Università degli Studi di Pavia
Facoltà di Relazioni Internazionali
Laurea Magistrale in Sviluppo Economico e Relazioni Internazionali

L'intelligenza artificiale tra innovazione, sostenibilità e governance: impatti ambientali, infrastrutture e regolazione

Candidato: Oliviero Protti

Relatrice: Prof.ssa Roberta Rabellotti
Correlatore: Prof. Andrea Morrison

Anno accademico 2025-2026

RINGRAZIAMENTI

Desidero ringraziare tutte le persone che, in modi diversi, hanno contribuito a questo percorso e alla realizzazione di questa tesi.
Un primo ringraziamento va alla mia relatrice, al mio correlatore, ai professori e ai compagni di corso con cui ho condiviso questi due anni.
Ringrazio poi chi ha accompagnato la mia formazione anche al di fuori dell'università. L'autista che suona mentre attraverso sulle strisce, ricordandomi con grande coerenza che la pazienza è una virtù rara. Trenord, che per due anni, un giorno sì e l'altro pure, ha saputo introdurre nelle mie giornate quella giusta dose di adrenalina. E ultimo, ma non per importanza, Marco per avermi garantito di non perdere mai una sveglia.
Un grazie sincero va agli amici e a tutte le persone che condividono con me parte della loro vita, rendendo più leggeri anche i momenti più impegnativi.

Ringrazio Annachiara, per la crescita condivisa.
La mia famiglia, che mi ha sempre spinto a fare meglio.
Tommi, che mi tiene a contatto con la realtà.
E infine Ghebbi, a cui va un ringraziamento speciale per avermi permesso di concepire questa tesi a dieci metri dal mare dell'Isola d'Elba.

INDICE

INTRODUZIONE

Negli ultimi anni, e in modo particolarmente evidente nell'ultimo periodo, l'intelligenza artificiale (IA) sta vivendo una fase di espansione rapidissima: nuovi modelli, nuovi casi d'uso e nuovi investimenti stanno trasformando l'IA in una tecnologia di uso generale, capace di attraversare settori e funzioni molto diverse.

A questa crescita, però, si accompagna un lato materiale spesso sottovalutato: diverse stime segnalano consumi energetici e idrici potenzialmente rilevanti, soprattutto per l'addestramento e l'uso su larga scala dei modelli. In questo contesto, l'obiettivo della tesi è descrivere il settore dell'IA con particolare attenzione alla sostenibilità, mettendo a fuoco le caratteristiche del settore, i costi ambientali e infrastrutturali e la progressiva attività di rendicontazione e regolazione.

La letteratura e il dibattito pubblico tendono a concentrarsi soprattutto sugli impatti operativi come le emissioni corrispondenti all'attività di allenamento o inferenza. Molto meno esplorata; e spesso più difficile da misurare, è la dimensione degli impatti incorporati (*embodied*) nella produzione di chip, server e infrastrutture, e più in generale nella *supply chain*. Questa tesi adotta quindi un approccio lungo l'intero ciclo di vita, con l'intento di integrare impatti operazionali e impatti incorporati e di individuare punti di intervento per una progettazione e una governance più sostenibile.

L'IA viene spesso raccontata come una "tecnologia immateriale": un insieme di algoritmi che vivono nel *cloud* e producono decisioni, testi e immagini con l'apparente leggerezza del software.

Eppure, proprio mentre l'IA diventa una tecnologia di uso generale la sua dimensione più concreta si impone con forza: l'IA è anche infrastruttura. È fatta di datacenter, chip, reti, energia e acqua; dipende da catene di approvvigionamento globali e da territori specifici; richiede capitale, competenze, e una capacità industriale che la avvicina più all'industria

pesante che al mito romantico dell'innovazione "da garage".

Questa tesi si colloca esattamente su questa frattura apparente tra astrazione algoritmica e materialità infrastrutturale, con un obiettivo: ricostruire come la crescita dell'IA, i suoi impatti ambientali e i tentativi di governance si intreccino in modo strutturale.

Il punto di partenza è la consapevolezza che la transizione digitale e la transizione verde non siano due traiettorie separate, ma due processi che si condizionano a vicenda. Da un lato, l'IA viene spesso presentata come strumento abilitante per l'efficienza, l'ottimizzazione e la riduzione delle emissioni in numerosi settori; dall'altro, il suo stesso sviluppo genera nuove pressioni su energia, risorse e infrastrutture (van Wynsberghe, 2021).

Questa tensione è affrontata in modo diretto: i dati disponibili sull'impatto dell'IA sono ancora spesso incompleti e disomogenei, talvolta dipendenti da *disclosure* aziendali, oltre che dal quadro normativo frammentato. Per questo, diventa centrale una metodologia capace di distinguere tra impatti diretti e indiretti, e di seguirli lungo il ciclo di vita dell'hardware, del *training* e del processo di inferenza.

Solo con misure condivise basate su: dei dati trasparenti e standardizzati, la \responsabilità nei confronti degli *stakeholders* e degli utenti finali e l'auditabilità da parte di terze parti; è possibile ottenere una governance che permetta lo sviluppo del settore in modo etico, sostenibile e democratico.

Per rendere visibile la struttura del fenomeno, la tesi procede in tre movimenti. Il primo capitolo costruisce una mappa del settore attraverso indicatori di innovazione: pubblicazioni e citazioni per la ricerca accademica, brevetti e investimenti per la ricerca privata, e infine il ruolo delle start-up come ponte tra conoscenza scientifica e mercato.

Questa lettura, apparentemente "quantitativa", ha in realtà una funzione interpretativa: mostra chi sta innovando, dove si concentrano capitale e capacità produttiva, e quali incentivi guidano le scelte tecnologiche. Questo quadro si connette direttamente alla dimensione geopolitica e industriale

dell'IA. Il secondo capitolo entra nella materialità: hardware, *supply chain*, energia e acqua. Qui la questione non è solo "quanto consuma l'IA", ma che cosa si sta misurando quando si parla di consumi.

L'analisi distingue tra impatti *embodied* nella produzione di chip, server e infrastrutture, e impatti operazionali legati al funzionamento quotidiano di datacenter e *workload* di *training* o inferenza. La tesi insiste su un punto metodologico cruciale: metriche locali e parziali possono produrre illusioni di sostenibilità.

Qui si innesta un altro tema chiave che attraversa l'intera tesi: la sostenibilità dell'IA non è solo un problema di efficienza tecnica, ma anche di attribuzione e di confini. Stabilire chi "paga" l'impatto, dove esso si accumula, e come ripartirlo tra attori e fasi del ciclo di vita è un'operazione tanto politica quanto tecnica.

Proprio perché l'IA dipende da filiere lunghe e opache, e perché gli impatti possono spostarsi tra categorie, diventa necessaria una prospettiva capace di evitare errori di "attribuzione" degli oneri ambientali: ciò che migliora un indicatore locale può peggiorare l'impatto totale, o semplicemente renderlo meno misurabile.

Il terzo capitolo porta questa complessità sul terreno della governance. Prima ancora delle norme, il capitolo ricostruisce il settore come arena di attori con potere asimmetrico: gli *incumbent* digitali dotati di capitale e infrastrutture e i nuovi attori che si muovono tra cooptazione, specializzazione, e posizionamenti geopolitici e simbolici.

In questo quadro, il concetto di trasparenza non viene trattato come mera "disponibilità di informazioni", ma come proprietà relazionale e istituzionale: ciò che conta non è solo pubblicare dati o documenti, ma renderli intelligibili, verificabili e contestabili da attori esterni. Senza questa reciprocità, la trasparenza può ridursi a una messa in scena tecnica o a una narrazione controllata. È qui che entra con forza il tema dell'auditabilità, e con esso l'idea che misurazione e regolazione non siano capitoli separati.

Il terzo capitolo propone una cornice che combina strumenti di contabilità climatica e approcci di *life-cycle thinking*, evidenziando tre livelli complementari di reporting: *corporate*, infrastrutturale e di *workload* per training e inferenza. L'argomento è chiaro: senza standard, metriche comparabili e procedure di verifica indipendente, la governance rischia di oscillare tra principi troppo generici e obblighi formali. In questo senso, anche gli standard ISO e i sistemi di gestione vengono letti come "infrastrutture" istituzionali non risolvono da soli i problemi, ma rendono praticabile una regolazione verificabile.

All'interno di questo impianto, il documento di Patterson et al. (2022) "*The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink*", fornisce un punto di snodo particolarmente utile, perché rende esplicito come la discussione sui consumi dell'IA dipenda dal modo in cui li si contabilizza. Gli autori distinguono tra emissioni operazionali ed emissioni di ciclo di vita, dichiarando di concentrarsi principalmente sulle prime.

Questa distinzione si incastra perfettamente con il secondo capitolo: da un lato, permette di capire come selezionate best practice ingegneristiche agiscono sugli impatti operazionali; dall'altro, evidenzia che una lettura completa richiederebbe di integrare anche gli impatti incorporati, che tendono a diventare relativamente più importante man mano che l'operativo viene decarbonizzato o reso più efficiente.

Per questa tesi, lo studio di Patterson et al. (2022) è importante per due ragioni complementari. La prima è metodologica: offre un linguaggio semplice e "datacenter-aware" per collegare consumo energetico, PUE e carbon intensity, mostrando che l'impatto non è solo funzione del modello, ma dell'infrastruttura e del contesto energetico.

La seconda è politico-istituzionale: gli autori chiedono che i provider pubblichino metriche come PUE, percentuale di energia carbon-free e CO_2e per MWh per location, così che gli utenti possano scegliere e ridurre il proprio impatto; e invitano i ricercatori a pubblicare consumi ed emissioni

per favorire competizione anche su dimensioni diverse dalla sola performance. In altre parole, vi è un suggerimento implicito ad una proposta di governance: senza trasparenza infrastrutturale e senza dati condivisi, non esiste "scelta efficiente" verificabile, e quindi nemmeno una regolazione credibile.

Alla luce di queste premesse, i capitoli che seguono sviluppano un percorso coerente: prima chiarire come l'IA si stia strutturando come settore innovativo e geopolitico; poi mostrare che la sua crescita ha un costo materiale misurabile e non sempre trasparente; infine discutere quali strumenti, tecnici e istituzionali, possono trasformare principi di sostenibilità e trasparenza in meccanismi praticabili di controllo, rendicontazione e responsabilità lungo la filiera.

L'ambizione non è offrire una risposta definitiva a un fenomeno in rapida evoluzione, ma costruire una lente interpretativa sufficientemente robusta: una lente che tenga insieme innovazione, infrastruttura e governance, evitando sia un entusiasmo cieco per il progresso, sia un pessimismo luddista che ignora le leve concrete di miglioramento.

1. Il contesto innovativo dell'Intelligenza Artificiale

Per analizzare in modo sistematico il settore dell'Intelligenza Artificiale, prenderemo in esame diversi indicatori che coprono tutti gli attori coinvolti attivamente nel comparto.

Per quanto riguarda la ricerca accademica, faremo riferimento al numero di pubblicazioni e alla loro rilevanza, misurata attraverso le citazioni. Analizzeremo come questi dati si siano evoluti nel tempo, le loro variazioni su base geografica, regionale e nazionale.

Per la ricerca privata, invece, considereremo il numero di brevetti depositati e gli investimenti nel settore, mantenendo anche in questo caso l'analisi spaziale e temporale.

Inoltre, data l'attualità del tema, offriremo una panoramica dei principali sistemi di intelligenza artificiale, approfondendone l'origine, lo sviluppo e le fonti di finanziamento (pubbliche o private).

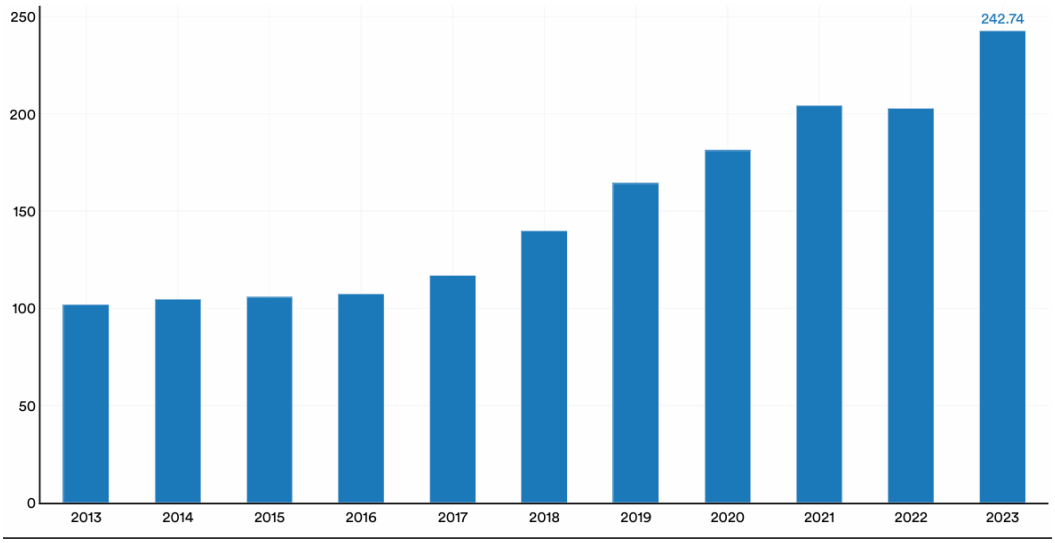
1.1 Le pubblicazioni

Il numero di pubblicazioni globali mostra chiaramente come il settore dell'IA sia in forte espansione da oltre un decennio: dal 2013 al 2023 le pubblicazioni sono più che raddoppiate, passando da 102.000 a 242.000, con un picco di crescita del 19,7% proprio nel 2023 (Stanford University HAI, 2025). Questi lavori coprono l'intero spettro delle applicazioni dell'IA, esplorandone ogni aspetto, dall'hardware e software engineering fino all'interazione uomo-macchina.

La natura intrinsecamente multidisciplinare del fenomeno riflette la pervasività di questa tecnologia, sia sul piano economico, sia su quello

sociale, attirando l'interesse di studiosi di numerose discipline e spiegando la crescita esponenziale della produzione scientifica.

Grafico 1.1 - Numero di pubblicazioni IA nel mondo per anno (Stanford University HAI, 2025).

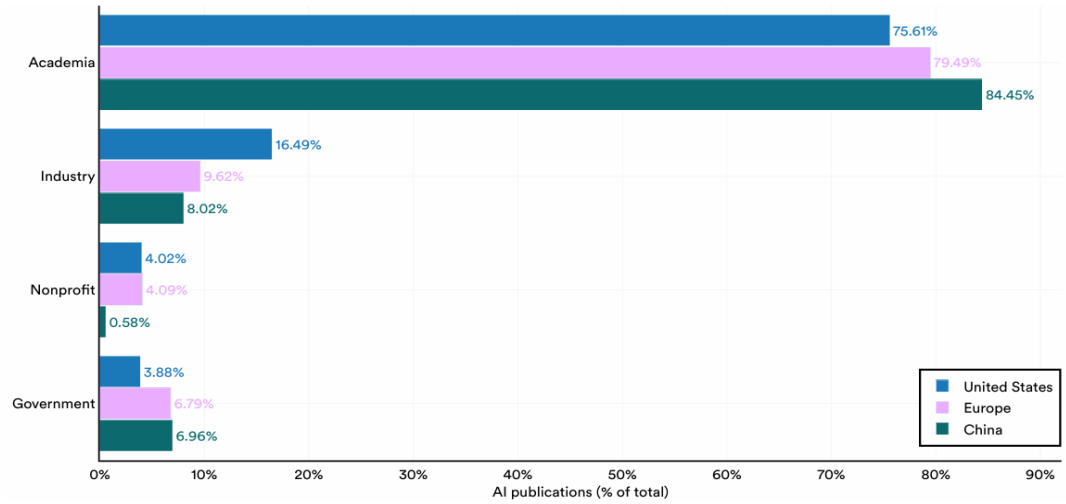


Scomponendo i dati su base geografica per il periodo 2013-2023, emerge che nel 2023 la quota principale di pubblicazioni proviene dall'Asia orientale e Pacifico, con il 34,5% del totale, seguita da Europa e Asia centrale con il 18,2% e dal Nord America con il 10,3%. Analizzando le citazioni, sempre nel 2023, l'Asia orientale e il Pacifico rappresentano il 37,1% del totale, mentre le quote europee e nordamericane sono diminuite rispettivamente al 21,88% e al 15,59% (Stanford University HAI, 2025).

A livello statale, la Cina è il principale produttore di pubblicazioni nel 2023, seguita da Unione Europea, India e Stati Uniti. Il primato cinese si spiega con la crescita costante avviata dal 2016, accompagnata da un lieve calo delle pubblicazioni europee. Quelle statunitensi sono rimaste relativamente stabili fino al 2021, anno a partire dal quale hanno iniziato a diminuire leggermente. In Cina, la ricerca è svolta prevalentemente in ambito accademico, che rappresenta l'84,5% delle pubblicazioni, mentre il settore privato contribuisce solamente per l'8% (Stanford University HAI, 2025).

Negli Stati Uniti la situazione è simile, sebbene la quota industriale sia più elevata (16,5%), grazie alla presenza di poli tecnologici ad alta capacità innovativa come la Silicon Valley o i laboratori del MIT.

