posteriori. Il rischio non è solo quello di discriminare singoli individui, ma anche di produrre nuove gerarchie dell'accesso e del riconoscimento, rendendo invisibili interi gruppi sociali o rafforzando posizioni di potere (non solo economico).

Tra i soggetti più esposti agli effetti distorsivi dell'intelligenza artificiale vi sono i **minori**, ossia coloro che già nella società analogica occupano posizioni di vulnerabilità (per un approfondimento sui rischi dell'IA per i giovani, sui profili educativi e sulla protezione dei minori nell'ecosistema digitale, cfr. Mauro Giusto, *Rapporto Comitato IA*, cap. 7). In questo caso, la questione è duplice: da un lato, i modelli generativi possono esporli a contenuti inappropriati o manipolativi, senza garanzie sufficienti di protezione; dall'altro, l'uso dell'IA in contesti educativi rischia di modellare i percorsi cognitivi e formativi su basi standardizzate, non adatte a promuovere sviluppo critico e pluralismo culturale. Queste ultime preoccupazioni sono corroborate dai risultati dei primi studi sull'impatto dell'IA sui sistemi cognitivi, così come illustrati nel paragrafo 5.1.

Il diritto a essere informati – nella sua doppia valenza di accesso attivo e passivo all'informazione – è un elemento strutturale delle democrazie costituzionali. Non solo perché garantisce la libertà individuale di conoscere, ma perché costituisce il presupposto materiale della partecipazione pubblica, del dibattito aperto e della legittimità delle decisioni collettive. L'intelligenza artificiale, assumendo un ruolo crescente come infrastruttura cognitiva, sta modificando radicalmente anche i meccanismi con cui l'**informazione** viene prodotta, distribuita e recepita. L'IA non si limita a selezionare contenuti: li genera, li ordina, li personalizza e li filtra, ristrutturando così il campo informativo secondo logiche opache e non pluralistiche.

In questo quadro, la **disinformazione** automatizzata rappresenta una minaccia emergente e trasversale. I modelli generativi di testo, immagine, audio e video permettono oggi di produrre a costo quasi nullo contenuti falsi, verosimili e personalizzati, con una velocità e una scala senza precedenti. Poiché questi sistemi superano ormai ampiamente i test di Turing (cfr. Box 1) per coerenza e naturalezza, la probabilità che utenti comuni prendano per autentiche informazioni completamente false è altissima, soprattutto in contesti emotivamente carichi o poco verificabili (per un approfondimento sui rischi sistemici connessi alla manipolazione informativa, alla degradazione discorsiva e ai meccanismi di amplificazione algoritmica nell'ecosistema digitale, cfr. Giovanni Boccia Altieri, *Rapporto Comitato IA*, cap. 5). La disinformazione diventa così più economica, più persuasiva, più difficile da tracciare[216].

Un ulteriore profilo critico, accentuato dall'adozione di sistemi generativi come infrastruttura di accesso all'informazione, riguarda il rischio che la fluidità linguistica dell'output ne accresca la credibilità percepita, anche quando il contenuto non è ancorato a fonti verificabili. Tale dinamica si innesta su criticità già richiamate nel § 5.2 con riferimento alle allucinazioni (risposte errate, inventate o incoerenti, spesso fornite con tono assertivo e verosimile), rendendo più difficile per l'utente distinguere tra plausibilità e verità. Parallelamente, il **pluralismo dell'informazione** può venire compromesso su almeno tre fronti. Primo, la personalizzazione algoritmica tende a chiudere gli individui in bolle cognitive, riducendo

[216] Per il secondo anno consecutivo, il report "The Global Risks Report 2025, 20th Edition", edito dal World Economic Forum, mette al primo posto – tra i rischi globali classificati in base alla gravità nel breve e lungo termine – la misinformazione e la disinformazione.

l'esposizione a punti di vista dissonanti. Secondo, i modelli sintetizzano la conoscenza in forme apparentemente neutre, ma fondate su scelte implicite nei dati e nei pesi dei modelli. Terzo, il controllo altamente concentrato delle piattaforme IA pone problemi sistemici di potere: chi decide cosa viene mostrato, con quale struttura e priorità?

Box 10 – L'uso dell'IA generativa come infrastruttura di ricerca

L'uso dell'IA generativa nella ricerca online segna un cambiamento strutturale nel modo di accedere all'informazione, spostando il baricentro dal tradizionale motore di ricerca, fondato su elenchi di collegamenti, a un vero e proprio motore di risposta, in cui l'utente riceve una sintesi immediata e contestualizzata direttamente nella pagina dei risultati. In questo paradigma, l'informazione non è più principalmente "cercata" attraverso la navigazione tra siti, ma viene elaborata, riassunta e presentata in forma discorsiva all'interno dell'interfaccia stessa.

Ad esempio, se si analizza il caso di Google, si osserva come funzionalità quali AI Overviews e AI Mode collochino la risposta generativa in posizione di massima evidenza nella SERP, sopra o prima dei risultati organici, integrandola in un'esperienza conversazionale attiva di default, senza una scelta preventiva esplicita da parte dell'utente.

Dal punto di vista tecnico, questo modello si basa su una pipeline che combina la classificazione della query, il recupero e l'ancoraggio delle fonti tramite i meccanismi di search tradizionale e i knowledge graph, e la sintesi operata da modelli linguistici di grandi dimensioni in modalità RAG, con l'applicazione di filtri di sicurezza e qualità prima dell'esposizione dell'output; il funzionamento risulta inoltre sensibile al contesto d'uso e alla cronologia delle ricerche, utilizzata per migliorare le prestazioni del sistema.

Lo spostamento dell'interazione verso la risposta "in pagina" ha effetti rilevanti sull'ecosistema informativo: riduce gli incentivi al clic e il traffico verso i siti di origine, accentuando il fenomeno dello zero–click e incidendo in modo particolare sugli editori medio–piccoli e sulle fonti minoritarie, con possibili ricadute sulla sostenibilità economica e sul pluralismo (per una ricostruzione sul tema della comunicazione, cfr. *Rapporto Comitato IA*, cap. 3.3, di Andrea Imperiali). Al tempo stesso, la priorità visiva delle risposte generative, la selettività delle fonti e l'integrazione verticale tra modello, servizio di ricerca e interfaccia rafforzano il potere delle grandi piattaforme, amplificando esternalità di rete e dinamiche di lock–in (coerentemente con quanto illustrato nei §§ 4.2 e 4.4).

In questo quadro, la ricerca generativa si configura come una nuova infrastruttura centrale di accesso all'informazione, rispetto alla quale i quadri regolatori europei – DSA, AI Act e DMA – assumono un ruolo cruciale per garantire trasparenza,

responsabilità e condizioni di equilibrio competitivo lungo l'intera catena del valore informativo.

Questi processi possono minare la qualità e la pluralità del **dibattito pubblico**, che diventa più fragile, frammentato, manipolabile. Se l'opinione pubblica è alimentata da contenuti opachi o simulati, se la voce degli attori sociali è mediata da algoritmi non trasparenti, la capacità collettiva di deliberare, dissentire e partecipare consapevolmente si riduce (per una lettura costituzionale del rapporto tra IA, libertà di informazione, pluralismo e ridefinizione del discorso pubblico nell'ecosistema digitale, cfr. Andrea Simoncini, *Rapporto Comitato IA*, cap. 4).

Inoltre, se il linguaggio politico e istituzionale viene progressivamente assorbito dai sistemi generativi, si rischia una depoliticizzazione strisciante del discorso pubblico: l'apparente neutralità dell'IA sostituisce la legittimità del conflitto democratico, con esiti negativi sul **confronto elettorale**.

Definizione	Descrizione	Natura	Tipo di rischio
Libero arbitrio	L'IA si può sostituire nei processi cognitivi individuale e orienta decisioni e preferenze tramite suggerimenti e strutture informative predefinite	Cognitiva, psicologica	Erosione dell'autonomia, dipendenza cognitiva
Libertà di comunicazione	L'IA può sostituire il confronto umano con risposte centralizzate, riducendo la costruzione sociale del sapere	Sociale, politica	Delega cognitiva generalizzata, perdita del confronto pluralistico
Non discriminazione e tutela minoranze	L'IA può amplificare distorsioni nei dati e trattare in modo iniquo gruppi svantaggiati (per genere, etnia, disabilità, ecc.)	Giuridica, sociale	Discriminazione automatizzata, esclusione sistemica
Tutela minori	I minori possono essere esposti a contenuti inappropriati o abusare di modelli cognitivi proprio in periodo di formazione culturale e sviluppo psicologico e cognitivo	Educativa, cognitiva	Disallineamento formativo, danni cognitivi e relazionali
Libertà di informazione e pluralismo	L'IA può personalizzare e filtrare i contenuti, riducendo l'accesso a fonti plurali e favorendo la produzione e diffusione di disinformazione	Informativa, sistemica	Bolle informative, omogeneizzazione cognitiva, diffusione massiva di disinformazione
Dibattito pubblico e partecipazione democratica	L'IA può alterare il discorso pubblico, simulando consenso e riducendo la legittimità del dissenso e del confronto	Democratica, politica	Depoliticizzazione, manipolazione del consenso, crisi di legittimazione deliberativa

Tabella 11 – Problematiche dell'IA relative ai diritti

6 Considerazioni conclusive

L'evoluzione dell'intelligenza artificiale, come ricostruito nel Capitolo 2, si caratterizza per una chiara traiettoria *path-dependent*: una lunga fase iniziale di ricerca accademica e sperimentazioni, alternata a momenti di entusiasmo e stagnazione, ha lasciato il posto – a partire dal 2010 circa – a un'accelerazione netta e irreversibile. Questa svolta è stata resa possibile dalla disponibilità di immense quantità di dati digitali e dall'accresciuta capacità computazionale necessaria per processarli, addestrare reti neurali complesse e far evolvere l'IA da sistema statico a infrastruttura dinamica di apprendimento.

Nel corso di questa transizione, tuttavia, l'intelligenza artificiale si è trasformata in una tecnologia pervasiva e di base, una vera *General Purpose Technology* (GPT), destinata a ridisegnare l'intero sistema socioeconomico. Come sottolineato anche dalla letteratura più recente[217], gli effetti di questa rivoluzione travalicano gli ambiti della produttività e dell'automazione, toccando strutturalmente la conoscenza, la *governance*, l'ambiente e la democrazia. Eppure, nonostante la natura sistemica dell'IA, il suo sviluppo si è rapidamente concentrato nelle mani di un numero ristretto di grandi piattaforme private, dando luogo a un processo di privatizzazione e mercificazione di un sistema supercognitivo.