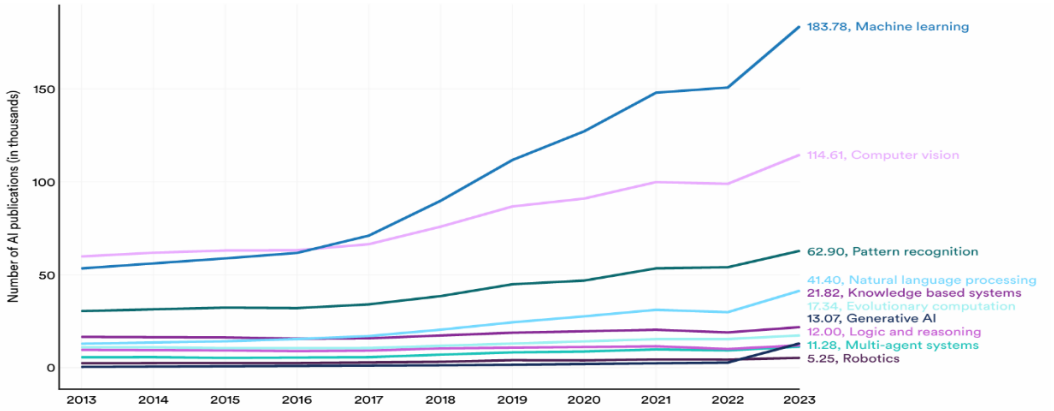
Grafico 1.2 - Numero di pubblicazioni per settore di ricerca, per area geografica (Stanford University HAI, 2025).



Tuttavia, dal 2023 si registra un drastico calo delle pubblicazioni provenienti dal settore privato, segnale di una crescente competizione: i laboratori industriali tendono infatti a divulgare meno dettagli e con minore frequenza per mantenere il vantaggio competitivo.

Per quanto riguarda gli ambiti di ricerca principali, al primo posto troviamo il machine learning (75,5%), seguito da computer vision (47,2%), pattern recognition (25,9%) e natural language processing (17,1%) (Stanford University HAI, 2025). Negli ultimi tre anni è cresciuto rapidamente soprattutto l'interesse per le IA generative.

Grafico 1.1 - Numero di pubblicazioni per topic (Stanford University HAI, 2025).



Su questo fronte emerge una differenza significativa tra USA e Cina: quest'ultima concentra la ricerca soprattutto in altri ambiti applicativi dell'IA, come biologia, medicina e robotica, mentre negli Stati Uniti la quota relativa alla GenAI raggiunge il 39% (Ding et al., 2025). Gli Stati Uniti si sono mossi più rapidamente della Cina nell'utilizzo della GenAI per la produzione scientifica fino al 2023.

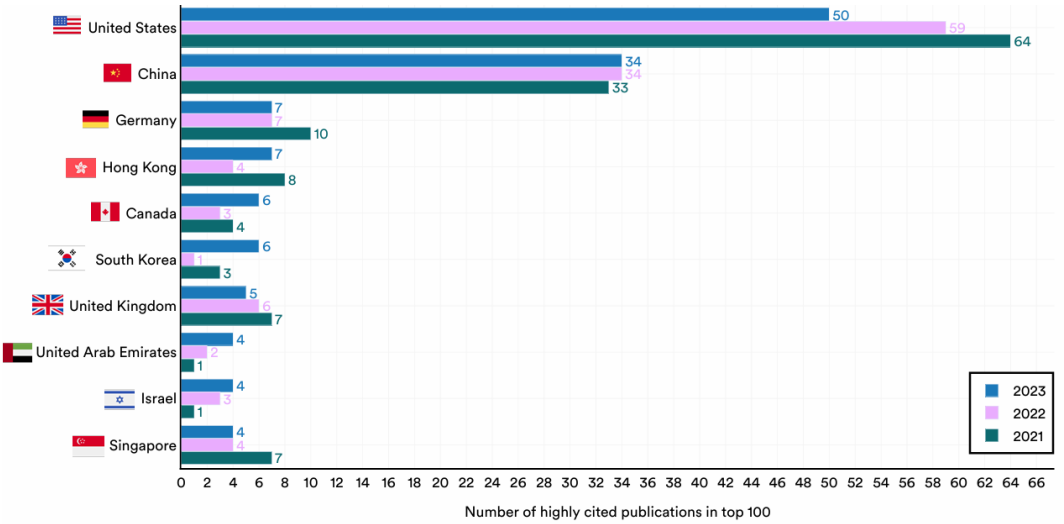
La comunità di ricerca statunitense ha sfruttato il vantaggio competitivo derivante da un ecosistema tecnologico dinamico, caratterizzato da alti livelli di investimenti in R&S e da ingenti finanziamenti sia da parte di grandi aziende tecnologiche, sia di nuovi operatori sostenuti da venture capitalist. La Cina, invece, riflette i suoi intensi sforzi in R&S e l'emergere di un ecosistema pubblico-privato più integrato, attraverso un'elevata produzione di pubblicazioni in altri ambiti dell'IA. Sebbene sia stata il secondo produttore di studi sulla GenAI fino al 2023, questo risultato è compensato da una base molto più ampia di pubblicazioni "Other-AI" (Ding et al., 2025). Il settore privato cinese, infatti, sta ora accelerando gli sforzi per sviluppare e applicare modelli generativi propri.

Altri Paesi con livelli elevati di utilizzo della GenAI nella ricerca includono Svezia, Lussemburgo e Corea del Sud. In Europa la maggior parte della ricerca proviene dal settore accademico (79,5%), mentre l'industria

contribuisce per il 9,6% (Stanford University HAI, 2025).

La quota destinata ai modelli generativi rimane contenuta, con l'Inghilterra come eccezione. Le pubblicazioni più citate in questo nuovo campo di ricerca riguardano ancora una volta i principali modelli di IA generativa, come i report tecnici su GPT-4 (OpenAI), Llama 2 (Meta) e PaLM-E (Google), sottolineando ancora di più l'interesse e l'attualità del settore.

Grafico 1.2 - Numero di pubblicazioni più citate (Stanford University HAI, 2025).



1.2. I brevetti

Oltre al numero di pubblicazioni e al sistema delle citazioni, un altro indicatore fondamentale della capacità innovativa di un settore è rappresentato dal numero di nuovi brevetti prodotti. Per comprendere come il comparto dell'Intelligenza Artificiale o, più nello specifico, del *deep learning* si interfacci con il sistema brevettuale, è utile partire dall'*European Patent Office*.

Per essere brevettabile, un'invenzione deve soddisfare sia i requisiti di ammissibilità sia quelli di brevettabilità. L'ammissibilità esclude, per definizione, materie considerate attività mentali o concetti astratti, come

metodi matematici, teorie scientifiche o metodi commerciali. La brevettabilità, invece, richiede che l'invenzione sia nuova e presenti un passo inventivo rispetto allo *state of the art*[1].

Nel caso dell'IA, la sfida principale risiede nel confine tra ciò che è astratto, un algoritmo matematico, e ciò che presenta un carattere tecnico, ossia un contributo concreto volto a risolvere un problema tecnico tramite mezzi tecnici. L'IA è quindi ammissibile alla protezione brevettuale non come algoritmo in sé, ma quando è applicata in un contesto tecnico, ad esempio: il controllo di un sistema o processo; la generazione o il miglioramento digitale di audio, immagini o video e la codifica, decodifica o compressione dei dati per una trasmissione o archiviazione affidabile ed efficiente.

La brevettabilità è riconosciuta solo se le caratteristiche tecniche che conferiscono ammissibilità distinguono l'invenzione dallo stato dell'arte. Queste caratteristiche devono essere nuove e rappresentare un avanzamento non ovvio per un tecnico del settore. In concreto, un processo non è brevettabile se si limita a eseguire un metodo di IA su un computer: per esserlo deve presentare uno scopo tecnico chiaro, un'architettura funzionale, input in forma di dati e output coerenti con l'obiettivo dichiarato.

Sebbene i dati di addestramento e le funzioni di costo non siano esplicitamente richiesti in sede di domanda, essi risultano spesso necessari per garantire sufficiente supporto tecnico, chiarezza e passo inventivo[2].

Negli Stati Uniti, l'USPTO[3] utilizza un criterio diverso: invece del *technical test* adotta l'*Alice/Mayo test*, un procedimento a più step che mira a garantire che l'innovazione non sia una mera astrazione. Dopo aver verificato che l'invenzione rientri in una delle quattro categorie brevettabili: *prodotto*, *processo*, *macchina* o *composizione di materia*, si valuta se essa superi lo

[1] Con "State of the art" ci si riferisce al livello di sviluppo più recente e sofisticato raggiunto da una tecnologia, una tecnica, un processo o una conoscenza, basato sui risultati e le pratiche più avanzate disponibili al momento.
[2] Come definiscono gli articoli 83-84-56 dell'European Patent Convention del 1973.
[3] United States Patent Office.

stadio di “idea astratta” e rappresenti un vero inventive concept. In questo senso, algoritmi per il marketing o strategie di business non sono considerati innovazioni brevettabili, mentre lo sono, ad esempio, algoritmi applicati all’ambito medico per il miglioramento dell’imaging attraverso l’eliminazione del rumore, in quanto traducono un’astrazione in un impiego concreto e utile.

Anche la CNIPA[4] ha pubblicato nel 2020 delle linee guida sulla brevettabilità dell’IA. Pur simili nella forma a quelle dell’EPO e dell’USPTO, nella pratica risultano più pragmatiche e meno restrittive.

In Cina, un’invenzione IA deve superare tre criteri:

1. Ammissibilità: vengono esclusi gli algoritmi puri, mentre sono accettati i metodi tecnici che implementano algoritmi, ossia combinazioni di caratteristiche tecniche e algoritmiche.

2. Requisito tecnico: sono richiesti input e output chiari associati a un’applicazione concreta.

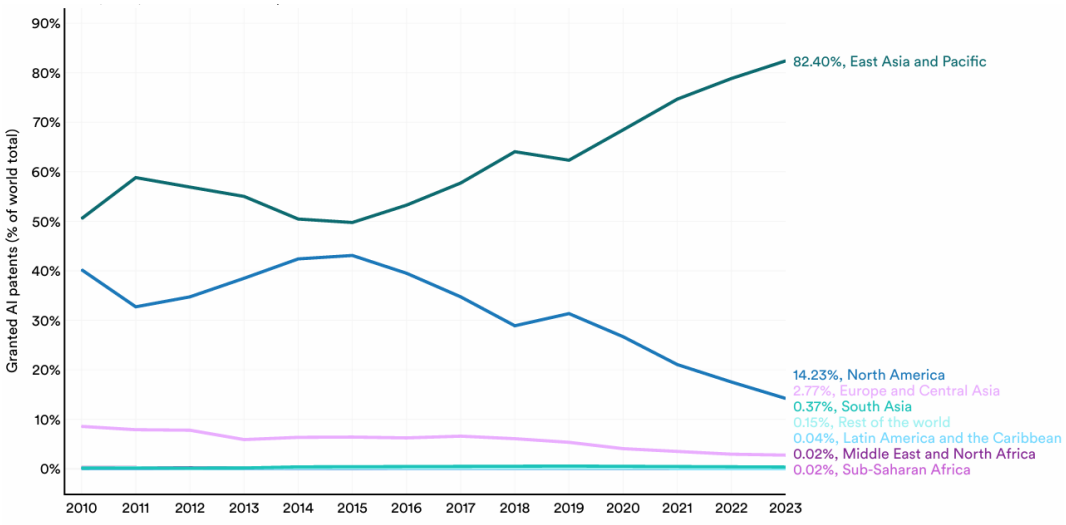
3. Novità e inventiva: l’invenzione deve essere sconosciuta allo *state of the art* e rappresentare un contributo tecnico non ovvio.

Proprio nell’interpretazione dell’ultimo criterio la CNIPA si distingue: considera brevettabili anche miglioramenti settoriali come l’applicazione dell’IA all’e-commerce o al riconoscimento facciale, elementi che l’EPO non riconoscerebbe come dotati di carattere tecnico. Osservando i dati dal 2010 al 2023, il numero di brevetti relativi all’IA è cresciuto in modo significativo: da 3.833 unità nel 2010 a 122.511 nel 2023, con un incremento del 29,6% (Stanford University HAI, 2025) solo nell’ultimo anno. Nel 2023, l’Asia

[4] China National Intellectual Property Administration.

orientale e il Pacifico hanno rappresentato l'82,4% dei brevetti di IA concessi a livello mondiale, seguiti dal Nord America con il 14,2%. Dal 2010 il divario tra queste due regioni si è progressivamente ampliato.

Grafico 1.3 - Numero di brevetti concessi per area geografica (Stanford University HAI, 2025).



A livello nazionale, la Cina detiene la quota maggiore (69,7%), mentre quella degli Stati Uniti è diminuita da un picco del 42,8% nel 2015 al 14,2% nel 2023 (Stanford University HAI, 2025).

Tuttavia, non è solo la quantità a contare: occorre considerare anche la qualità dei brevetti, misurata attraverso parametri come durata, grado di commercializzazione e valore di mercato. Questi elementi evidenziano come l'attività brevettuale cinese, pur intensa e in crescita, non sempre corrisponda a un reale avanzamento tecnico.

Il numero elevato di brevetti è spesso spiegato dagli incentivi governativi volti a colmare il divario con i paesi più sviluppati. Tali brevetti risultano frequentemente di basso valore innovativo, con scarsa durabilità e, in molti casi, non vengono mai commercializzati. Un fenomeno diffuso è quello della brevettazione duplicata, in cui vengono depositati brevetti simili con leggere variazioni. Oppure della brevettazione strategica, pratica globale che consiste

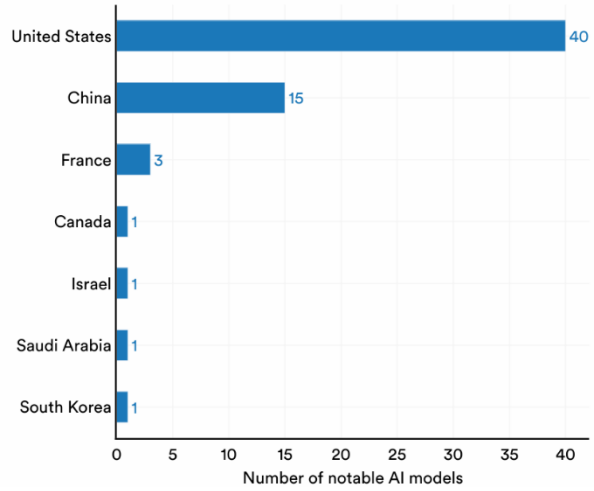
nel circondare un'invenzione principale con "brevetti satellite" per ostacolare i concorrenti. Questi meccanismi mostrano come politiche di incentivo mal calibrate possano produrre effetti indesiderati: ad esempio, i sussidi per le tasse di deposito possono incoraggiare la presentazione di brevetti di scarso valore e aumentare il carico di lavoro degli esaminatori.

Per mitigare il problema, alcune amministrazioni locali, come quelle di Hunan e Zhejiang (Dang e Motohashi, 2015) hanno sostituito i sussidi con finanziamenti condizionati alla concessione del brevetto. Anche in questo caso, tuttavia, permane il rischio di depositi strategici volti esclusivamente a ottenere fondi e creare un groviglio brevettuale per ostacolare la concorrenza.

Un'ultima peculiarità del mercato cinese dei brevetti è il suo forte orientamento domestico. Ciò è dovuto, da un lato, agli incentivi e ai sistemi di punteggio aziendale legati agli appalti pubblici, e dall'altro alle difficoltà di accesso ai mercati esteri. Per entrare nei sistemi brevettuali dell'EPO e dell'USPTO, le imprese cinesi devono affrontare procedure più lunghe (in media 3-5 anni) e costi più elevati (tasse di deposito ed esame, traduzioni tecniche, spese legali, rappresentanza locale, costi di mantenimento).

Di conseguenza, la maggior parte dei brevetti cinesi non arriva sui mercati esteri, sia perché non supera i criteri più severi di esame europei e statunitensi, sia per i costi economici elevati che tali procedure comportano.

Grafico 1.4 - Numero di modelli IA di rilievo (Stanford University HAI, 2025).



1.3. Gli investimenti Pubblici

Nel complesso, i governi di tutto il mondo stanno trattando l'IA come infrastruttura strategica, ma la misurazione comparabile dei fondi che stanziano a riguardo resta difficile. Le linee di bilancio IA sono spesso inglobate nella spesa generale in R&S: l'OCSE sta quindi sviluppando metodi ad hoc, come Fundstat[5], proprio per identificare la "quota IA" dei progetti finanziati con fondi pubblici.

Questo perché grazie alla crescita della spesa in R&S, proseguita in modo eterogeneo nel 2023–2025 e l'espansione del perimetro IA verso sempre più numerosi programmi governativi settoriali, si intuisce subito come questa tecnologia sia radicale e destinata a portare forti cambiamenti.

In questo quadro, gli Stati Uniti hanno introdotto una rendicontazione federale specifica: il *Supplement to the President's FY2025 Budget* (2025) che quantifica in 3,3 miliardi di dollari la richiesta per AI R&D sulle agenzie

[5] Fundstat è un prototipo di infrastruttura analitica distribuita che raccoglie dati sui progetti R&D finanziati dai governi di vari paesi OCSE e nell'UE. Non si occupa di misurare i finanziamenti solo a livello quantitativo, ma ne studia anche la "direzionalità".

NITRD[6], in aumento del 6,5% rispetto alla richiesta FY2024, a conferma di una traiettoria al rialzo della spesa pubblica dedicata.

L'Unione Europea ha adottato un'impostazione "a portafoglio" che combina programmi a gestione diretta e cofinanziamento degli Stati membri: la Commissione indirizza almeno 1 miliardo di euro l'anno su *Horizon Europe* e *Digital Europe* per iniziative IA, con l'obiettivo politico di mobilitare 20 miliardi di euro l'anno tra pubblico e privato lungo il decennio digitale (Commissione Europea [CE], n.d.).

Per quanto riguarda la Cina invece la strategia si orienta verso un mix di piani centrali, fondi dedicati e partnership pubblico-privati. Al piano nazionale del 2017 sono state affiancate nuove iniziative come il *National AI Industry Investment Fund* (Ministero dell'Industria e della Tecnologia cinese, 2025) da 60 miliardi di yuan, annunciato nel 2025 e, soprattutto, il varo nel marzo 2025 di un fondo guida di venture capital con capacità di circa un trilione di yuan (Consiglio di Stato cinese, 7 marzo 2025).

In sintesi, il trend globale mostra una normalizzazione dell'IA nei bilanci pubblici di ricerca e sviluppo con incrementi nominali nelle economie avanzate, oltre a una crescente ibridazione pubblico-privato che rende più complessi i confronti internazionali.

1.3.1. Stati Uniti

Negli Stati Uniti, gli investimenti pubblici dedicati all'intelligenza artificiale rappresentano una componente sempre più rilevante della strategia federale per l'innovazione tecnologica e la competitività nazionale. Il governo statunitense ha consolidato la propria azione attraverso il *National AI*

[6] Networking and Information Technology Research and Development Program: un programma interagenzia creato dal governo degli Stati Uniti per coordinare gli investimenti federali in ricerca e sviluppo IT. In particolare, si occupa di evitare duplicazioni e di favorire la collaborazione tra le varie agenzie che finanziano o conducono ricerca nel campo delle tecnologie digitali.

Initiative Act (Congresso degli Stati Uniti d'America, 2020) del 2020, che coordina le attività delle principali agenzie federali, tra cui il *National Science Foundation* (NSF), il *Department of Energy*, il *National Institutes of Health* e la *Defense Advanced Research Projects Agency*, con l'obiettivo di sostenere la ricerca di base, accelerare lo sviluppo di applicazioni avanzate e favorire il trasferimento tecnologico verso il settore privato.

Proprio per quanto riguarda la NSF, gli stanziamenti in ambiti specifici prevedono 500 milioni per l'espansione degli *AI Research Institutes*; ovvero centri interdisciplinari dedicati all'IA distribuiti su tutto il territorio nazionale. Questa infrastruttura federale lanciata nel 2023 vuole fornire risorse computazionali e dataset *opensource* a università e start-up, in modo da democratizzare l'accesso alla potenza di calcolo e promuovere la competitività dell'ecosistema.

In questa direzione, il governo ha finanziato la costruzione di supercomputer specializzati in IA e iniziative di condivisione di dataset pubblici per l'addestramento di modelli di *machine learning*, al fine di ridurre la dipendenza da risorse private.

Il modello statunitense si caratterizza quindi per l'approccio stimolante e decentrato, dove la spesa pubblica non mira a sostituire quella privata ma a mobilitare investimenti complementari. Parallelamente il settore della difesa e della sicurezza nazionale rimane uno dei principali destinatari degli investimenti governativi, coerentemente con la strategia di mantenere un vantaggio tecnologico sugli avversari geopolitici. Proprio per questo l'agenzia DARPA ha ricevuto somme importanti destinate alla difesa avanzata pari a 600 milioni di dollari per lo sviluppo di nuovi programmi all'avanguardia.

Inoltre, va considerato che al Dipartimento della Difesa si attesta il 75% (Stanford University HAI, 2025) dei contratti riguardanti l'IA per portare

avanti iniziative come il CDAO[7] per integrare i sistemi di comando, la logistica militare e l'analisi predittiva dei conflitti.
Proprio per l'importanza che ha nella società statunitense l'apparato militare anche il Dipartimento per i Veterani ha ottenuto significativi investimenti, pari al 6,8% (Stanford University HAI, 2025), utilizzati per la ricerca riguardo diagnosi, protesi e salute mentale. Tutto questo insieme di politiche mostra come gli Stati Uniti abbiano consolidato nel tempo un modello di intervento pubblico basato su: investimenti mirati, cooperazione pubblico-privata e sostegno infrastrutturale.
Bisogna inoltre sottolineare come il settore della difesa sia terreno fertile per la messa a punto di nuove tecnologie, spesso con anche applicazioni civili. Sicuramente l'investimento pubblico nella ricerca di base e in quella militare favorisce gli attori privati che svolgono ricerche più avanzate. Quest'ultimi contribuiscono enormemente alla crescita economica assicurandosi di preservare la leadership tecnologica e industriale nel lungo periodo.

1.3.2. Unione Europea

Nel contesto europeo, gli investimenti pubblici per l'intelligenza artificiale si inseriscono in una strategia più ampia volta in primis a colmare il divario tecnologico con Stati Uniti e Cina e in secondo luogo a garantire la sovranità digitale dell'Unione. La Commissione Europea ha definito le linee guida di questa politica con il *Coordinated Plan on Artificial Intelligence*, aggiornato nel 2021, che mira a rafforzare la ricerca, stimolare l'adozione industriale e promuovere un'IA affidabile ed etica.
L'UE si è posta l'obiettivo di mobilitare almeno 20 miliardi di euro all'anno di investimenti pubblici e privati combinati nel settore entro il 2030 (CE, 21

[7] Chief Digital and Artificial Intelligence Office per coordinare le varie forze armate e agenzie nella messa a punto di tecnologie basate sull'IA.

aprile 2021), attraverso una combinazione di fondi europei diretti, risorse nazionali e cofinanziamenti privati. Solo tramite il programma *Horizon Europe*, il principale strumento dell'UE per il finanziamento della ricerca, sono stati destinati oltre 2,1 miliardi di euro a progetti riguardanti l'IA nel periodo 2021-2027.

Inoltre, il *Digital Europe Programme* ha stanziato circa 2,5 miliardi di euro per la creazione di infrastrutture digitali e piattaforme di intelligenza artificiale su tutto il territorio (CE, n.d.). Un elemento distintivo dell'azione europea è l'attenzione alla dimensione etica, regolatoria e sociale dell'IA. A differenza degli approcci statunitense e cinese, l'UE tenta di investire non solo nello sviluppo tecnologico, ma anche nella costruzione di un quadro normativo solido.

In particolare, con l'AI Act (Parlamento europeo, 2024), si tenta di stabilire degli standard di sicurezza, trasparenza e responsabilità, sebbene sia opinione diffusa che con questo atto si potesse fare di più vista la pervasività di questa tecnologia e le sfide che pone alla società.

È infatti interessante sottolineare come nonostante l'attenzione destinata dall'Unione Europea verso questioni ambientali ed etiche, all'interno di questa norma non si trovino riferimenti all'impatto ambientale dell'IA. Tale discrepanza si spiega tramite l'azione di lobbying volta ad erodere la regolamentazione di un settore ancora in via di sviluppo per non scoraggiarne gli investimenti.

Un'ulteriore prova dello svuotamento di significato dell'AI Act la si trova nelle pressioni del governo francese sulla Commissione Europea in modo da favorire Mistral, unica vera start-up europea operante nel campo dell'IA generativa, e le big tech come Microsoft con cui la realtà francese ha concluso un accordo nel 2024. Quadro normativo a parte, che ritroveremo nel terzo capitolo, consideriamo la questione degli investimenti pubblici in Europa. Parte di questi fondi viene destinata al sostegno dell'implementazione delle nuove normative, oltre che per il finanziamento di centri di eccellenza

specializzati e alla creazione di *testbed*[8] europei per la sperimentazione di tecnologie AI in ogni settore.

L'Unione finanzia inoltre le *European Digital Innovation Hubs* (CE, 2021), ovvero una rete di poli tecnologici diffusi in tutti gli Stati membri con la funzione di supportare piccole e medie imprese e pubbliche amministrazioni nell'adozione dell'intelligenza artificiale e di favorire il trasferimento tecnologico dalla ricerca all'industria.

L'approccio europeo si basa su una forte cooperazione multilivello tra istituzioni comunitarie, governi nazionali, realtà locali e attori privati, ed è caratterizzato da una distribuzione geografica estesa e capillare degli investimenti. Tuttavia, nel sistema europeo persistono delle criticità strutturali come: la frammentazione del mercato unico, la complessità delle procedure di finanziamento e la minore disponibilità complessiva di capitale pubblico rispetto ad altri concorrenti globali.

Nonostante queste sfide, l'Unione sta registrando un'accelerazione significativa nella spesa per IA: secondo i dati della Commissione, i fondi destinati alla ricerca e sviluppo nell'intelligenza artificiale sono aumentati di oltre il 70% dal 2018 al 2020 (Joint Research Centre Technical Report, 2022), inoltre secondo AI Watch/JRC, nel 2020 gli investimenti in IA nell'UE erano composti per circa 33% fondi pubblici e 67% privati (JRC Technical Report, 2022); la Commissione indica inoltre che dal 2021 *Horizon Europe* ha già destinato oltre 8 miliardi di euro a iniziative IA.

Questo modello, incentrato su un'IA sicura, affidabile e soprattutto etica, riflette la volontà dell'UE di costruire un ecosistema tecnologico competitivo ma coerente con i propri valori fondamentali e con le esigenze del proprio mercato interno.

[8] Un ambiente di prova controllato messo a disposizione per sperimentare, validare e confrontare tecnologie, prototipi o soluzioni prima dell'adozione in produzione o su larga scala.

1.3.3. Cina

La Cina rappresenta oggi il caso più emblematico di come gli investimenti pubblici possano diventare il motore principale di una strategia nazionale per l'intelligenza artificiale. Il governo di Pechino ad oggi considera l'IA un pilastro della propria politica industriale e tecnologica, come stabilito nel *New Generation Artificial Intelligence Development Plan* (Consigli di Stato cinese, 2017), pubblicato nel 2017, che fissa l'obiettivo di trasformare la Cina nel principale hub mondiale dell'intelligenza artificiale entro il 2030. Questo piano fa da cornice strategica per una vasta gamma di iniziative governative e di strumenti finanziari, che combinano fondi centrali, risorse provinciali e capitali pubblico-privati.

Secondo i dati dell'*OECD.AI Policy Observatory* e del *China Academy of Information and Communications Technology*, la spesa pubblica cinese in ricerca e sviluppo legata all'IA ha superato i 14 miliardi di dollari annui negli ultimi anni, registrando un tasso di crescita medio dell'8,7% tra il 2020 e il 2023 e raggiungendo circa il 96% della spesa statunitense se considerata in termini di parità di potere d'acquisto (Yamashita et al., 2021).

L'intervento pubblico in Cina si distingue per l'approccio fortemente centralizzato e pianificato, dove lo Stato svolge un ruolo attivo nel finanziamento diretto della ricerca, ma soprattutto nell'orientamento delle priorità scientifiche e industriali. I fondi pubblici vengono infatti canalizzati secondo vari step: attraverso programmi di ricerca strategica, verso grandi progetti infrastrutturali come centri nazionali per l'intelligenza artificiale e infine poli industriali specializzati.

Un ruolo sempre più rilevante è svolto dai fondi guida statali, strumenti finanziari pubblici creati per catalizzare investimenti privati e sostenere la crescita dell'ecosistema: tra questi, il *National AI Industry Investment Fund*, annunciato nel 2025, (Ministero dell'Industria e della Tecnologia cinese,

2025) e un nuovo fondo di venture capital da circa 1.000 miliardi di yuan[9] destinato allo sviluppo di tecnologie emergenti *hard tech*, tra cui intelligenza artificiale, semiconduttori e quantum computing. Questi fondi agiscono come moltiplicatori di capitale, mobilitando risorse aggiuntive da parte delle imprese e incentivando la partecipazione del settore privato a progetti strategici.

L'intervento statale non si limita al livello centrale ma si articola anche su scala locale e municipale, dove città come Pechino, Shanghai, Shenzhen e Hangzhou hanno varato propri piani per l'IA con budget dedicati e incentivi fiscali per attrarre talenti, start-up e investitori privati (Municipality of Hangzhou, 5 giugno 2025).

A Pechino, ad esempio, il programma municipale per l'IA prevede un investimento pubblico diretto di oltre 13 miliardi di yuan entro il 2025 (Yicai Global, 28 febbraio 2025), mentre Shanghai ha istituito il proprio fondo speciale per la creazione di un hub globale di innovazione nell'IA. Questa combinazione di politiche centrali e locali consente alla Cina di costruire un ecosistema altamente coordinato e integrato, nel quale ricerca pubblica, università, grandi imprese statali e settore privato collaborano per raggiungere gli obiettivi strategici fissati dal governo.

La scala e la rapidità dell'intervento cinese rappresentano oggi un fattore competitivo di primo piano nel panorama globale. Tuttavia, questo modello centralizzato presenta anche delle criticità, tra cui la minore trasparenza sull'allocazione delle risorse, il rischio di sovrapposizioni tra programmi locali e nazionali e la tendenza a finanziare un numero elevato di progetti a basso contenuto innovativo per soddisfare obiettivi quantitativi imposti dall'alto.

Nonostante ciò, la Cina continua a rafforzare il proprio ruolo come principale competitor degli Stati Uniti nel campo dell'intelligenza artificiale, e

[9] Circa 140 miliardi di dollari.

l'ampiezza delle risorse pubbliche mobilitate suggerisce che il divario tecnologico tra i due paesi potrebbe ridursi ulteriormente entro la fine del decennio.

1.4. Gli investimenti privati

Un altro indicatore fondamentale per comprendere appieno l'attività innovativa di un settore è rappresentato dagli investimenti privati che esso riesce ad attrarre. Nel caso dell'IA, questi sono cresciuti in modo straordinario: negli ultimi dieci anni sono aumentati di tredici volte, raggiungendo nel 2024 i 252,3 miliardi di dollari (Stanford University HAI, 2025), con l'incremento maggiore nella componente privata.

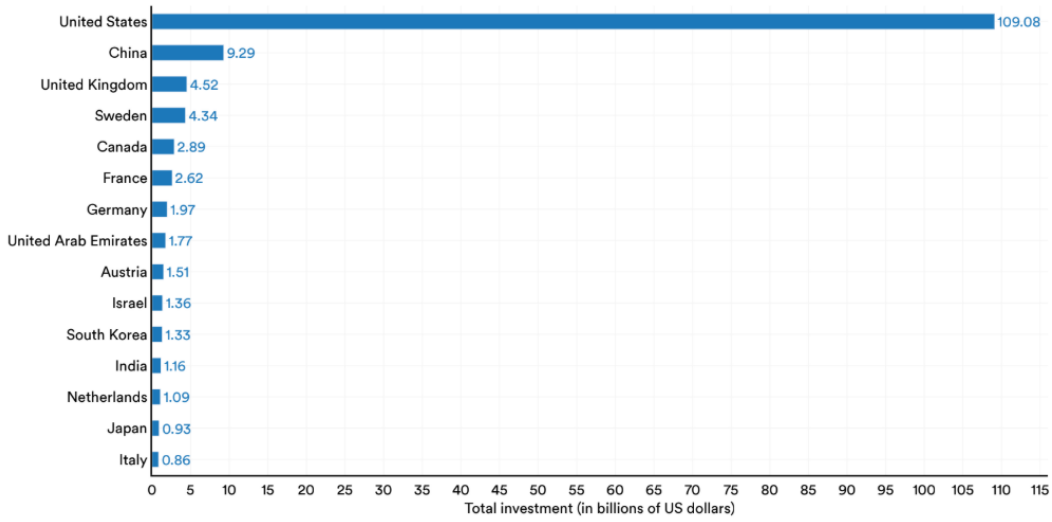
Gli investimenti privati hanno registrato un aumento del 44,5% tra il 2023 e il 2024, segnando così la prima crescita annuale dal 2021 (Stanford University HAI, 2025). In particolare, i finanziamenti destinati alla IA generativa hanno continuato a crescere rapidamente. Nel 2024 il settore ha attratto 33,9 miliardi di dollari (Ding et al., 2025), con un aumento del 18,7% rispetto al 2023 e oltre 8,5 volte rispetto al 2022.

La GenAI ha rappresentato più di un quinto di tutti gli investimenti privati correlati all'IA nel 2024. Questi forti flussi di capitale hanno favorito la nascita di un gran numero di nuove aziende e start-up altamente innovative. Solo nel 2024 sono state fondate 2.049 nuove imprese di IA, con un aumento dell'8,4% (Stanford University HAI, 2025) rispetto all'anno precedente. Particolarmente significativa è anche la crescita delle start-up di IA generativa, salite a 214 nel 2024, rispetto alle 179 del 2023 e alle appena 31 del 2019.