Come descritto nel Capitolo 3, ciò è anche il risultato di precise caratteristiche tecnico-economiche della moderna IA: i costi di addestramento sono elevati, irreversibili (*sunk*) e in rapida crescita; i costi fissi legati all'infrastruttura (data center, chip, energia) sono consistenti e tendenzialmente crescenti. Ne deriva una struttura industriale dominata da rendimenti di scala crescenti dal lato dell'offerta, come analizzato nel Capitolo 4, che tende a favorire i pochi attori globali con capacità finanziaria di sostenere tali investimenti.

Parallelamente, i mercati dell'IA si configurano come mercati multiversante (*multi-sided*), dove gli operatori agiscono da piattaforme e sfruttano esternalità di rete dirette (numero di utenti) e indirette (qualità dei dati e dei servizi connessi), integrando progressivamente i propri servizi

[217] Cfr., ad esempio, Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., ... & Viale, R. (2024). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS Nexus*, 3(6), pag. 191.

sia in senso verticale (dall'hardware all'applicazione) sia in senso orizzontale e trasversale (tra ambiti diversi: linguaggio, visione, programmazione, ecc.).

Questo contesto pone interrogativi cruciali, come discusso nel Capitolo 5. Si tratta di questioni tanto teoriche (il libero arbitrio, i limiti della delega tecnologica, il rapporto tra umano e macchina, l'evoluzione dei sistemi cognitivi e culturali individuali e collettivi), quanto pratiche (tecniche, economiche, ambientali e giuridiche): dalla tutela dei diritti individuali alla concorrenza, dalla sostenibilità energetica all'informazione e al dibattito pubblico.

Proprio per affrontare queste sfide, l'Unione Europea ha adottato l'AI Act[218], un quadro regolatorio che punta a bilanciare innovazione e sicurezza. Il suo approccio è di tipo "*light but effective*": evitare di distorcere gli incentivi economici che alimentano il progresso, ma offrire al contempo prime risposte normative a rischi concreti ed emergenti, in particolare per le applicazioni ad alto impatto.

Parallelamente, il Digital Services Act (DSA) ha ridefinito l'architettura istituzionale della *governance* del digitale nell'Unione Europea, prevedendo che ogni Stato membro designi un Digital Services Coordinator (DSC), responsabile dell'attuazione del Regolamento e della cooperazione tra le autorità competenti. In Italia, tale funzione è stata attribuita ad AGCOM. E oggi, alla luce della crescente compenetrazione tra intelligenza artificiale e servizi digitali, questo ruolo assume una rilevanza ancor più strategica, dato che le piattaforme soggette al DSA integrano sistemi di IA generativa e non, motori di raccomandazione, algoritmi decisionali e interfacce intelligenti che incidono sull'accesso all'informazione, sulla visibilità dei contenuti e sulla fruizione dei servizi.

L'intelligenza artificiale, pur presentando una struttura tecnica sostanzialmente unitaria – in quanto i modelli sono definiti "a monte" sotto il profilo architetturale e funzionale – non produce effetti uniformi nei diversi contesti di utilizzo. Se l'AI Act fornisce una cornice europea comune, prevalentemente orientata al prodotto e ai requisiti di sicurezza, affidabilità e conformità, nella fase di impiego l'IA conosce una declinazione inevitabilmente nazionale, adattandosi alle specificità linguistiche, culturali, informative e sociali delle platee di riferimento. È proprio questa dimensione contestuale a rendere centrale il ruolo delle autorità nazionali, soprattutto

[218] Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence.

quando i sistemi di IA incidono su diritti costituzionalmente rilevanti, quali il pluralismo dell'informazione. Quando l'IA interviene nella selezione, organizzazione o raccomandazione dei contenuti informativi, essa tende infatti a riflettere – e talora ad amplificare – le caratteristiche degli ecosistemi mediatici nazionali, rendendo necessaria una lettura dell'IA non solo come tecnologia neutra, ma come infrastruttura che si modella sulle comunità democratiche cui è destinata. In tale prospettiva, la disciplina europea dell'IA si configura come condizione necessaria ma non sufficiente, richiedendo una declinazione nazionale sul piano degli usi concreti e delle pratiche applicative (sul punto, cfr. Andrea Renda, *Rapporto Comitato IA*, cap. 9, sul ruolo delle autorità di regolazione settoriale nei diversi contesti nazionali ai fini dell'applicazione dell'IA nei settori di rispettiva competenza).

Il **presente rapporto** nasce in questo contesto e intende offrire un primo contributo conoscitivo volto a esplorare l'evoluzione dell'IA in relazione alla missione istituzionale di AGCOM quale DSC, in vista di una futura e più approfondita riflessione sui punti di intersezione tra AI Act e DSA. Tale funzione si esercita in raccordo con le Autorità designate per l'attuazione dell'AI Act, nell'ambito di un modello di cooperazione in cui AGCOM è chiamata a valorizzare la propria capacità istituzionale di analizzare i fenomeni in chiave trasversale e sistemica, propria del settore delle comunicazioni, offrendo così un contributo alla definizione di una governance multilivello dell'IA coerente con la crescente centralità del DSC nello spazio digitale europeo. La condivisione delle competenze e delle prospettive tra autorità è tanto più necessaria in un contesto in cui molte questioni rimangono aperte e richiedono un monitoraggio continuo, anche alla luce delle evidenze – emerse nel Capitolo 3 – che segnalano l'avvicinarsi di forme sempre più avanzate di intelligenza artificiale generale (AGI - *Artificial General Intelligence*). In tale scenario, si rafforza l'esigenza di un approccio pubblico proattivo e di una capacità regolatoria più incisiva, come sottolineato dalla comunità scientifica[219], da istituzioni globali come l'ONU[220] o il World Economic Forum[221], nonché da personalità[222] come Geoffrey Hinton – premio Nobel

[219] Baronchelli, A. (2024). Shaping new norms for AI. *Philosophical Transactions of the Royal Society B*, *379*(1897), 20230028.

[220] United Nations, AI Advisory Body, Governing AI for Humanity, settembre 2024.

[221] World Economic Forum, AI in Action: Beyond Experimentation to Transform Industry, gennaio 2025.

[222] La necessità di un intervento a livello regolamentare inizia a essere richiesta anche dalla stessa industria. Poche settimane orsono Dario Amodei, CEO della società Anthropic che ha sviluppato il sistema di IA denominato *Claude*, ha pubblicato un lungo appello affinché si investa in procedure finalizzate all'interpretabilità del funzionamento dei modelli di intelligenza

per la Fisica e vincitore del premio Turing – che ha recentemente richiamato la necessità di una regolazione pubblica più forte a fronte della crescente potenza dei sistemi di IA[223].

In questa prospettiva, la **prima parte** dell'analisi – curata dall'Ufficio Intelligenza Artificiale – svolge una funzione fondativa e orientativa, offrendo una base comune di conoscenza tecnica, economica e istituzionale. Essa ricostruisce il percorso storico dell'IA, ne analizza il funzionamento tecnico e infrastrutturale, ne mette in luce le caratteristiche economiche e di mercato e individua le principali tensioni che essa solleva sul piano dei diritti, della concorrenza, della sostenibilità e della qualità del dibattito pubblico.

Su questa base si innestano i contributi del Comitato sull'Intelligenza Artificiale, raccolti nella **seconda parte**, che ampliano e approfondiscono il quadro ricostruttivo attraverso una lettura interdisciplinare, giuridica, regolamentare e sociologica delle implicazioni dell'IA in rapporto ai mercati, ai diritti fondamentali e agli ambiti di intervento dell'AGCOM. La seconda parte non si collega così alla prima in termini meramente aggiuntivi, ma come suo naturale sviluppo applicativo: dalle categorie ricostruttive generali si passa infatti all'analisi dei nodi critici che l'IA pone nei diversi settori e nelle materie di competenza dell'Autorità.

⁂

artificiale generativa. In questo contesto ha tra l'altro affermato che: "*governments can use light-touch rules to encourage the development of interpretability research and its application to addressing problems with frontier AI models*" Dario Amodei, The Urgency of Interpretability, aprile 2025).

[223] In *Godfather of AI' shortens odds of the technology wiping out humanity over next 30 years*, pubblicato il 28 dicembre del 2024, Hinton ha affemato: "*My worry is that the invisible hand is not going to keep us safe. So just leaving it to the profit motive of large companies is not going to be sufficient to make sure they develop it safely. The only thing that can force those big companies to do more research on safety is government regulation*".

7 Bibliografia

Adam, M., & Hockuard, C. (2023). Artificial intelligence, democracy and elections. European Parliament Briefing.

Amodei, D. (2024). Machines of Loving Grace: How AI Could Transform the World for the Better.

Amodei, D (2026). The Adolescence of Technology, Confronting and Overcoming the Risks of Powerful AI.

Amodei, D. (2025). The Urgency of Interpretability.

Anson Ho, Tamay Besiroglu, Ege Erdil, David Owen, Robi Rahman, Zifan Carl Guo, David Atkinson, Neil Thompson, and Jaime Sevilla. Algorithmic progress in language models. *ArXiv*, 2024.

Anthropic technical report (2025), System Card: Claude Opus 4 & Claude Sonnet 4.

Aresu, A. (2024). *Geopolitica dell'intelligenza artificiale*. Feltrinelli Editore.

Arrow, K. J. (1972). Economic welfare and the allocation of resources for invention. *Macmillan Education UK.*

Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, and National Research Council Publication, *Language and Machines Computers in Translation and Linguistic*, Publication 1416, 1966.

Baronchelli, A. (2024). Shaping new norms for AI. *Philosophical Transactions of the Royal Society B*, *379*(1897), 20230028.

Bengio, Y. (Chair). (2026). *International AI Safety Report 2026*. UK Department for Science, Innovation and Technology, on behalf of the international Expert Advisory Panel.

Biever, C. (2023). ChatGPT broke the Turing test–the race is on for new ways to assess AI. *Nature*, *619*(7971), 686–689.

Boden, M. A. (2008). *Mind as machine: A history of cognitive science*. Oxford University Press.

Bryson, A. E., Ho, (1969), *Applied optimal control*, Routledge, 2018.

Buchanan, B. G. (2005). A (very) brief history of artificial intelligence, *AI Magazine*, *26*(4), 53–53.

Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., ... & Viale, R. (2024). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS Nexus*, 3(6), pgae191.

Cellini, P., Ibarra, M., (2024), *AI Impact*, Luiss University Press.