Analizzando la distribuzione degli investimenti privati su base nazionale, gli Stati Uniti si confermano il principale polo mondiale. Nel 2024 hanno registrato 109,1 miliardi di dollari di investimenti, un valore 11,7 volte

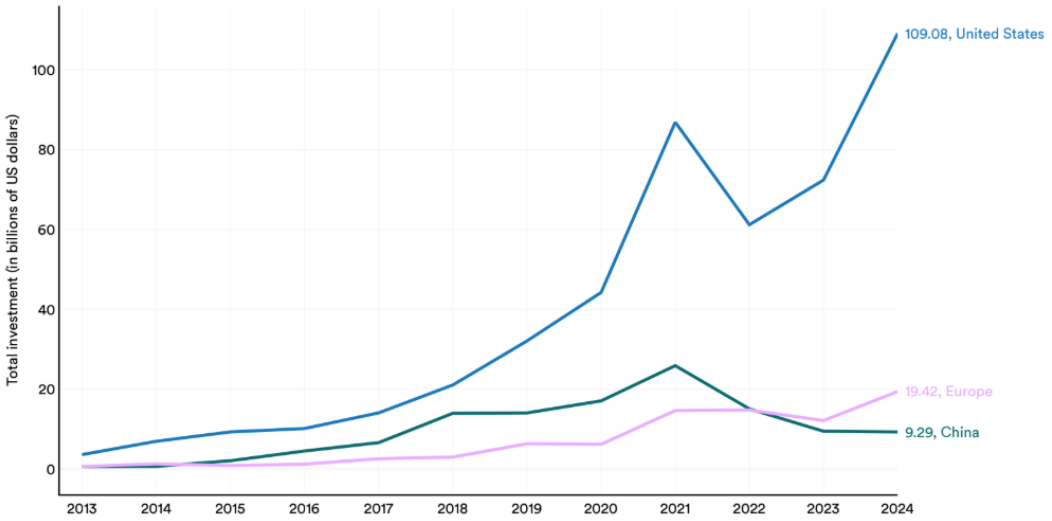
superiore a quello del secondo paese, la Cina, con 9,3 miliardi, e 24,1 volte superiore a quello del Regno Unito, con 4,5 miliardi (Stanford University HAI, 2025). Altri Paesi che hanno attratto capitali significativi nell'ultimo decennio includono Israele (15 miliardi di dollari), Singapore (7,3 miliardi) e Svezia (7,3 miliardi).

Grafico 1.5 - Investimenti privati per Paese (Stanford University HAI, 2025).



L'analisi dei trend recenti mostra come il divario negli investimenti privati tra Stati Uniti e resto del mondo stia crescendo rapidamente. Mentre gli investimenti privati in IA sono diminuiti in Cina dell'1,9% e aumentati in Europa del 60% dal 2023, gli Stati Uniti hanno registrato un incremento del 50,7% nello stesso periodo e del 78,3% rispetto al 2022. La disparità è ancora più evidente nel campo della IA generativa: nel 2024, gli Stati Uniti hanno superato la somma combinata degli investimenti di Cina ed Europa in GenAI di circa 25,4 miliardi di dollari.

Grafico 1.6 - Andamento degli investimenti privati 2013-2023 (Stanford University HAI, 2025).



Sebbene la GenAI rappresenti una quota importante, anche altre aree dell'IA continuano ad attirare grandi volumi di capitale: 37,3 miliardi di dollari (Stanford University HAI, 2025) sono stati destinati alla ricerca su infrastrutture e governance dell'IA, con investimenti significativi verso aziende come OpenAI, Anthropic e xAI, focalizzate sullo sviluppo di applicazioni avanzate; 16,6 miliardi di dollari sono stati investiti nel miglioramento della gestione ed elaborazione dei dati, con start-up come Lambda, Impetus Technologies e Databricks; infine, 11 miliardi di dollari sono confluiti nel settore sanitario, finanziando aziende come Tempus, Abridge e Innovaccer negli Stati Uniti, e Baichuan in Cina, quest'ultima sostenuta da Alibaba e attiva nelle tecnologie mediche.

1.5. Le start-up

Le start-up rappresentano una componente cruciale dell'ecosistema innovativo, poiché incarnano la capacità di un sistema economico di trasformare la conoscenza in valore economico. Questa attività nevralgica si

spiega grazie alle caratteristiche delle start-up in primis, ma anche del contesto in cui operano. Unendo fattori strutturali, economici e tecnologici alla spinta innovatrice di queste realtà si ottiene ciò che Schumpeter vedeva come motore dell'attività innovativa: "il processo di distruzione creativa".
È proprio tramite l'introduzione di innovazioni radicali, che vanno a "sfidare" le grandi aziende già affermate, spingendo per il rinnovamento tecnologico e competitivo; che le start-up diventano entità fondamentali per il processo di sviluppo di un settore.
Oltre ad essere quindi motore dell'avanzamento tecnologico, queste piccole società fanno da ponte tra l'attività di ricerca accademica e il mercato come si può verificare nell'*European Innovation Scoreboard* (2025) della Commissione Europea. Tanto che il numero di start up nate in ambiti accademici o da centri di ricerca, è un indicatore proxy dell'efficienza dei sistemi di innovazione.
La trasformazione di risultati scientifici in prodotti e processi scalabili permette di catalizzare gli investimenti privati, contribuendo alla formazione e all'attrazione di capitale umano specializzato.
Grazie a questa dinamica si ha come effetto ulteriore quello di spillover a livello territoriale, in grado di assicurare il trasferimento di conoscenze tra aziende e settori aumentando così la capacità innovativa di una regione (CE, 2025). Proprio nello studio del numero di start-up, della loro longevità e nella quantità di capitale attratto, si ottengono indicatori quantitativi e qualitativi indispensabili per lo studio dell'innovazione e dello sviluppo economico.

1.5.1. Stati Uniti

Il panorama dei finanziamenti all'IA negli Stati Uniti è caratterizzato da un'abbondanza di capitale senza paragoni, dalla frequente presenza di mega-round, da un ecosistema di venture capital altamente dinamico e da una

generale vivacità del tessuto imprenditoriale. Questi elementi permettono alle start-up di crescere rapidamente e di alimentare un'intensa attività innovativa. Le start-up statunitensi nel campo dell'IA mostrano una notevole capacità di attrarre finanziamenti superiori ai 100 milioni di dollari, grazie a una combinazione di modelli di business scalabili, differenziazione tecnologica e forte trazione sul mercato.

Nel 2025 le aziende legate all'intelligenza artificiale hanno raccolto 118 miliardi di dollari fino al 15 agosto, mostrando un significativo aumento rispetto ai 108 miliardi di dollari dell'intero 2024 (Crunchbase News, 20 agosto 2025). I finanziamenti al settore sono più che raddoppiati nello stesso periodo. Anche le percentuali sono in aumento: il 48% dei finanziamenti di venture capital globali dall'inizio dell'anno è stato investito in aziende legate all'intelligenza artificiale, rispetto a un terzo nel 2024. Di questi 118 miliardi di dollari, otto aziende hanno raccolto complessivamente 73 miliardi di dollari tramite round da un miliardo di dollari, rappresentando il 62% dei finanziamenti destinati alle aziende del settore.

Tra i casi più rilevanti: OpenAI, con un round impressionante da 40 miliardi di dollari nel marzo 2025; Databricks, con 1 miliardo raccolto nel 2025; Anthropic, con 16,5 miliardi nel 2025 e Infinite Reality, con 3 miliardi a gennaio 2025 (Crunchbase News, 3 aprile 2025). Il divario tra il mega-round di OpenAI e gli altri competitor evidenzia la formazione di una dinamica da "*winner-takes-most*", in cui gli investitori concentrano capitali enormi sui leader già affermati, piuttosto che distribuire i fondi tra più concorrenti. Grazie a questa dinamica possiamo notare come il grado di maturità di questo settore stia aumentando e come stiano emergendo aziende leader sempre più solide. La California, e in particolare la *Bay Area*, continua a svolgere un ruolo predominante nella raccolta di finanziamenti per l'IA, dimostrando che, nonostante le tendenze verso il lavoro da remoto e l'aumento dei costi operativi, i vantaggi dell'ecosistema locale rimangono fortemente attrattivi per start-up e investitori (Brookings Metropolitan Policy Program, 2021). Un

altro elemento di rilievo è la rapidità con cui le nuove aziende riescono a raccogliere capitali significativi: molte di quelle fondate tra il 2022 e il 2024 hanno ottenuto round a nove cifre entro 1 o 2 anni dal lancio, a dimostrazione sia dei cicli di sviluppo accelerati sia della disponibilità degli investitori a scommettere presto su progetti promettenti.

Le società focalizzate sull'IA generativa, tra cui OpenAI, Anthropic e Runway, hanno attirato la maggior parte dei finanziamenti, riflettendo l'entusiasmo e le aspettative elevate attorno a questo settore. Un aspetto chiave è anche la partecipazione strategica di grandi player tecnologici come NVIDIA e Google nel capitale di start-up di IA. Questi investimenti rivelano strategie di lungo periodo volte a garantire che i nuovi attori del mercato utilizzino le loro infrastrutture hardware e i loro servizi.

Tali operazioni non sono solo investimenti, ma veri e propri strumenti per creare partnership strategiche e rafforzare l'interconnessione tra start-up e aziende già affermate. Tutti questi elementi indicano un settore dell'IA ormai in fase di maturazione, in cui la specializzazione e l'integrazione con settori tradizionali diventano prioritarie. La concentrazione di capitali su pochi attori suggerisce che gli investitori stiano diventando più selettivi, privilegiando aziende con tecnologie consolidate, strategie chiare e piani di business ben definiti per la scalabilità e monetizzazione su larga scala.

Con l'interesse per l'IA in costante crescita, la soglia per accedere ai cosiddetti "mega-round" continua a salire. Le start-up che sapranno puntare su piattaforme scalabili, proprietà intellettuale difendibile e pratiche di IA responsabile saranno le più adatte ad attrarre grandi investimenti e a guidare trasformazioni significative in molteplici settori. Questa evoluzione è ben rappresentata dal cambiamento osservato a partire dal 2024, quando la maggior parte dei capitali si è spostata verso la *late-stage* e la *growth stage* delle aziende, ossia le fasi che precedono l'uscita sul mercato e la quotazione in borsa.

Il trend riflette la rapida maturazione del settore: i primi "unicorni" dell'IA

stanno consolidando la loro posizione di leadership e attirando i maggiori flussi di capitale, confermando il ruolo centrale degli Stati Uniti come principale motore globale dell'innovazione nel campo dell'intelligenza artificiale.

1.5.2. Unione Europea

Per comprendere il contesto europeo dell'innovazione nel settore dell'IA è utile partire dal rapporto sulla competitività dell'Unione Europea redatto da Mario Draghi, secondo cui l'UE si trova oggi a scontare il ritardo accumulato nella precedente rivoluzione tecnologica, quella di Internet. Proprio a causa di questa mancata leadership, l'Europa conta oggi solo quattro aziende tra le prime cinquanta al mondo nel settore tecnologico.

Come sottolinea l'ex presidente della BCE, *"l'Europa deve rifocalizzare profondamente i propri sforzi collettivi per colmare il divario di innovazione con Stati Uniti e Cina"*. Proprio a causa di questo mancato dinamismo Draghi continua nel suo discorso di presentazione del report al Parlamento Europeo di Strasburgo: "*Di conseguenza, molti imprenditori europei preferiscono cercare finanziamenti da venture capitalist statunitensi e ampliare la propria attività sul mercato statunitense. Tra il 2008 e il 2021, quasi il 30% degli "unicorni" fondati in Europa, ovvero startup con un valore di mercato superiore a 1 miliardo di dollari, ha trasferito la propria sede all'estero*".

Nonostante queste premesse, l'ecosistema europeo dell'innovazione è in crescita, sebbene con caratteristiche peculiari. Più che nello sviluppo di modelli di inferenza, hardware o infrastrutture, l'Europa eccelle nell'applicazione dell'IA per la trasformazione dei processi e dei flussi di lavoro in vari settori.

Negli ultimi anni sono nate decine di start-up in tutto il continente con l'obiettivo di riprogettare operazioni e servizi grazie all'intelligenza

artificiale. L'ecosistema europeo si distingue inoltre per la sua diffusione geografica e istituzionale: esistono circa 150 hub di innovazione, inseriti nel programma *European digital innovation hubs,* che spaziano da centri legati alle università a iniziative completamente commerciali. Questi hub hanno contribuito a far nascere imprese di rilievo globale come Klarna, Celonis, Hugging Face e Isar Aerospace, attive rispettivamente nei settori della finanza, del software, dello spazio e dell'intelligenza artificiale.

Un esempio significativo è l'*UnternehmerTUM* di Monaco di Baviera, fondato nel 2002 da Susanne Klatten per promuovere una nuova cultura imprenditoriale in Germania. Associato alla *Technische Universität München*, l'hub può contare su un'ampia rete di scienziati, imprenditori e investitori in grado di supportare un'azienda dal lancio fino alla quotazione in borsa. Ha incubato oltre 1.000 imprese, tra cui il gruppo di trasporti FlixMobility e la start-up di IA Konux. Sempre a Monaco opera Start2 Group, strettamente legato al Ministero dell'Economia tedesco e caratterizzato da una forte vocazione internazionale, con presenza in 18 Paesi e filiali negli Stati Uniti e in Asia.

In Francia, uno degli hub più dinamici è Station F, fondato nel 2017 da Xavier Niel. Con oltre 1.000 start-up attive, tra le 40 più promettenti ben 32 hanno l'IA al centro del proprio modello di business.

Nel Regno Unito invece *Founders Factory* si distingue per il modello ibrido di venture builder e investitore early-stage, collaborando con circa 60 grandi partner aziendali su quattro continenti.

Nel complesso, l'Europa sta compiendo progressi significativi nell'innovazione legata all'IA, ma continua ad affrontare sfide strutturali importanti: un contesto di finanziamenti più cauto, una regolamentazione più complessa e un mercato più frammentato rispetto a quello statunitense. Questi fattori si traducono in investimenti complessivamente inferiori e in un numero minore di round di finanziamento di grandi dimensioni, rallentando in parte la crescita dell'ecosistema. Come per lo sviluppo di Mistral AI, una

start-up francese specializzata in IA generativa, nota per aver creato un modello LLM di dimensioni ridotte ma estremamente preciso. Questa realtà ha dimostrato come, partendo da un'architettura compatta, che vede sette miliardi di parametri contro i cento trilioni di GPT-4, sia possibile ottenere un modello efficiente e con costi di sviluppo e inferenza più contenuti. Nonostante l'ottimismo che circonda Mistral, il radicamento nel contesto europeo comporta alcuni limiti strutturali. Pur avendo ricevuto ingenti investimenti e sostegno governativo dalla Francia, l'azienda ha dovuto comunque stringere accordi con i grandi attori internazionali, ottenendo finanziamenti da Eric Schmidt, ex presidente di Google, e siglando un importante accordo con Microsoft nel 2024. Ad oggi, Mistral è valutata circa 11 miliardi di euro e detiene circa il 2% del mercato globale del settore, secondo i dati di OpenRouter[10]. Tuttavia, in un contesto competitivo dove, ad esempio, Meta ha investito 14 miliardi di dollari per acquisire il 49% di Scale AI (Reuters, 13 giugno 2025), la start-up francese rimane una piccola realtà in mezzo a colossi globali.

1.5.3 Cina

Nel corso del 2025 il settore dell'IA in Cina ha attraversato profondi cambiamenti strategici, soprattutto per quanto riguarda l'approccio delle principali aziende e start-up. Questi mutamenti sono in larga parte riconducibili alla posizione di leadership assunta da DeepSeek, che ha ricevuto anche l'endorsement ufficiale del governo cinese.

Il lancio del modello R1 ha infatti spinto molte imprese concorrenti a rivedere le proprie strategie, riallocando fondi e investimenti per adattarsi al nuovo contesto competitivo. In particolare, numerose aziende hanno scelto di

[10] Bisogna tenere conto del fatto che OpenRouter misura l'attività solamente sopra la propria piattaforma, per cui il dato citato rimane una proiezione e non un valore ufficiale di mercato.

concentrarsi sullo sviluppo di applicazioni basate sul modello DeepSeek, piuttosto che sull'addestramento di modelli di base proprietari. Un esempio significativo è 01.ai, che ha annunciato l'intenzione di vendere soluzioni di IA personalizzate per le imprese, puntando sulla propria esperienza nel paradigma dei *mixture of experts* (MoE) come vantaggio competitivo. A differenza dell'addestramento di un unico modello "denso" su vasti dataset raccolti dal web, la metodologia MoE combina più modelli più piccoli, ciascuno addestrato su dati specifici di settore.

Questo approccio consente così alle aziende con risorse hardware limitate, di addestrare modelli complessi con una potenza di calcolo ridotta. Anche la già citata Baichuan ha modificato la propria strategia, decidendo di ridurre gli investimenti nelle applicazioni finanziarie dell'IA per focalizzarsi sulle tecnologie destinate al settore sanitario e al supporto alla diagnosi medica. Altre due realtà di rilievo, Moonshot e Zhipu, continuano a inseguire DeepSeek sul terreno dell'addestramento di modelli, pur introducendo aggiustamenti strategici: Moonshot ha tagliato le spese di marketing legate al proprio chatbot Kimi, mentre Zhipu ha mantenuto diverse linee di business, lanciando sia applicazioni *consumer*, sia un ramo *enterprise* che offre soluzioni di IA personalizzate a governi locali e imprese. Quest'ultimo segmento rappresenta un mercato altamente competitivo e a margini ridotti in Cina.

1.6. Conclusioni

In questo capitolo abbiamo cercato di ricostruire il settore dell'intelligenza artificiale come un ambito innovativo in fortissima espansione, mettendone in evidenza alcuni elementi centrali. L'analisi delle pubblicazioni scientifiche mostra non solo la rapidità con cui cresce la ricerca in questo campo, ma anche il carattere multidisciplinare del fenomeno e la competizione sempre

più marcata tra aree geografiche. Lo studio dei brevetti evidenzia come questa corsa all'innovazione si traduca in una competizione industriale per l'appropriazione e il controllo della conoscenza tecnologica.
Infine, l'esame degli investimenti pubblici e privati, insieme al ruolo delle start-up, permette di osservare come l'IA si stia consolidando come settore strategico, capace di attrarre capitali crescenti e di occupare una posizione sempre più centrale nelle politiche industriali e nelle strategie di sviluppo. Nel complesso, emerge quindi l'immagine di un settore dinamico, competitivo e destinato ad ampliare ulteriormente il proprio raggio d'azione in numerosi ambiti economici e produttivi.
Tuttavia, proprio questa espansione solleva un interrogativo cruciale, che non può essere eluso: se l'IA si configura come un'innovazione radicale capace di ridefinire gli equilibri economici e tecnologici globali, diventa necessario interrogarsi anche sulle condizioni materiali che ne rendono possibile lo sviluppo. L'aumento della diffusione di questa tecnologia comporterà inevitabilmente una crescita della domanda di capacità computazionale, di infrastrutture e di energia e risorse naturali; rendendo sempre più urgente una riflessione sulla sostenibilità complessiva di questo paradigma tecnologico. È proprio su questa tensione, tra accelerazione dell'innovazione e crescita dei consumi, che si innesta il secondo capitolo, dedicato ad analizzare i costi ambientali e infrastrutturali dell'intelligenza artificiale, andando oltre la narrazione della sua apparente immaterialità.

2. L'impatto ambientale dell'Intelligenza artificiale

Le economie e le società di tutto il mondo stanno affrontando due grandi sfide globali: la transizione verde da un lato e quella digitale dall'altro. L'avvento di una tecnologia di uso generale come l'IA deve quindi essere visto come

un'opportunità per sfruttare la transizione digitale per un futuro e uno sviluppo sostenibile. Considerare queste due direttrici simultaneamente può essere di aiuto per affrontare una moltitudine di settori e aspetti differenti, che tramite tecnologie integrate a IA, possono aumentare l'efficienza e ridurre i costi ambientali e le emissioni.

Per quanto concerne invece l'impatto dell'IA stessa, la questione diventa critica: i dati riguardanti l'impatto diretto e indiretto di questo tipo di tecnologia sono spesso, ancora imparziali o non del tutto autorevoli venendo principalmente dalle dichiarazioni delle aziende coinvolte, il tutto accompagnato da un quadro normativo tutt'altro che completo.

In questo lavoro l'attenzione è volutamente concentrata sugli impatti materiali dell'IA, cercando quando possibile di affidarsi a fonti che indaghino la materialità dei problemi. Solo tramite uno studio effettivo e basato su numeri reali, e non su stime e proiezioni, si potrà ottenere una IA etica e sostenibile. Pertanto, costruire una metodologia di studio del settore diviene fondamentale: concentrarsi su emissioni dirette e indirette di tutte le fasi di "produzione" di questa tecnologia, analizzare le esternalità positive e negative e infine sostenere questa grande opera con una serie di provvedimenti che ne regolino l'attività.

Per quanto riguarda le *policies* e la regolamentazione ne discuteremo nel dettaglio nel prossimo capitolo, per ora ci focalizzeremo su quelli che sembrano essere i fattori più impattanti sull'ambiente: dalla produzione dell'hardware e l'allenamento degli algoritmi, fino all'utilizzo da parte dei consumatori e alla quantità di acqua consumata nei datacenter.

2.1 La componente *embodied* dell'impatto ambientale

A proposito della componentistica hardware per l'IA ci troviamo a descrivere prodotti, processi e infrastrutture con scale di misura che variano moltissimo,

possiamo infatti passare dal nanometro del transistor ai chilometri dei cavi elettrici che collegano i datacenter alla rete. All'interno dell'hardware partiamo con l'analizzare la componentistica più essenziale, grazie alla quale è stata messa a punto la tecnologia dell'IA.

Partiamo con la CPU che gestisce tutte le operazioni logiche e di controllo e coordina il lavoro delle altre unità (come GPU, memoria e dischi); in particolare avvia i processi di calcolo, distribuendo il lavoro a GPU o TPU, e gestisce i dati di input e output. Passiamo quindi alle GPU, create per gestire la grafica dei videogiochi e capaci di svolgere milioni di calcoli in parallelo. Quest'ultima caratteristica è ciò che le ha rese perfette per l'IA: in particolare per addestrare le reti neurali con dei dataset e per eseguire le inferenze, ovvero per il funzionamento vero e proprio del modello addestrato.

Mentre le CPU hanno pochi core (unità di calcolo) molto intelligenti, le GPU si basano sulla quantità e la semplicità, che permette di svolgere operazioni matematiche ripetitive come sommare e moltiplicare matrici di numeri in poco tempo. Grazie a questo "esercito" di piccole unità coordinate è possibile gestire trilioni di operazioni in pochi giorni quando una CPU a parità di calcoli impiegherebbe anni. Un ulteriore passo avanti rispetto alle GPU sono le TPU, un tipo di chip progettato da Google, e i chip ASIC: entrambi componenti specializzate nella gestione di matrici e vettori per l'IA. Questa specializzazione rende meno flessibili le TPU e ancor meno gli ASIC, ma garantisce di ridurre il consumo energetico per la gestione di enormi quantità di calcoli, migliorando ulteriormente l'efficienza. Allo stesso tempo essa introduce anche vincoli significativi in termini di flessibilità e dipendenza da specifiche architetture proprietarie[11], un aspetto spesso sottovalutato nel dibattito sull'IA sostenibile, ma che ne lede la trasparenza e la possibilità di eseguire valutazioni ambientali comparative.

[11] soluzioni hardware o software sviluppate e controllate da singoli attori industriali, il cui funzionamento interno non è completamente accessibile al pubblico.

Altra componente è la memoria: divisa in HBM, ad alta velocità posta vicino alle TPU/GPU per accedere in modo immediato ai dati, e la DRAM, una memoria di lavoro generale della CPU impiegata per la gestione dei dataset e delle operazioni generali. Infine, troviamo l'HDD ovvero l'hard disk drive, dove vengono memorizzati dataset, modelli e log di allenamento in modo permanente[12].

Oltre a queste parti fondamentali, l'hardware si compone anche di tutti quegli elementi di supporto: dalla PCBA, la scheda verde con piste di rame su cui sono montati chip, resistori e connettori; alle componenti termiche come ventole, dissipatori o piastre termiche; o ancora dagli elementi meccanici ed elettromeccanici come telai, alimentatori e cavi elettrici. Oltre agli elementi citati chiaramente ve ne sono ulteriori, ma la panoramica ci aiuta a comprendere quanta complessità richieda stimare un impatto sull'ambiente di questo tipo di attività.

Proprio la difficoltà tecnica di questo tipo di studi rende evidente come ogni tentativo di sintesi dell'impatto ambientale dell'hardware IA rischi di semplificare eccessivamente una realtà fatta di processi diversissimi e difficilmente comparabili.

La produzione di tutti questi elementi per l'intelligenza artificiale si basa sull'estrazione fisica e sul consumo di risorse naturali per costruire hardware, inclusi chip, semiconduttori, GPU e CPU. La produzione di hardware e componenti si articola con diverse fasi: dall'estrazione mineraria, alla fusione e raffinazione, fino alla produzione di componenti e l'assemblaggio. L'impatto ambientale lungo questa catena del valore non si limita all'emissione di anidride carbonica o di altri gas climalteranti, ma include problematiche dirette come: la contaminazione del suolo, la deforestazione, l'erosione, il degrado della biodiversità, lo smaltimento di rifiuti tossici, l'inquinamento delle falde acquifere, l'uso dell'acqua, i rifiuti radioattivi e

[12] Per i server a IA si utilizzano gli SSD, solid state drive, più veloci ed efficienti per questo tipo di tecnologia.

l'inquinamento atmosferico.

Questo genere di esternalità negative a livello micro comporta una difficoltà intrinseca nel quantificare l'impatto dei siti di produzione, rendendo impossibile formare un dato aggregato comprensivo. Inoltre, l'impatto singolo di ogni fase della produzione viene spesso trascurato a causa della difficoltà di attribuzione (Henderson et al., 2020) per cui è molto complesso operare un *Life-Cycle-Assesment* dalla prima fase di produzione al fine vita. È importante evidenziare che, mentre l'uso crescente di energie rinnovabili e il miglioramento dell'efficienza energetica nei data center riducono la percentuale delle emissioni dirette in loco, aumenta invece la quota di emissioni attribuibile alla produzione dei componenti fisici, ovvero la componente indiretta, sull'impatto complessivo dell'intelligenza artificiale (Gupta et al., 2020). Pertanto, ci si aspetta che le fasi non operative, come quelle di estrazione, raffinazione e produzione delle componenti fisiche, diventino sempre più importanti nello studio della sostenibilità per l'IA (IEA, 2021). Attualmente, tuttavia, molti fornitori di server IA segnalano solo la componente operativa, escludendo quella derivante dalla produzione proprio a causa delle problematiche viste in precedenza.

Secondo una stima nel 2015 l'impronta di carbonio derivante dalla produzione globale di data center si attestava a 20 megatoni di CO_2e, pari al 15% delle emissioni totali del settore (Malmodin e Lundén, 2018). Più recentemente, Meta ha dichiarato che le emissioni derivanti dalla produzione dei propri datacenter rappresentano circa il 30% delle emissioni totali dell'azienda per il 2022 (Wu et al., 2022). In linea con le previsioni finora viste, si stima che la quota di emissioni derivante dalla produzione dei datacenter potrebbe aumentare fino a oltre l'80% (Gupta et al., 2020). Sempre secondo Meta, la loro quota di emissioni della catena del valore rispetto alle emissioni totali di gas serra è aumentata dal 44% nel 2017 al 99% nel 2020 (Meta, 2021), con un conseguente "significativo costo del carbonio incorporato pagato in anticipo" per i nuovi componenti di sistema basati

sull'IA per i data center (Wu et al., 2022). Analogamente, la quota di emissioni di gas serra della catena del valore secondo Google è aumentata dal 45% nel 2016 a oltre il 90% nel 2020 (Google, 2022).

Questi aumenti delle emissioni sono dovuti a un maggiore approvvigionamento di energia pulita, che riduce la percentuale di emissioni derivanti dal consumo diretto di energia.

Nonostante le numerose criticità che caratterizzano la catena produttiva dell'hardware per l'IA, il settore attraversa una fase di forte espansione, attirando ingenti flussi di investimento. Tali investimenti si traducono non solo in un aumento delle spese in ricerca e sviluppo, ma soprattutto nella costruzione e nell'ampliamento di infrastrutture dedicate.

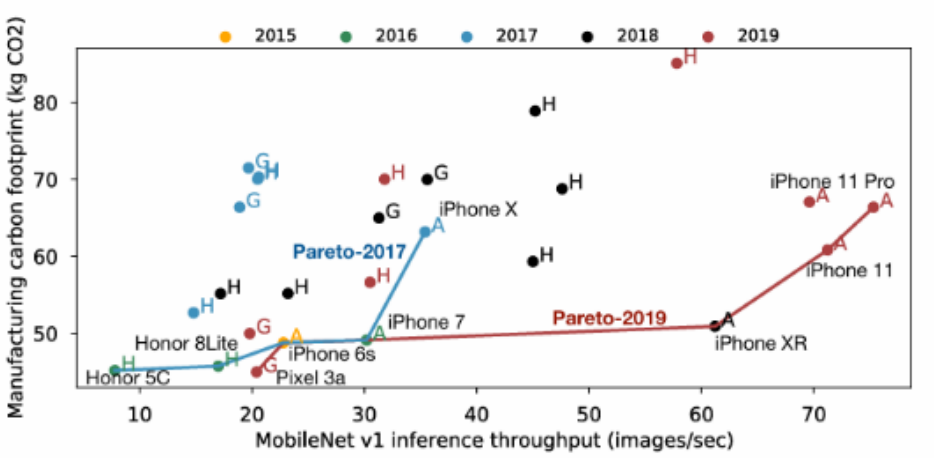
Con l'affermarsi dell'IA su larga scala, i principali attori del settore *high-tech* hanno infatti investito massicciamente nel potenziamento delle proprie infrastrutture hardware. Un esempio emblematico è rappresentato da Meta, che nel 2024 ha accelerato significativamente l'espansione della propria infrastruttura per l'IA generativa, annunciando due cluster da 24.576 GPU e l'obiettivo di raggiungere 350.000 GPU NVIDIA H100 entro la fine dell'anno (Meta, 2024).

Dal punto di vista ambientale, la fase più impattante rimane la costruzione dell'infrastruttura fisica e la produzione delle componenti hardware più complesse. Per quanto riguarda i chip, sebbene essi diventino progressivamente più potenti ed efficienti dal punto di vista computazionale, la loro impronta ecologica tende ad aumentare a causa di processi produttivi estremamente sofisticati, ad alta intensità energetica e caratterizzati dall'uso di materiali critici. Ne deriva una tensione strutturale tra il miglioramento delle prestazioni e la riduzione dell'impatto ambientale, che non sempre evolve in modo bilanciato.

In questo contesto risulta particolarmente utile l'analisi proposta da Gupta et al. (2020) sull'elusività dell'impronta ecologica nel settore dell'informatica. Gli autori mostrano come, nel periodo compreso tra il 2017 e il 2019, le

innovazioni hardware si siano concentrate prevalentemente sull'aumento dell'efficienza e delle prestazioni, piuttosto che su una riduzione significativa dell'impatto ambientale delle componenti. Lo studio analizza il compromesso tra le prestazioni di inferenza e l'impronta di carbonio attraverso il concetto di ottimo paretiano, evidenziando come, nei due anni considerati, tale ottimo si sia spostato sensibilmente verso un aumento delle prestazioni, praticamente raddoppiate, a fronte di una riduzione marginale delle emissioni, pari a circa 3 kg di CO_2.

Grafico 2.7 - Ottimo Paretiano tra impronta ecologica e prestazioni. Gupta et al. (2020).



Sebbene l'analisi di Gupta et al. (2020) sia condotta su specifici *workload* di inferenza mobile, essa risulta particolarmente utile per mettere in luce una tendenza strutturale del progresso computazionale: lo spostamento dell'ottimo paretiano privilegia sistematicamente l'aumento delle prestazioni rispetto alla riduzione dell'impronta di carbonio.

Tale dinamica appare rilevante anche per l'evoluzione delle infrastrutture dedicate all'intelligenza artificiale su larga scala, dove l'incremento delle capacità di calcolo non è stato accompagnato da una riduzione proporzionale delle emissioni associate alla produzione dell'hardware. Nel complesso, questa evidenza suggerisce che gli sforzi orientati esclusivamente alla razionalizzazione dei processi e al miglioramento dell'efficienza consentano

di raggiungere capacità di calcolo sempre più elevate, ma risultino insufficienti ad affrontare in modo diretto il problema delle emissioni di CO_2 legate alla produzione delle componenti hardware.
Ne emerge quindi la necessità di un cambio di prospettiva, il concetto di ottimo paretiano, infatti, rischia di normalizzare un *trade-off* che dovrebbe essere messo in discussione a monte, intervenendo sulla progettazione e sulla domanda.

2.1.1. La catena di approvvigionamento

La gestione delle infrastrutture IA richiede quantità rilevanti di metalli e materiali specifici, tra cui rame, cobalto, litio, palladio, grafite e terre rare. L'estrazione di queste risorse è concentrata in un numero limitato di Paesi, come la Cina, la Repubblica Democratica del Congo, il Cile, il Brasile, l'Australia, la Russia, il Sud Africa, gli Stati Uniti e l'Indonesia.
In molti casi, soprattutto nei paesi ricchi di risorse naturali ma ancora in via di sviluppo industriale, le capacità di raffinazione e di trasformazione risultano limitate. Va tuttavia tenuto presente che molte delle stime disponibili su queste fasi produttive presentano margini di incertezza elevati, rendendo difficile un confronto diretto tra Paesi e tecnologie. Per quanto riguarda le fasi della catena del valore a più alto contenuto tecnologico come: la raffinazione, la produzione di componenti intermedi e la manifattura avanzata; queste tendono a concentrarsi in Paesi dotati di infrastrutture industriali più mature, come ad esempio Cina, il Giappone e la Corea del Sud. Questa distribuzione delle attività lungo la filiera dà origine a una catena di approvvigionamento fortemente interdipendente e complessa, esposta a fattori geopolitici, normativi e commerciali.
In tale contesto, la diversificazione sia delle fonti di approvvigionamento sia delle capacità di trasformazione viene spesso indicata come una strategia fondamentale per ridurre il rischio di interruzioni nelle forniture. Negli ultimi decenni, la Cina ha assunto un ruolo sempre più centrale nella lavorazione di

numerosi minerali critici, grazie alla forte domanda interna proveniente dall'industria manifatturiera e alla rapida crescita dei settori tecnologico e ingegneristico. Va comunque considerato come, nel dibattito pubblico occidentale, questa concentrazione venga rappresentata come una vulnerabilità strategica sorta improvvisamente, quando la realtà dei fatti dimostra che essa sia il risultato di precise scelte industriali compiute nel corso di diversi decenni.