Chander, B., John, C., Warrier, L., & Gopalakrishnan, K. (2025). Toward trustworthy artificial intelligence (TAI) in the context of explainability and robustness. *ACM Computing Surveys*, *57*(6), 1–49.

Coppin, B. (2004). *Artificial intelligence illuminated*. Jones & Bartlett Learning.

Cottier, B., Rahman, R., Fattorini, L., Maslej, N., Besiroglu, T., & Owen, D. (2024). The rising costs of training frontier AI models. *arXiv preprint arXiv:2405.21015*.

De Vries, A. (2023). The growing energy footprint of artificial intelligence. *Joule*, *7*(10), 2191–2194.

Ding, C., Ke, J., Levine, M., & Zhou, N. (2024). Potential of artificial intelligence in reducing energy and carbon emissions of commercial buildings at scale. *Nature Communications*, *15*(1), 5916.

Eeckhout, L. (2017). Is moore's law slowing down? what's next?. *IEEE Micro*, *37*(04), 4–5.

Electric Power Research Institute – EPRI (2024). Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption.

European Data Protection Supervisor – EDPS. (2023). TechDispatch #2/2023: Explainable Artificial Intelligence.

European Parliamentary Technology Assessment – EPTA (2024). Artificial Intelligence and Democracy.

Farrell, J., & Klemperer, P. (2007). Coordination and lock–in: Competition with switching costs and network effects. *Handbook of industrial organization*, *3*, 1967–2072.

Federal Trade Commission, FTC. (2025). *Partnerships between cloud service providers and AI developers*. FTC staff report on AI partnerships & investments.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Helpman, E., & Trajtenberg, M. (1998). *A time to sow and a time to reap: Growth based on general purpose technologies*. In E. Helpman (Ed.), *General Purpose Technologies and Economic Growth*. MIT Press.

Ho, A., Besiroglu, T., Erdil, E., Owen, D., Rahman, R., Guo, Z. C., ... & Sevilla, J. (2024). Algorithmic progress in language models. *Advances in Neural Information Processing Systems*, *37*, 58245–58283.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... & Sifre, L. (2022). Training compute–optimal large language models. *arXiv preprint arXiv:2203.15556*.

International Energy Agency – IEA. (2024). Electricity 2024 – Analysis and forecast to 2026.

International Energy Agency – IEA. (2025). Energy and AI – World Energy Outlook Special Report.

International Energy Agency – IEA. (2025). AI and climate change.

International Energy Agency. (2024). Global critical minerals outlook 2024.

Jevons, W. S. (1866). *The coal question; an inquiry concerning the progress of the nation and the probable exhaustion of our coal–mines*. Macmillan.

Jones, C. R., & Bergen, B. K. (2024). People cannot distinguish GPT–4 from a human in a Turing test. *arXiv preprint arXiv:2405.08007*.

Kautz, H. (2022). "The third AI summer: AAAI Robert S. Engelmore memorial lecture", *AI magazine*, 43(1), 105–125.

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.–H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). *Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task*. *arXiv*.

Lee, D., & Lin, C. (2024). Universal artificial intelligence workflow for factory energy saving: Ten case studies. *Journal of Cleaner Production*, *468*, 143049.

Lighthill Report (1973). Science Research Council (SRC).

Mana, A. A., Allouhi, A., Hamrani, A., Rehman, S., El Jamaoui, I., & Jayachandran, K. (2024). Sustainable AI–based production agriculture: Exploring AI applications and implications in agricultural practices. *Smart Agricultural Technology*, *7*, 100416.

Massenkoff, M., & McCrory, P., (2026). Labor market impacts of AI: A new measure and early evidence. Anthropic.

McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E. (1955) A proposal for the Dartmouth summer research project on artificial intelligence.

McCarthy, John. (2007). What is artificial intelligence.

McKinsey Global Institute (2023). *The economic potential of generative AI: The next productivity frontier.*

Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, *121*(9).

Minsky, M. L., & Papert, S. A. (1988). *Perceptrons: expanded edition*.

Minsky, M., & Papert, S. A. (1972). Artificial intelligence progress report.

Muthukrishnan, N., Maleki, F., Ovens, K., Reinhold, C., Forghani, B., & Forghani, R. (2020). Brief history of artificial intelligence, *Neuroimaging Clinics of North America*, *30*(4), 393–399.

Newell, A., Shaw, J. C., & Simon, H. A. (1959). Report on a general problem solving program, in *IFIP Congress* (vol. 256, p. 64).

Ng, A. (2018). Machine learning yearning: Technical strategy for AI engineers, in the era of deep learning.

Nøland, J. K., Hjelmeland, M., & Korpås, M. (2024). Will Energy–Hungry AI create a baseload power demand boom?. IEEE Access.

Organisation for Economic Co–operation and Development – OECD (2023). *Artificial Intelligence Outlook 2023: Enabling Trust and Innovation*.

Organisation for Economic Co–operation and Development – OECD (2025). Competition in artificial intelligence infrastructure, OECD Roundtables on Competition Policy Papers, No. 330, OECD Publishing, Paris.

Organisation for Economic Co–operation and Development – OECD (2024). Defining AI incidents and related terms, *OECD Artificial Intelligence Papers*, No. 16, OECD Publishing, Paris.

Organisation for Economic Co–operation and Development – OECD (2024). Digital Economy Outlook 2024 (Volume 1): Embracing the Technology Frontier.