Al rapido sviluppo interno verificatosi in Cina si è affiancata un'espansione all'estero delle sue imprese, incluse società a partecipazione statale come la Zijin Mining Group Ltd., che hanno investito in attività estrattive in diverse aree del mondo, come Africa e America Latina (Reuters, 21 maggio 2024). Secondo diverse stime IEA (23 ottobre 2025), la Cina detiene oggi una quota predominante della capacità globale di raffinazione delle terre rare e occupa una posizione di rilievo anche in altre filiere strategiche, come quelle della grafite, del litio, del rame e dell'alluminio; tutti materiali centrali per la transizione digitale ed energetica.

L'impatto ambientale della fase estrattiva varia in modo significativo a seconda delle tecniche adottate, della tipologia di giacimento e del contesto normativo del Paese in cui l'attività ha luogo. Nel caso delle terre rare, le specificità risultano particolarmente accentuate: questi elementi sono infatti presenti in basse concentrazioni nel suolo e, a causa delle loro proprietà chimiche simili, risultano difficili da separare. Ciò rende i processi di lavorazione più complessi, energivori e potenzialmente inquinanti.

Proprio in risposta a questi temi, negli ultimi anni sono stati compiuti alcuni passi avanti sia sul piano procedurale sia su quello normativo. In questo contesto si inserisce, ad esempio, il *Responsible Mining Index Report* (World Resources Forum, 2025), un'indagine che ogni due anni coinvolge quaranta aziende minerarie operanti a livello globale e che offre una panoramica utile sull'evoluzione del settore verso pratiche più sostenibili. I risultati più evidenti riguardano soprattutto il miglioramento delle condizioni legate ai

diritti dei lavoratori, alle politiche di contrasto alla corruzione e alla trasparenza, in particolare tra le imprese di dimensioni minori. Allo stesso tempo, il rapporto mette in luce una distanza significativa tra le linee guida dichiarate dalle grandi corporation e la loro effettiva applicazione nei siti operativi.

Pertanto, rimane aperta la questione se i miglioramenti ottenuti riflettano un effettivo cambiamento strutturale delle pratiche operative o se rappresentino solamente una risposta a requisiti di reporting sempre più stringenti. Dei circa 250 siti minerari analizzati, infatti, solo una minoranza risulta pienamente conforme ai requisiti della *survey*. Nel complesso, il miglioramento delle pratiche estrattive appare come un processo lento e graduale. Le aziende tendono a concentrarsi sugli aspetti ESG per i quali dispongono di maggiori margini di intervento, mentre gli obiettivi più complessi e costosi vengono spesso rinviati (Garst et al., 2022).

Inoltre, i progressi osservati sembrano essere legati soprattutto all'introduzione di normative più stringenti, a nuovi obblighi regolatori e a un sistema di reporting sempre più strutturato. Al contrario, le iniziative volontarie e i codici di condotta non vincolanti mostrano, nella pratica, un impatto limitato sulle reali modalità operative delle imprese.

Nel settore della raffinazione, i principali problemi ambientali derivano dai processi di separazione del metallo dai materiali di partenza, che comportano l'utilizzo intensivo di reagenti chimici e possono portare alla contaminazione di ingenti volumi d'acqua con sostanze tossiche e metalli pesanti. Nel caso delle terre rare, tali procedure risultano particolarmente idroesigenti e generano quantità significative di acque reflue contaminate.

Assieme agli impatti legati all'uso dell'acqua, la fase di raffinazione presenta ulteriori aspetti complessi connessi al mix energetico impiegato nei processi produttivi. L'uso di combustibili fossili comporta infatti emissioni di gas inquinanti quali diossido di carbonio, diossido di azoto e diossido di zolfo, questi ultimi due responsabili della formazione di piogge acide con effetti

potenzialmente dannosi sugli ecosistemi e sulle falde acquifere.
Oltre alla disponibilità dei materiali raffinati, risulta inoltre fondamentale il possesso di competenze tecnologiche e infrastrutturali avanzate per la loro trasformazione in prodotti ad alto valore aggiunto, come i semiconduttori.

2.1.2. Il contesto geopolitico

Lungo questa catena che va dall'estrazione alla raffinazione, fino alla produzione di componenti strategici, si innestano molte delle dinamiche geopolitiche alla base dell'attuale competizione globale nel campo dell'intelligenza artificiale.
All'interno del dibattito pubblico emerge con chiarezza come l'attuale competizione tecnologica tra Stati Uniti e Cina assuma sempre più i contorni di una vera e propria guerra commerciale. Da un lato, gli Stati Uniti possono contare sulla presenza sul loro territorio delle principali aziende high-tech globali e su solidi rapporti strategici con Paesi chiave della filiera dei semiconduttori, quali Taiwan, Corea del Sud e Giappone, fondamentali per l'approvvigionamento dei chip più avanzati.
Dall'altro lato, la Cina si trova in una posizione di rincorsa rispetto al *know-how* necessario per la prodduzione di queste componenti ad alta complessità tecnologica. I Paesi asiatici appena citati occupano infatti una posizione di leadership nello sviluppo e nella produzione di semiconduttori di ultima generazione, garantendo alle imprese statunitensi un vantaggio competitivo significativo rispetto a quelle cinesi. Proprio facendo leva su tali relazioni industriali e strategiche, gli Stati Uniti hanno introdotto una serie di controlli sulle esportazioni di componenti a *dual-use*[13], con l'obiettivo di limitare l'accesso cinese alle tecnologie più avanzate.
Queste misure hanno costretto Pechino a intensificare gli investimenti nella produzione domestica di chip di fascia alta, rallentandone al contempo il

[13] Con applicazioni sia civili che militari.

progresso nel breve periodo e obbligandola a destinare ingenti risorse a un settore particolarmente complesso e *capital-intensive*. Come conseguenza diretta di tali restrizioni, molte aziende cinesi hanno dovuto operare con hardware meno performante, come le GPU NVIDIA H20, progettate per rientrare nei limiti imposti dalla normativa statunitense. In risposta a questi vincoli, l'industria cinese ha progressivamente spostato l'attenzione sullo sviluppo software, cercando di compensare le limitazioni hardware attraverso l'ottimizzazione degli algoritmi e delle architetture dei modelli.

In questo contesto, modell come *DeepSeek* e *Hunyuan-Large* hanno mostrato come sia possibile ottenere prestazioni competitive anche utilizzando hardware considerato obsoleto, evidenziando il ruolo centrale del software nel superamento dei limiti fisici delle infrastrutture di calcolo.

Tuttavia, le restrizioni messe in atto nei confronti della Cina hanno evidenziato alcune criticità applicative. Le aziende cinesi, ad esempio, hanno avuto la possibilità di accumulare scorte di chip in previsione dell'entrata in vigore delle misure restrittive, o nei periodi in cui queste venivano annunciate ma non ancora pienamente implementate, come avvenuto nel caso delle GPU NVIDIA H100. A ciò si aggiunge la possibilità, seppur limitata, di reperire componenti attraverso canali informali o mercati paralleli. Un caso spesso citato a sostegno dei limiti di efficacia delle restrizioni è quello di Tencent, che nel maggio 2024 ha presentato il modello *HunyuanDiT*, dichiarandone l'addestramento tramite GPU NVIDIA A100 e V100, componenti già soggette a controlli sulle esportazioni.

Anche dal lato delle aziende occidentali produttrici di semiconduttori, come NVIDIA, AMD e Intel, le restrizioni hanno generato reazioni significative. Il mercato cinese rappresenta infatti una quota rilevante delle vendite di chip ad alte prestazioni, e le limitazioni all'export incidono sulle prospettive di crescita, sui contratti futuri e sulla presenza a lungo termine in uno dei mercati più dinamici del settore. Tra le principali strategie adottate da queste *corporation* figurano la progettazione di chip specificamente adattati al

mercato cinese[14], in modo da rispettare i parametri normativi imposti; la diversificazione verso altri mercati[15]; l'esclusione del mercato cinese dalle previsioni finanziarie a causa dei rischi legati all'export; e un'intensa attività di lobbying finalizzata a ottenere condizioni più favorevoli nell'applicazione delle restrizioni.

2.1.3. I datacenter

La componentistica hardware è essenziale per ottimizzare sia l'allenamento che la gestione energetica di un modello; tuttavia, oltre all'attività principale, è importante considerare anche l'energia impiegata dai servizi ausiliari e i costi legati al raffreddamento. L'Agenzia Internazionale dell'Energia (IEA, 2024) attesta che mediamente in un datacenter il 40% dell'energia raggiunge effettivamente i server, mentre tra il 7% e il 30% viene impiegato per il raffreddamento[16] e la restante percentuale viene associata agli altri servizi necessari al funzionamento del sistema.

L'efficienza dei datacenter viene calcolata tramite il *power usage effectiveness* (PUE) un indice in cui si rapportano il quantitativo totale di energia utilizzata dal datacenter e il quantitativo che raggiunge effettivamente le attrezzature informatiche. Più questo indice si avvicina di valore all'unità, minore sarà l'energia elettrica dispersa e di conseguenza minore sarà l'impatto energetico. Datacenter con PUE di 1,2 sono considerati molto efficienti mentre con valori compresi tra 1,5 e 2 vengono considerati scarsamente efficienti, quando il PUE supera il valore due si arriva a una grave inefficienza. Oltre ai datacenter tradizionali con minore potenza per rack, un PUE tendenzialmente sopra l'1,4 e destinati ad applicazioni aziendali, sono stati sviluppati centri più potenti e in grado di supportare in

[14] Come le RTX 4090D, GPU prodotte da NVIDIA per rientrare nelle specifiche della regolamentazione.
[15] Stati Uniti, Europa, India e Vietnam.
[16] L'IEA associa al 7% le prestazioni di hyperscale molto efficienti mentre il 30% è associato a datacenter meno recenti.

particolare l'IA.

Partiamo quindi dai datacenter specializzati: questi godono di un'alta intensità computazionale e di un'alta efficienza per Watt (PUE 1,1-1,2). Ottenere questi risultati in termine di efficientamento energetico significa utilizzare architetture di rete specifiche e molto veloci, numerose GPU/TPU di ultimo modello e sistemi di raffreddamento all'avanguardia. Inoltre, per abbassare le emissioni collegate a queste facilities ad alta intensità[17] vengono adottate interazioni con reti elettriche locali specializzate in mix energetici rinnovabili.

Passiamo quindi agli h*yperscale*, datacenter con estensione maggiore ai 50.000 m^2, gestiti da *big tech* come Google, Microsoft, AWS e Meta. Questi centri fanno della scalabilità la propria strategia migliore: è infatti dimostrato come all'aumentare della capacità di un datacenter il PUE migliori[18]. Proprio per questa caratteristica l'architettura è modulare, in modo tale da poter aggiungere potenza facilmente, e progettata per gestire milioni di server. All'interno di queste infrastrutture multiservizi vengono ospitati molteplici realtà: servizi web come youtube e Office365, cloud computing generale come storage e database, servizi edge o interi cluster dedicati all'IA. Il PUE degli *hyperscale* si attesta tra l'1,1 e l'1,3, grazie anche all'ottimizzazione della rete su cui operano e alla continua manutenzione a cui sono sottoposti. Anche in questo caso per provvedere al fabbisogno energetico si ricorre all'utilizzo di fonti rinnovabili o energia nucleare; oltre al massiccio ricorso ai *power purchase agreement* (PPA) in modo da rientrare negli obiettivi ESG e rincorrere la *carbon neutrality.*

[17] Sono impianti che possono variare per potenza tra i 100MW e i 1000MW.
[18] L'italia dei datacenter, energia, efficienza, sostenibilità per la transizione digitale 5 settembre 2025 – The European House Ambrosetti.

2.1.4. *End-of-life*

Infine, un ulteriore approccio per ammortizzare le emissioni associate allo sviluppo dell'intelligenza artificiale consiste nell'aumentare quanto più possibile la durata della componentistica hardware, così da utilizzare i dispositivi per un periodo sufficiente a giustificarne l'impronta ecologica. Proprio in questo ambito emergono alcune questioni particolarmente: il rapido ritmo dell'innovazione tecnologica riduce infatti la vita media dei dispositivi, favorendo un ricambio continuo che porta alla dismissione di hardware ancora funzionante e spesso in buone condizioni.

Sebbene tali dinamiche fossero già presenti prima dell'affermazione dell'IA, in relazione alla diffusione di computer, smartphone e altri dispositivi elettronici, la costruzione di nuovi data center dedicati all'intelligenza artificiale rischia di amplificare ulteriormente il problema dei rifiuti elettronici.

Secondo le stime del report *The Global E-waste Monitor* (2024), redatto dall'UNITAR, nel 2022 sono stati prodotti circa 62 miliardi di chilogrammi di rifiuti elettronici a livello globale, di cui solo il 17–20% è stato riciclato ufficialmente e in modo tracciabile. Considerata la rapidità con cui le infrastrutture per l'IA vengono aggiornate o sostituite, con cicli di vita stimati tra i tre e i cinque anni, le proiezioni per il 2030 indicano un ulteriore e significativo aumento dei rifiuti elettronici complessi. In questo contesto, migliorare le capacità di riciclo assume un ruolo centrale, poiché consente di ridurre la domanda di nuove materie prime e, di conseguenza, l'impatto ecologico diretto sugli ecosistemi.

All'interno dell'hardware dei data center sono presenti materiali di natura molto diversa: si va da metalli relativamente semplici da recuperare e riciclare, a plastiche termoindurenti difficilmente riutilizzabili, fino a metalli preziosi presenti in quantità minime, il cui recupero risulta economicamente complesso.

Negli ultimi anni, tuttavia, il settore ha iniziato a ricevere una crescente attenzione istituzionale, con l'adozione di strategie sempre più strutturate volte a migliorare la gestione del fine vita delle infrastrutture. Un primo passo in questa direzione riguarda il design modulare dei server, che consente di semplificare le operazioni di disassemblaggio e di sostituire esclusivamente specifiche componenti, come GPU e moduli di memoria.

Queste componenti possono così essere riutilizzate in contesti applicativi meno intensivi dal punto di vista computazionale, oppure avviate a processi di riciclo più efficienti. A supporto di questo modello si stanno diffondendo anche contratti di *hardware leasing*, attraverso i quali i produttori mantengono la proprietà delle tecnologie fornite e si assumono la responsabilità dello smaltimento e del riciclo a fine vita.

Tali pratiche favoriscono una maggiore circolarità delle risorse e rappresentano un passo importante verso la riduzione dell'impatto ambientale complessivo dell'infrastruttura per l'IA. Proprio in questo campo risulta necessario un maggiore impegno da parte del settore privato nel disincentivare lo smaltimento scorretto oltre a fenomeni di esportazione o di delocalizzazione verso Paesi caratterizzati da normative ambientali meno stringenti.

In questo contesto, l'economia circolare rappresenta una soluzione efficace in numerosi ambiti produttivi e assume un ruolo particolarmente rilevante nel settore dell'IA, caratterizzato da un rapido e continuo ricambio tecnologico.

A testimonianza di come il settore stia cercando di muoversi in questa direzione, può essere considerato il caso di Microsoft. Nel 2024, l'azienda statunitense ha dichiarato di aver raggiunto un tasso di riuso e riciclo pari al 90,9% per i server e l'hardware impiegati nei propri data center (Microsoft, 2025). Sebbene il raggiungimento anticipato di un obiettivo previsto dalla strategia "zero waste 2030" rappresenti, in linea generale, un segnale positivo, tale affermazione presenta alcune criticità che meritano di essere esaminate con attenzione.

In primo luogo, non risultano disponibili resoconti pubblici a supporto di questo dato: l'informazione non è stata sottoposta a un audit esterno indipendente e non è stata resa nota la metodologia utilizzata per il calcolo della percentuale dichiarata. In assenza di sistemi e di unità di misura chiaramente definiti ed espliciti, non è possibile stabilire se il valore si riferisca alla massa fisica dei materiali recuperati, al loro valore economico o al numero di server e componenti coinvolti.