Organisation for Economic Co–operation and Development – OECD. (2026). *Venture capital investments in artificial intelligence through 2025* (OECD Policy Briefs, No. 50). OECD Publishing. https://doi.org/10.1787/a13752f5–en

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Pan, X., Dai, J., Fan, Y., & Yang, M. (2024). Frontier AI systems have surpassed the self–replicating red line. *arXiv:2412.12140*.

Patil, R., Boit, S., Gudivada, V., & Nandigam, J. (2023). A survey of text representation and embedding techniques in NLP, *IEEE Access*, 11, 36120–36146.

Perry, T. S. (2018). Move over, Moore's law. Make way for Huang's law [Spectral Lines]. *IEEE Spectrum*, *55*(5), 7–7.

Polit, S. (1984). "R1 and beyond: AI technology transfer at digital equipment corporation", *AI Magazine*, *5*(4), 76–76.

Quattrociocchi, W. (2025). *Sistemi algoritmici delle piattaforme digitali*. Presentazione tenuta nel corso del seminario "*Le piattaforme online: caratteristiche tecnico-economiche, impatto sociale e tutela delle libertà fondamentali*", Autorità per le Garanzie nelle Comunicazioni – AGCOM.

Rochet, J. C., & Tirole, J. (2003). Platform competition in two–sided markets. *Journal of the European Economic Association*, 1(4), 990–1029.

Rosenblatt, F. (1957). *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Representations by Back–Propagating Errors, *Nature*, *323*(6088), 533–536.

Salvi, F., Horta Ribeiro, M., Gallotti, R., West R. (2025). On the conversational persuasiveness of GPT–4, Nature Human Behaviour.

Samuelson, P. (2023). Generative AI meets copyright. *Science*, *381*(6654), 158–161.

Sejnowski, T. J. (2023). Large language models and the reverse turing test. *Neural computation*, *35*(3), 309–342.

Shapiro, C., & Varian, H. R. (1999). *Information rules: A strategic guide to the network economy*. Harvard Business Press.

Sheikh, H., Prins, C., & Schrijvers, E. (2023). Mission AI: The new system technology, Springer Nature, 410.

Sheikh, H., Prins, C., & Schrijvers, E. (2023). *Mission AI: The new system technology*.

Shick, A.A., Webber, C.M., Kiarashi, N. et al. (2024). Transparency of artificial intelligence/machine learning–enabled medical devices. *npj Digit. Med.* 7, 21.

Shortliffe, E. H., & Buchanan, B. G. (1975). "A model of inexact reasoning in medicine", *Mathematical biosciences*, *23*(3–4), 31–379.

Silvestri, F. (2026). *Architetture e funzionamento dei sistemi di IA.* Presentazione tenuta nel corso del seminario "*Intelligenza artificiale e servizi digitali: tecnologie, impatti e prospettive future*", Autorità per le Garanzie nelle Comunicazioni – AGCOM.

Stern, N., Romani, M., Pierfederici, R., Braun, M., Barraclough, D., Lingeswaran, S., ... & Niemann, N. (2025). Green and intelligent: the role of AI in the climate transition. *npj Climate Action*, *4*(1), 1–7.

Sun, Y., Sheng, D., Zhou, Z., & Wu, Y. (2024). AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence–generated content. *Humanities and Social Sciences Communications*, *11*(1), 1–14.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks, *Advances in Neural Information Processing Systems*, 27.

The 2025 AI Index Report (2025). Stanford University, Human Centered Artificial Intelligence – HAI.

The 2026 AI Index Report (2026). Stanford University, Human Centered Artificial Intelligence – HAI.

Toosi, A., Bottino, A. G., Saboury, B., Siegel, E., & Rahmim, A. (2021). A brief history of AI: how to prevent another winter (a critical review), *PET Clinics*, *16*(4), 449–469.

Turing, A.M., (1950), Computing machinery and intelligence, *Mind*, 49.

UNESCO (2024), Artificial intelligence and democracy.

United Nations, AI Advisory Body (2024). Governing AI for Humanity.

U.S. Department of Defense (2025), *Developmental test and evaluation of autonomous systems guidebook*, Office of the Under Secretary of Defense for Research and Engineering.

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

World Economic Forum (2025), *AI in Action: Beyond Experimentation to Transform Industry*.

World Economic Forum (2025), *The Global Risks Report 2025, 20th Edition*.

World Economic Forum. (2025). *Intelligent Transport, Greener Future: AI as a Catalyst to Decarbonize Global Logistics*.

Xu, F. F., Song, Y., Li, B., Tang, Y., Jain, K., Bao, M., ... & Neubig, G. (2024). Theagentcompany: benchmarking LLM agents on consequential real world tasks. *arXiv*:2412.14161.

Yakura, H., Lopez–Lopez, E., Brinkmann, L., Serna, I., Gupta, P., & Rahwan, I. (2024). Empirical evidence of Large Language Model's influence on human spoken communication. *arXiv preprint arXiv:2409.01754*.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., ... & Qiu, Z. (2024). Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115,* (DeepSeek–V3 Technical Report).

Zhang, J., Hu, S., Lu, C., Lange, R., & Clune, J. (2025). Darwin Godel Machine: Open–Ended Evolution of Self–Improving Agents.

Zhuhadar, L. P., and M. D. Lytras, (2023). The Application of AutoML Techniques in Diabetes diagnosis: Current approaches, performance, and future directions, *Sustainability*, 15 (18), 13484.

8 Indice dei box

9 Indice delle figure

10 Indice delle tabelle

11 Glossario tecnico

Addestramento (*training*): fase in cui un sistema di intelligenza artificiale apprende a partire da grandi quantità di dati, modificando progressivamente i propri parametri interni fino a individuare regolarità, correlazioni e schemi utili a svolgere un determinato compito. Si tratta della fase "formativa" del modello, distinta da quella di inferenza (si veda la relativa voce), nella quale il modello già addestrato viene invece utilizzato per generare output o assumere decisioni. Proprio per questa ragione, mentre l'addestramento serve a costruire il modello, l'inferenza ne costituisce l'impiego operativo. Sotto il profilo ambientale, l'addestramento richiede ingenti risorse computazionali e può protrarsi per settimane o mesi, con elevati consumi di energia e di acqua. Per la distinzione tra addestramento e inferenza si veda § 5.4.

Agente IA (*AI Agent*): nel contesto più attuale dell'intelligenza artificiale, soprattutto con la diffusione dei modelli generativi, si definisce *agente IA* un sistema capace di combinare generazione di contenuti, presa di decisioni e capacità operative. Questi agenti non si limitano a fornire risposte testuali, ma interagiscono con strumenti digitali, accedono a dati esterni e portano a termine compiti articolati in più passaggi. Ne è un esempio ChatGPT, applicazione che può automatizzare flussi di lavoro, programmare, consultare database o esplorare il web.

Accanto a questa accezione operativa, esiste una definizione più ampia e teorica: un *agente intelligente* è qualunque entità artificiale in grado di osservare l'ambiente, agire in modo autonomo e apprendere per raggiungere uno scopo. Si tratta di un concetto generale, alla base di molti sistemi IA, anche semplici, che non necessariamente interagiscono con ambienti esterni complessi.

Allucinazione: nel contesto dell'intelligenza artificiale, un'allucinazione è un contenuto errato, inesatto o inventato generato da un modello, tipicamente linguistico, in assenza di una base fattuale o logica adeguata. Questo fenomeno deriva dalla natura probabilistica e non deterministica dei modelli generativi: tali sistemi non accedono direttamente alla verità oggettiva, ma predicono la sequenza di output più

verosimile dato un certo input, sulla base delle regolarità apprese nei dati di addestramento (§ 5.2).

Apprendimento, tipologie di (§ 3.1.4):

- **Supervisionato**: il sistema apprende da esempi già etichettati, utile per compiti in cui si conosce l'output corretto (es. classificazioni, previsioni).
- **Non supervisionato**: il sistema scopre autonomamente strutture o pattern nei dati non etichettati (es. clustering, riduzione di dimensionalità).
- **Per trasferimento**: conoscenze apprese in un contesto vengono riutilizzate in un altro (es. un modello visivo addestrato su foto può essere adattato a immagini mediche).
- **Per rinforzo**: l'agente impara esplorando l'ambiente e ricevendo ricompense o penalità, affinando la strategia per massimizzare il risultato complessivo (es. giochi, robotica, controllo).

Apprendimento automatico (*Machine Learning* - ML): metodo mediante il quale i sistemi informatici apprendono dai dati, individuando regolarità e relazioni tra essi senza essere programmati in modo rigido. Grazie all'esperienza, migliorano le prestazioni e riescono ad applicare quanto appreso anche a nuovi dati (§ 3.1.2). Le applicazioni del *Machine Learning* sono molteplici, e tipologie di algoritmi variano in base agli obiettivi, tra di essi vi sono:

- **Algoritmi di classificazione**, utilizzati per assegnare etichette a nuovi dati, come nel filtraggio antispam, nell'analisi del sentiment o nella rilevazione di attacchi DDoS;
- **Algoritmi di regressione**, stimano relazioni tra variabili e realizzare previsioni in contesti socio–economici, ad esempio per la domanda energetica, le vendite azionarie o il valore degli immobili;

- **Algoritmi di clustering**, impiegati per raggruppare dati in insiemi omogenei in base a criteri di similarità, come nella segmentazione dei clienti nell'e–commerce o nella classificazione dei pazienti in ambito sanitario.

***Deep Learning* (DL)**: ramo del *Machine Learning* che utilizza reti neurali profonde per apprendere rappresentazioni complesse dei dati. Le reti sono composte da strati di "neuroni" artificiali connessi da pesi che si aggiornano durante l'addestramento. Gli strati successivi catturano strutture sempre più astratte: dai contorni in un'immagine fino al riconoscimento di oggetti o contesti. È la tecnologia alla base dei progressi più avanzati nel linguaggio, nella visione artificiale e nella robotica (§ 3.1.3).

Distillazione (*Knowledge Distillation*): tecnica di compressione dei modelli di intelligenza artificiale che consiste nel trasferire le conoscenze da un modello di grandi dimensioni (detto *teacher*) a uno più piccolo e leggero (detto *student*). Il modello studente non apprende direttamente dai dati grezzi, ma cerca di imitare le risposte del modello insegnante, replicandone le decisioni e le probabilità di output. In questo modo si ottiene un modello più efficiente, adatto a essere utilizzato in ambienti con risorse limitate (come dispositivi mobili o applicazioni in tempo reale), senza perdere troppo in accuratezza (§ 4.3).