Analogamente, non è disponibile un dettaglio che distingua tra componenti riutilizzate e componenti effettivamente riciclate, ne informazioni sulla quantità e sulla qualità dei materiali recuperati. Ulteriori elementi di incertezza riguardano i confini geografici e temporali dell'analisi. Non è infatti chiaro se siano state applicate esclusioni su base regionale, se il dato includa esclusivamente l'hardware di proprietà di Microsoft o anche quello gestito da terzi, né quale sia il periodo temporale di riferimento considerato. In assenza di tali informazioni, risulta difficile valutare in modo accurato la portata e l'effettiva significatività del risultato comunicato.

In conclusione, la comunicazione in materia di sostenibilità richiede elevati livelli di trasparenza e chiarezza metodologica. In mancanza di tali elementi, anche iniziative potenzialmente virtuose rischiano di essere percepite come pratiche di greenwashing o come semplici strategie di posizionamento comunicativo, piuttosto che come indicatori affidabili di un reale progresso ambientale.

In questo contesto, il settore high-tech, e in particolare quello dell'IA, è chiamato a un ulteriore processo di maturazione, volto a garantire livelli più elevati di affidabilità, precisione dei dati e trasparenza nelle pratiche adottate. Dall'analisi condotta emerge come una delle principali sfide consista nel definire meccanismi di rendicontazione in grado di offrire garanzie sulla qualità e sull'affidabilità delle informazioni pubblicate, senza compromettere il vantaggio competitivo delle imprese legato alle attività di ricerca e sviluppo.

Allo stesso tempo, risulta evidente che un rafforzamento dei quadri regolatori e degli standard di reporting rappresenta un elemento imprescindibile per orientare il mercato verso pratiche più sostenibili e verificabili. In assenza di criteri chiari e condivisi, il rischio è che la comunicazione in ambito ambientale rimanga frammentaria o prevalentemente orientata al posizionamento strategico, piuttosto che a una reale misurazione dei progressi compiuti.

In conclusione, la componente *embodied* dell'impatto ambientale dell'IA fatica ancora a trovare spazio nella ricerca e nella rendicontazione sostenibile. Lungo il paragrafo abbiamo valutato i processi di produzione principali coinvolti per la messa a punto dell'IA materiale, e le principali criticità che ne derivano. Solo tramite una maggior attenzione all'impronta ambientale dell'hardware specializzato e alla progettazione e alla gestione dei datacenter in un'ottica *green* sarà possibile affrontare le esternalità negative e i costi ambientali nascosti di questo tipo di attività.

Nella prossima sezione l'oggetto di analisi passerà dalla componente materiale dell'IA a quella "immateriale", valutandone gli impatti operativi dovuti alle fasi più energivore del processo di messa a punto e utilizzo del software.

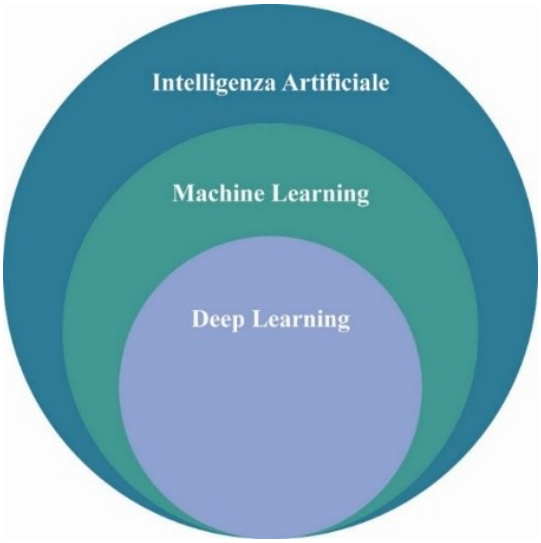
2.2. Gli impatti operativi dell'IA: allenamento, retraining e inferenza

L'intelligenza artificiale (IA) indica, in senso generale, l'insieme di metodi con cui una macchina riproduce alcune capacità cognitive umane, ad esempio attraverso sistemi che simulano processi decisionali. Oggi le soluzioni di IA sono impiegate in molti ambiti e, in particolare, si basano spesso su tecniche di apprendimento automatico.

I termini “Intelligenza Artificiale”, “*Machine Learning*” e “*Deep Learning*” vengono talvolta usati come sinonimi, ma non lo sono. *Machine learning* e *deep learning* sono infatti sottocampi dell’IA. Il machine learning raccoglie algoritmi che, tramite metodi statistici, consentono ai sistemi informatici di apprendere dai dati e di prendere decisioni. A differenza degli approcci tradizionali, in cui le regole sono definite esplicitamente nel codice, qui il comportamento del modello deriva dall’individuazione di schemi e regolarità all’interno dei dati.

Il *deep learning* è una famiglia specifica di modelli di machine learning basata su reti neurali artificiali caratterizzate da architetture particolarmente profonde e complesse, con un elevato numero di parametri da ottimizzare durante l’addestramento.

Figura 2.1 - Rappresentazione insieme Intelligenza Artificiale e sottoinsiemi. Elaborazione dell'autore.

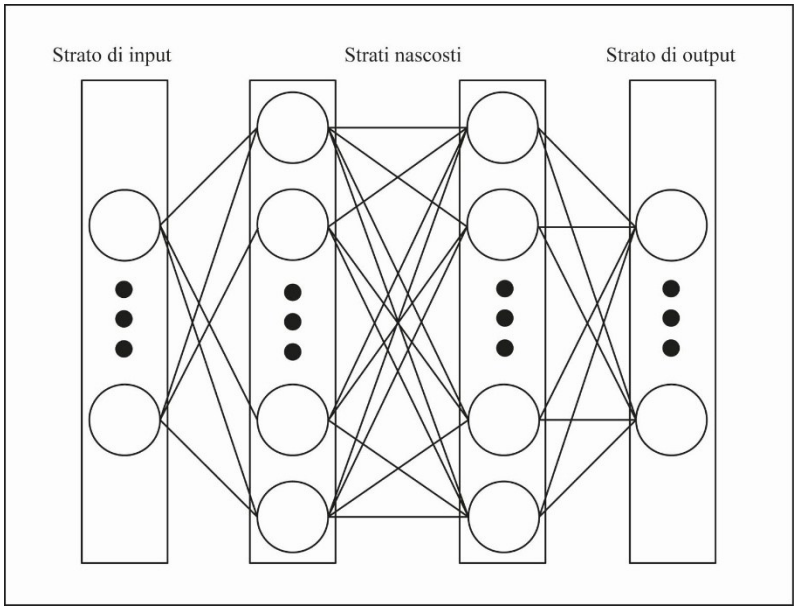


Per molti problemi sono sufficienti algoritmi relativamente semplici; tuttavia, in ambito industriale le reti neurali sono oggi tra le soluzioni più diffuse perché risultano efficaci in una gamma molto ampia di applicazioni. Le dimensioni e la complessità dei modelli possono variare enormemente: si va da reti piccole, addestrabili in pochi secondi su un comune laptop, fino a modelli molto grandi come GPT-3, che richiedono giorni o settimane di

calcolo su hardware specializzato.

Le reti neurali artificiali si ispirano (in modo semplificato) al funzionamento del cervello e puntano a riprodurre meccanismi di apprendimento. Una rete è composta da unità di calcolo ("neuroni") collegate tra loro: a partire da un segnale di ingresso, ogni neurone combina gli input ricevuti, applica una funzione di attivazione e produce un'uscita che viene trasmessa agli elementi successivi. Questi neuroni sono organizzati in strati: uno strato di input, uno o più strati nascosti e uno strato di output.

Figura 1.2 - Rappresentazione reti neurali artificiali. Elaborazione dell'autore.



Le connessioni tra neuroni sono descritte dai pesi, valori che determinano quanto ogni segnale influenzi i calcoli successivi. Durante l'addestramento, i pesi vengono modificati per ridurre l'errore del modello: in questo senso, essi rappresentano la "conoscenza" appresa dalla rete sulla relazione tra input e output. Anche la struttura del modello, ad esempio il numero di strati nascosti e il numero di neuroni per strato, incide direttamente sulla sua capacità di rappresentare relazioni complesse e costituisce una scelta progettuale fondamentale.

Nei modelli più semplici, i neuroni di uno strato sono spesso completamente connessi a quelli dello strato successivo e l'informazione scorre in un'unica

direzione. Nelle reti profonde moderne, invece, si adottano architetture e strategie diverse per evitare che il numero di parametri cresca troppo rapidamente, rendendo l'addestramento inefficiente o impraticabile.
Le reti neurali, soprattutto quelle di deep learning, apprendono modificando i pesi delle connessioni interne sulla base dei dati. Questo processo di apprendimento, insieme all'utilizzo successivo del modello, comporta costi operativi misurabili in termini di consumo energetico ed emissioni. In particolare, gli impatti operativi dell'IA si concentrano in tre attività principali: addestramento, ri-addestramento o aggiornamento e inferenza.

2.2.1. L'Allenamento

Per quanto riguarda l'allenamento dei modelli IA procediamo analizzando le variabili che determinano i consumi di questo tipo di attività.
Per prima cosa troviamo la dimensione, ovvero da quanti parametri è composta l'architettura, se sono tra l'uno e i sette miliardi sono considerati modelli di piccole dimensioni, per cui basterà un numero ridotto di GPU per essere addestrato. Quando i parametri sono oltre i settanta miliardi si parla di modelli enormi per il cui allenamento vengono impiegate centinaia o migliaia di GPU in funzione per settimane o addirittura mesi.
Sebbene questa variabile sia importante soprattutto per i modelli meno recenti, dal 2023 le grandi aziende del settore come OpenAI, Google, Anthropic e Meta hanno smesso di pubblicare il numero di parametri dei loro modelli. Questo perché attraverso il veloce progresso tecnologico e tecniche come il m*ixture of experts*[19], o dataset migliori; alcuni modelli recenti hanno ottenuto risultati migliori a livello di prestazioni con architetture più leggere. Ad oggi il numero di parametri non è più correlato linearmente con il consumo energetico, sebbene questa informazione possa ancora fornire un'idea approssimativa sulla scala del modello (Luccioni et al., 2024).

[19] Che permette di non avere tutti i parametri sempre attivi per ogni token.

Seconda variabile per importanza troviamo la quantità di dati: un modello deve infatti leggere e processare enormi quantità di testo o immagini durante l'allenamento. In particolare, i token sono l'unità minima di testo che un modello di linguaggio usa per leggere, capire e generare frasi. Sono utilizzati i token perché più efficienti rispetto alle parole o alle lettere, sono infatti "pezzi" di testo che il modello sa gestire bene. Questi token possono presentarsi come: parole, caratteri, spazi, emoticon o ancora pezzi di parole con cui si facilita il modello nel riconoscere pattern e radici comuni in differenti lingue e contesti.

A livello di addestramento di un modello il quantitativo di token utilizzati contribuisce alla crescita lineare del costo computazionale: più è lungo un testo più GPU saranno utilizzate per processarlo e quindi maggiore sarà l'energia richiesta. Proprio per ovviare a questa problematica si è andati ad agire sui dataset in modo da ridurre il quantitativo di token necessari, puntando sulla qualità dell'informazione.

I dataset sono infatti filtrati dai contenuti di bassa qualità[20]; sottoposti a deduplicazione[21] e organizzati per importanza[22] in modo tale da concentrare il più possibile le informazioni utili al modello. Altre pratiche utili in questo senso sono il *curriculum training*[23] e l'addestramento multi-stage per cui si sottopone inizialmente un dataset più ampio e generico, per poi in una seconda fase utilizzare dati specializzati per sviluppare competenze complesse.

[20] Come spam, testi generati automaticamente, commenti tossici o ancora micro-frasi e frasi eccessivamente lunghe.

[21] Con cui si scartano le copie dello stesso testo. Può essere deduplicazione semantica se si tratta di testi diversi riguardo lo stesso argomento, o di pagine di documentazione che ripetono concetti identici.

[22] Con sistemi di ranking del dataset in cui token provenienti da materiale didattico o accademico sono considerati maggiormente rispetto a lunghi log di chat casuali fra utenti.

[23] Metodi avanzati di questa pratica come la self-paced learning, permettono al modello stesso di regolare in autonomia il ritmo di somministrazione dei dataset per l'addestramento in base alle sue prestazioni.

Con questa attenzione alla qualità dei dati e un'efficienza sempre maggiore è possibile scartare tra il 50% e il 90% dei token provenienti dal web, migliorando l'efficienza di modelli futuri e riducendo il consumo energetico fino al 92% (Verdecchia et al, 2022).

Ora che è chiaro cosa sono i token possiamo introdurre un'altra nozione fondamentale per ottenere un'unità di misura che permetta di confrontare modelli per dimensione o architettura, indipendentemente dall'hardware su cui girano. Introduciamo quindi i FLOPs (*Floating Point Operations*): operazioni in virgola mobile[24] come somme, sottrazioni, moltiplicazioni e divisioni. I modelli neurali sono in grado di compiere miliardi di calcoli in parallelo; proprio il numero di FLOPs necessari alla generazione di un token indica il costo computazionale intrinseco di un modello. Tramite l'informazione FLOPs/token è infatti possibile capire quale modello sia il più dispendioso a livello di quantità di operazioni, paragonando di fatto architetture diverse per efficienza in modo più diretto e pulito.

Allo stesso tempo come *trade-off* bisogna tenere a mente che l'uso di quest'ultima metrica sia utile, ma che essa rischia di semplificare eccessivamente la complessità reale dei consumi. Questo può portare a paragoni del lavoro di modelli diversi per hardware, software e infrastruttura perdendo le informazioni che ne conseguono, soprattutto in un settore in cui gli scenari di produzione sono estremamente eterogenei.

L'ultima componente che va ad impattare notevolmente sui costi di allenamento è la tipologia di hardware che si utilizza. Numerose ricerche mostrano come impiegare hardware specifico per IA, come le TPU o gli ASIC specializzati, offra spesso un vantaggio energetico significativo nei carichi di lavoro di *deep learning* (Patterson et al, 2022). La scelta dell'hardware influisce direttamente sia sui costi operativi che sull'impatto ambientale dei modelli su larga scala. Ultima variabile, ma non per

[24] Un'operazione in virgola mobile è un calcolo matematico eseguito su numeri decimali.

importanza, del costo energetico riguardante l'allenamento dell'IA proviene dai datacenter in cui si svolge l'attività computazionale. Infatti, oltre ai costi computazionali delle operazioni svolte, bisogna tenere conto anche dell'energia utilizzata da servizi ausiliari e dei costi di raffreddamento della struttura in cui lavora l'hardware. A livello quantitativo, l'addestramento di una singola architettura con 213 milioni di parametri può produrre circa 300 tCO_2e, equivalenti alle emissioni di 125 voli di andata e ritorno tra Pechino e New York (Jiang et al, 2024).

Inoltre, per ottenere un modello finale con prestazioni ottimizzate possono servire fino a 5000 modelli di prova pre-addestrati. Nonostante lo sviluppo tecnologico della componentistica hardware ne migliori l'efficienza energetica, l'ultima sessione di *training* di GPT-3[25] sembra aver consumato 1287 MWh di elettricità, con un'impronta di carbonio di 552 tCO_2e (Jiang et al, 2024). Va inoltre specificato che la stima non tiene conto del consumo energetico e delle emissioni per tutte le operazioni di ricerca e sviluppo, né di quelle derivanti da aggiornamenti o *retraining*. Nel seguente sottoparagrafo saranno esplorate le procedure e ler condizioni che permettono di ottimizzare l'allenamento di modelli IA dal punto di vista energetico e di emissioni.

2.2.1.1. *Best-practice* in allenamento

Nel delineare le *best practice* per ridurre l'impatto climatico dell'addestramento, i lavori di Strubell et al. (2019) e Patterson et al. (2022) forniscono due prospettive complementari che aiutano a capire sia "dove" si accumulano le emissioni sia "quali leve operative" le possano ridurre in modo più efficace.

Nello studio di Strubell et al. (2019) il problema viene inquadrato come una diagnosi della pratica scientifica: l'impatto infatti non deriva soltanto dal

[25] Con 175 miliardi di parametri.

training “conclusivo”, ma anche dal percorso sperimentale che lo precede. Tutte le prove, le riformulazioni e le replicazioni per ottenere un modello finale possono amplificare significativamente il consumo energetico se si include l’attività di ricerca e sviluppo all’interno dei confini del sistema. Patterson et al. (2022) spostano invece l’attenzione sulle condizioni concrete dell’addestramento e mostrano che molte stime diffuse risultano fuori scala. Ciò che viene criticato nella formulazione di queste stime è che siano costruite su assunzioni “medie” o sfavorevoli come l’utilizzo di hardware superato o di datacenter non IA-specifici; oltre ai mix energetici medi. In questa cornice viene sostenuto che, a parità di obiettivo, l’impronta dipende in modo determinante da scelte tecniche e infrastrutturali sintetizzabili nelle 4M:

1- *Model*: questa prima dimensione dipende dalla scelta dell’architettura e dalle tecniche algoritmiche impiegate; in questo modo viene definita “quanta computazione” serve per raggiungere un certo livello di prestazioni. L’ottimizzazione permette così di raggiungere la stessa qualità con meno passi di training, meno parametri attivi, o meno calcolo per token, così da agire sui consumi. Un esempio di tecniche sono i modelli sparsi[26], il *mixture of experts*[27] e in generale le architetture che diminuiscono le interazioni tra neuroni.