Elaborazione del Linguaggio Naturale (*Natural Language Processing* – NLP): campo dell'IA che si occupa della comprensione, generazione e traduzione del linguaggio umano da parte dei computer. Interviene in attività come la sintesi automatica, l'analisi del sentiment, la ricerca semantica e la generazione di testo coerente (§ 3.1.3).

Fine-tuning: tecnica di apprendimento automatico che consiste nell'adattare un modello pre–addestrato a un nuovo compito o a un nuovo insieme di dati: si parte da un modello che ha già appreso conoscenze generali su un compito simile e si prosegue l'addestramento con un numero minore di dati specifici per specializzarlo su un problema più ristretto o più mirato (§ 4.3).

Inferenza: fase in cui un modello di intelligenza artificiale, una volta addestrato, viene impiegato per produrre un risultato concreto: ad esempio rispondere a una domanda, classificare un contenuto, formulare una previsione o generare un testo. Essa si distingue dall'addestramento (si veda la relativa voce), che è la fase precedente nella quale il modello apprende dai dati e acquisisce la propria struttura funzionale. In altri termini, l'addestramento costruisce il modello, mentre l'inferenza lo mette al lavoro. Sebbene tradizionalmente l'attenzione si sia concentrata soprattutto sui costi energetici dell'addestramento, oggi è spesso l'inferenza, proprio per la sua ripetizione continua e su vasta scala, a rappresentare la componente prevalente del consumo complessivo associato all'IA. Per la distinzione tra addestramento e inferenza si veda § 5.4.

Hyperscale (data center iperscalabili): grande centro dati progettato per aumentare rapidamente (e in modo modulare) capacità di calcolo e archiviazione (storage) a supporto di carico di lavoro (workload) su vasta scala – tipicamente cloud e IA – grazie a un'elevata automazione e ad architetture infrastrutturali ottimizzate (§4.3).

Hyperscaler: grande operatore che costruisce o gestisce quell'infrastruttura.

Intelligenza Artificiale Agentica (*Agentic Artificial Intelligence*): indica una classe di sistemi progettati per perseguire un obiettivo definito con supervisione umana limitata, selezionando e organizzando in modo autonomo le azioni necessarie al suo conseguimento. Il sistema può essere composto da uno o più agenti di IA, ossia componenti basate su modelli di apprendimento automatico capaci di percepire informazioni di contesto, pianificare e decidere sequenze di azioni in tempo (quasi) reale. Nei sistemi multi–agente, ciascun agente esegue una sotto–attività specializzata funzionale all'obiettivo complessivo; il coordinamento è assicurato da meccanismi di orchestrazione che assegnano compiti, gestiscono dipendenze e integrano gli output.

Intelligenza Artificiale Generale (*Artificial General Intelligence* – AGI): rappresenta l'ideale di una IA capace di eseguire qualsiasi compito cognitivo che un essere umano saprebbe affrontare. Non è limitata a domini specifici, ma può ragionare,

apprendere, adattarsi a contesti nuovi e perfino mostrare forme di autoconsapevolezza. Al momento, resta un obiettivo di ricerca e non una realtà compiuta (§ 3.3).

Iperpersuasione (*Hyperpersuasion*): capacità delle intelligenze artificiali generative di influenzare opinioni e comportamenti umani in modo progressivo e adattivo, rispondendo alle aspettative dell'utente e rafforzandone le convinzioni. A differenza della persuasione **tradizionale**, sfrutta l'**allineamento**, cioè il processo attraverso cui l'IA modella le proprie risposte in base ai desideri, ai valori e allo stile comunicativo dell'interlocutore, ottenendo un effetto più profondo e meno visibile (§ 5.5).

Modelli linguistici di grandi dimensioni (*Large Language Model* – LLM): sono reti neurali profonde specializzate nel trattamento del linguaggio. Addestrati su enormi quantità di testo, riescono a cogliere il significato, il tono e il contesto delle frasi, generando risposte articolate e comprensibili a una richiesta utente (prompt). Consentono traduzioni più precise, sintesi complesse e interazioni simili al dialogo umano (§ 3.1.3).

Opacità degli algoritmi di IA (*black box*): difficoltà nel comprendere il processo attraverso cui un sistema di *machine learning* perviene a una determinata decisione o previsione. Questi modelli, in particolare quelli più complessi come le reti neurali profonde, processano grandi quantità di dati mediante calcoli opachi e difficilmente interpretabili dagli esseri umani. Tale mancanza di trasparenza genera preoccupazioni etiche e pratiche, poiché l'assenza di una motivazione chiara delle decisioni algoritmiche rischia di compromettere la responsabilità, la fiducia e l'equità (§ 3.5, § 5.2 e Tabella 7 – Problematiche generali dell'IA).

Token: rappresenta l'unità fondamentale di elaborazione di un modello linguistico (*Large Language Model*). A differenza dei sistemi tradizionali che leggono il testo parola per parola, i modelli moderni scompongono i dati in segmenti atomici chiamati token, che possono corrispondere a intere parole, sillabe, singoli caratteri o persino frammenti di codice informatico. Questo processo, chiamato **tokenizzazione**,

permette al modello di gestire con efficienza le lingue complesse, i neologismi e gli errori ortografici, trasformando il linguaggio naturale in una sequenza numerica (vettori) che la macchina può elaborare statisticamente. In media, nei modelli più diffusi, 1.000 token equivalgono a circa 750 parole.

12 Indice analitico