2- *Machine*: in questo caso si valuta l’hardware su cui avviene l’allenamento e quanto è adatto al carico di lavoro. Un hardware più moderno e specializzato compie più lavoro per watt risultando in meno KWh consumati.

3- *Mechanization*: per cui si agisce sull’efficienza dell’infrastruttura dove avviene il training: valori associati ai datacenter, alle tecnologie di

[26] I modelli sparsi sono modelli in cui solo una parte dei parametri o delle connessioni viene usata in modo significativo. L’idea è evitare di “accendere tutto il cervello” ogni volta.
[27] Un Mixture of Experts (MoE) è un’architettura in cui, invece di usare sempre tutto il modello per ogni input, il sistema è composto da sottoreti specializzate (experts) e da un meccanismo che decide quali expert attivare per ciascun token.

raffreddamento, alle dinamiche di distribuzione elettrica, e di overhead; diventano fondamentali per agire attivamente sull'efficienza dell'allenamento. Passare da un datacenter generico a uno IA-specifico, molto ottimizzato, può ridurre parecchio la quota di energia non direttamente computazionale. Da qui l'importanza che assume l'infrastruttura: non basta scegliere il modello e la GPU, conta anche dove stanno le macchine e quanto è efficiente la struttura che le alimenta e raffredda.

4- *Map:* ultima direttrice per ottimizzare l'allenamento dei modelli è la localizzazione del training, sia geografica che temporale. Infatti, oltre ad avere differenze per mix energetici regionali, anche il momento in cui avviene un allenamento può influenzare i consumi[28]. A parità di energia consumata, il CO_2e cambia con l'intensità di carbonio della rete elettrica locale per cui allenare in una regione con mix energetico più "pulito" può abbattere le emissioni di CO_2e anche se l'energia totale rimane la stessa.

A fare da cerniera metodologica tra queste due letture interviene Henderson et al. (2020), che sottolinea come l'eterogeneità delle assunzioni renda i confronti fragili e sottolinea come il reporting sistematico di energia e CO_2e tramite strumenti e appendici standardizzate sia fondamentale per l'analisi dei consumi energetici. In questa prospettiva, le *best practice* non operano solo sull'ottimizzare il training, ma richiedono anche misurazione e documentazione per ottenere replicabilità.
Queste considerazioni sull'importanza della standardizzazione per l'attività di reporting verranno affrontate in modo diretto nel corso del terzo capitolo.
Infine, lo studio di Jiang et al. (2024), prepara direttamente il terreno al tema del paragrafo successivo: anche quando il singolo addestramento è ottimizzato, l'impatto può diventare rilevante se si considera la dimensione cumulativa della moltiplicazione dei cicli di sviluppo e aggiornamento su

[28] La rete elettrica può essere più o meno "verde" in ore diverse.

larga scala, dove frequenza e modalità di *retraining* finiscono per pesare quanto, o più, dell'esecuzione "una tantum".

2.2.2. Il *Retraining* e aggiornamento

Prendiamo ora in considerazione una variabile che spesso non viene considerata: il riaddestramento e l'aggiornamento dei modelli. Queste due pratiche possono arrivare ad avere un costo pari a quello di allenamento a seconda delle operazioni che devono essere compiute. Partiamo con il *full-retraining* che prevede la ripetizione di tutto l'addestramento da zero con dataset e architetture aggiornate, oltre all'adozione procedure più efficienti. Questo è il tipo di *retraining* più costoso e viene generalmente effettuato dalle aziende più grandi solo per le *major releases*, come da GTP-4 a 5, ogni due anni proprio per l'importanza del costo.
Abbiamo poi il *fine-tuning*, un raffinamento mirato su un sottoinsieme di dati per ottenere un miglioramento delle capacità logiche; ha un costo molto minore dell'allenamento iniziale, ma comunque significativo per modelli sopra i 70B di parametri. Più sarà grande il modello, più GPU saranno utilizzate per il *fine-tuning;* per cui un aggiornamento anche minore può risultare in migliaia di GPU-ore diventando estremamente costoso[29]. Arriviamo quindi agli aggiornamenti di tipo *reinforcement learning,* utilizzati per allineare maggiormente le risposte dei modelli con le preferenze umane. A differenza dei *fine tuning*, non si opera su un dataset specifico per migliorare determinate *task*; ma si allena invece il modello tramite dei feedback sulle risposte così da allenarlo a scegliere quelle più allineate con il comportamento umano. Può essere di più tipi, i principali sono il *reinforcement learning from human feedback* (RLHF) in cui l'allenamento interattivo avviene su dati etichettati da esseri umani che esprimono una

[29] Un parameter efficient fine tuning permette, tramite l'aggiornamento di pochi parametri (1-2%), una riduzione del costo computazionale rispetto a un supervised fine tuning, dove si aggiornano tuti i parametri, da 4 milioni di GPU ore a 400 mila GPU ore.

preferenza sulla risposta, oppure il *reinforcement learning from AI feedback* (RLAIF) in cui non sono più gli esseri umani a valutare le risposte, ma un altro modello IA in modo da ridurre i costi e velocizzare gli aggiornamenti (Lee et al, 2024).

Per quanto riguarda il RLHF l'interazione tra uomo e macchina determina dei costi maggiori, una velocità più bassa di aggiornamento e una qualità molto alta, seppur estremamente limitata. Per i RLAIF invece i costi sono più contenuti grazie all'opera del "modello giudice"[30], a cui è associata una velocità maggiore e una qualità variabile ma su grande scala. Queste due pratiche vengono implementate insieme per ottenere qualità dai RLHF e quantità dai RLAIF, in questo modo il reward model dell'LLM ottiene dei dati di qualità standard scalabili su grandi quantitativi. Inoltre, vengono corretti gli eventuali bias del modello giudice o dell'essere umano, migliorandone la coerenza e rendendo il modello più dialogico, intuitivo e di aiuto per gli utenti finali.Nel contesto dei modelli di grandi dimensioni, l'uso combinato di RLHF e RLAIF appare attualmente come il compromesso più pragmatico, pur non potendo esimersi da problemi di scalabilità e continuo controllo.

Implementare queste tecniche comporta un forte consumo energetico, da una parte perché utilizzare feedback umani è un processo costoso e lento e dall'altra perché per ottenere un buon rendimento con i RLAIF bisogna passare attraverso un "modello giudice" solido e di qualità altissima su cui va investito tempo, denaro ed energia.

Infine, troviamo la modalità di addestramento continuo per cui il modello non è addestrato per poi essere congelato, ma viene continuamente aggiornato con nuovi dati, regole e conoscenze. Con il passare del tempo, infatti, un modello per funzionare al meglio deve poter accedere a materiale, il più attuale possibile, per essere aggiornato riguardo leggi, eventi, scoperte

[30] Quello che valuta le risposte come se fosse un essere umano.

scientifiche, slang e via dicendo. Tramite una continua somministrazione di dataset, puliti e ordinati per dominio, con pratiche come *fine-tuning, continual pre-training,* RLHF e RLAIF*,* e aggiornamenti modulari (come MoE), è possibile aggiustare il modello su ciò che è cambiato e correggerne gli eventuali errori. Il principale rischio di questo tipo di aggiornamento risiede nel *catastrophic forgetting* per cui il ri-addestramento su dati nuovi porta il modello a non essere più in grado di fare ciò che faceva in precedenza. Per evitare questo fenomeno si utilizzano degli algoritmi chiamati *elastic weight consolidation* (EWC) con cui è possibile rallentare l'apprendimento di determinati pesi in base alla loro importanza rispetto alle attività svolte in precedenza (Kirkpatrick et al, 2017).

In conclusione, possiamo affermare che l'attività di riaddestramento comporta un consumo energetico ridotto rispetto all'attività di allenamento base, ma che non è in alcun modo trascurabile. Mahadevan e Mathioudakis (2024) mostrano che il *retraining* completo su tutto lo storico dei dati comporta costi elevati spesso non giustificati dai benefici marginali in termini di accuratezza sulle *query* sottoposte. Gli studiosi suggeriscono un approccio *cost-aware* che limiti o ritardi il *retraining*, supportando l'idea che aggiornamenti di piccole dimensioni, somministrati più volte nel tempo e basati su dati più recenti possano essere preferibili a un *retraining* completo. Pertanto, la frequenza di questi aggiornamenti ha un impatto significativo sul quantitativo di energia necessario al funzionamento dei LLM, per questo motivo introdurre un "*drift detector"*[31] accuratamente selezionato può ridurne il consumo energetico e migliorare le performance (Poenaru-Olaru et al., 2023).

In particolare, passare da un sistema di aggiornamento a frequenza fissa (ogni 2 anni, ogni 6 mesi, ogni due settimane) ad uno in cui gli aggiornamenti vengono somministrati solamente quando il drift detector lo reputa

[31] Un algoritmo progettato per rilevare quando la distribuzione dei dati cambia nel tempo rispetto a quella su cui un modello è stato addestrato.

necessario ne migliora l'efficienza energetica. In seguito, l'analisi passerà dai consumi operativi per ottenere un modello accurato e scalabile, a quelli derivanti dall'utilizzo vero e proprio. Il processo di inferenza si dimostra sempre più importante a livello di consumi man mano che l'adozione dell'IA si diffonde.

2.2.3. L'Inferenza

Analizziamo quindi la seconda componente più importante, dopo l'attività di allenamento, nel consumo di energia elettrica da parte dell'IA: l'inferenza. Per inferenza si intende l'utilizzo di un modello già addestrato per fare previsioni, nel momento in cui si presenta una domanda viene generata una risposta tramite moltiplicazioni di matrici e lettura/scrittura in memoria, eseguite su GPU o TPU specializzate.

Mentre riguardo i costi energetici dell'allenamento è già presente una nutrita letteratura accademica, l'inferenza ha suscitato meno attenzione attraendo relativamente poca attenzione da parte della comunità scientifica. Questo passaggio in sordina di una fase critica per il conteggio dei consumi evidenzia, oltre a una concreta difficoltà di calcolo, una mancanza di standardizzazione e di approcci formulati per lo studio di questo fenomeno. Tutto ciò è in parte dovuto alle caratteristiche intrinseche dell'inferenza: il ridotto consumo energetico per *query*, la varietà di attività che un modello può compiere e la volatilità dovuta a consumi fluttuanti.

Andiamo ad analizzarle per ordine: nel generare una risposta i modelli impiegano una potenza di calcolo molto alta[32], ma questo processo avendo una durata di millisecondi, genera un consumo di energia elettrica contenuto. Pertanto, conviene concentrarsi sul consumo di energia più che sulla potenza istantanea delle GPU, andando a valutare il totale richiesto per gestire più inferenze nel tempo. Se andiamo a scalare questi dati su milioni o miliardi di

[32] Come riferimento abbiamo una GPU che impiega 200 W mentre genera una risposta, che a livello di potenza istantanea può sembrare un'operazione pesante.

richieste otteniamo così una cifra stimabile di quanto impatti questo tipo di attività. Con il dato aggregato possiamo quindi determinare i costi, le emissioni e la capacità di rete necessarie al funzionamento dell'IA. Ad oggi le stime riguardo l'implementazione dell'IA portano a rivalutare ulteriormente l'importanza dell'inferenza nel conto delle emissioni totali, una previsione in particolare riporta che i consumi raggiungeranno l'ordine dei petawattora[33] già nel 2026 (IEA, 2024).

Accanto al ridotto consumo energetico per singola richiesta, un ulteriore elemento che contribuisce a rendere complessa l'analisi dell'inferenza riguarda la forte variabilità dei consumi nel tempo. A differenza dell'allenamento, che avviene in contesti relativamente controllati e ripetibili, l'inferenza si svolge in condizioni operative dinamiche, influenzate dal carico istantaneo del sistema, dalle modalità di utilizzo dei modelli e dalle scelte implementative adottate.

Come evidenziato da Fernandez et al. (2025), il consumo energetico associato a una singola inferenza non è una grandezza stabile, ma può variare in modo significativo anche a parità di modello, a seconda di fattori quali la lunghezza degli input e degli output, le strategie di generazione, il software di *serving* e l'architettura hardware impiegata. Gli autori mostrano inoltre come una parte rilevante di questa variabilità sia legata alla distinzione tra la fase di *prefill*, nella quale l'elaborazione dell'input è fortemente parallelizzabile, e la fase di *decoding*, che procede in modo sequenziale e risulta meno efficiente dal punto di vista energetico (Fernandez et al., 2025). Nei modelli linguistici di grandi dimensioni, proprio la fase di generazione del testo tende a incidere in misura significativa sul consumo complessivo, contribuendo a rendere i carichi energetici dell'inferenza particolarmente sensibili alla natura delle richieste e alle condizioni di utilizzo. Questa volatilità rende difficoltosa l'adozione di metriche aggregate semplici o di

[33] 10^{15} Wh.

valori medi rappresentativi. Fernandez et al. (2025) evidenziano infatti come approcci basati su stime teoriche, quali il consumo nominale delle GPU o il conteggio delle operazioni computazionali, tendano a sottostimare l'energia effettivamente utilizzata in scenari reali. Ne deriva la necessità di valutare l'impatto energetico dell'inferenza in modo più contestualizzato, tenendo conto non solo del numero totale di richieste, ma anche della loro eterogeneità e della distribuzione temporale con cui vengono elaborate.

Passiamo ora alla seconda caratteristica chiave dell'inferenza, ovvero la varietà di richieste che un LLM può gestire: attività brevi o semplici come rispondere alla domanda "che tempo c'è oggi?", tradurre un breve testo; e attività complesse come la generazione di codici, compiere ragionamenti multipasso, oltre a tutto ciò che inizia ad avere un numero importante di token e di passaggi.

Abbiamo inoltre le diversità nelle *query* per tipo di output come la generazione di immagini o video e l'elaborazione di file audio. Chiaramente queste attività, molto diverse tra loro, comportano consumi differenti a seconda della complicatezza della task richiesta, ad esempio l'inferenza dovuta ad una semplice operazione di classificazione di testo può consumare 0,002 Wh, mentre la generazione *text-to-text* può arrivare a consumare fino a 0,047 Wh, per quanto riguarda invece task più pesanti come la generazione *text-to-image* si arriva al picco di 2,907 Wh (Luccioni et al., 2025).

Con questo range molto ampio di valori diviene difficile stimare una media affidabile dei consumi, poiché l'energia consumata per singola inferenza varia enormemente. Queste differenze comportano una distribuzione dei consumi altamente asimmetrica: con numerose inferenze leggere con output contenuti e una piccola frazione di richieste pesanti con output impegnativi e ragionamenti multi-step.

Nonostante per numero quest'ultime siano inferiori, a livello di complessità comportano un peso altamente impattante sul consumo totale: da qui la

necessità di studiare il modello tramite i percentili[34], con particolare attenzione ai percentili più alti[35], dove le code pesanti di questo tipo di distribuzione hanno valori molto diversi dal valore medio (Barroso e Dean, 2013).

Anche a livello di hardware la relazione con il carico di lavoro non è lineare per cui non è possibile stimare una media affidabile dei consumi. Il lavoro delle GPU moderne, infatti, dipende da moltissime variabili[36] per cui una richiesta complessa non è solo più lunga da eseguire, ma va ad attivare veri e propri pattern di potenza diversi, eseguiti su parti differenti della stessa scheda grafica.

Altra considerazione di cui tenere conto è la volatilità dell'energia dovuta ai consumi fluttuanti: le richieste per l'inferenza, infatti, non arrivano in modo regolare, poiché gli utenti non utilizzano i modelli sempre con lo stesso ritmo. Questa irregolarità porta ad avere dei picchi e delle cadute improvvise nel carico di lavoro, di conseguenza anche l'energia necessaria a sostenere le variazioni di potenza oscilla rapidamente (Chen et al., 2025). In particolare, nell'articolo di Chen et al., vengono introdotti dei meccanismi di controllo del workload per renderlo più regolare così da contenere l'instabilità, diminuire i picchi e rendere più semplice previsione e gestione dei flussi.

Ad ogni modo anche se presenti dei meccanismi di controllo l'infrastruttura deve essere sempre in grado di gestire i picchi di attività, anche se questo significa mantenere accesi server e GPU/TPU quando il numero di *query* sottoposte è basso.

Per gestire le richieste degli utenti i grandi provider come OpenAi, Meta, Google e Anthropic, si affidano ad un'infrastruttura digitale, complessa e ottimizzata in modo tale da massimizzare la velocità di risposta. In una tipica architettura di inferenza LLM, le richieste vengono prima smistate da un

[34] I percentili sono misure statistiche che indicano la posizione relativa di un valore all'interno di una distribuzione. Il *k-esimo percentile* è il valore sotto il quale cade il k% dei dati.
[35] P95-P99.
[36] Dai FLOPs, alla memoria attiva, dall'uso di tensor cores e dalla frequenza attiva.

livello di *load balancing*[37] e pre-processate da servizi di front-end, dove possono essere aggregate tramite meccanismi di *batching*[38] prima di essere inoltrate ai server che ospitano le GPU. Le GPU eseguono quindi l'inferenza, restituendo i risultati a ritroso.

Nella pratica, queste fasi possono essere accorpate o implementate in modo differente a seconda dello *stack*[39] e del *runtime*[40] adottati. Il numero di richieste che arriva ad un modello IA va quindi ad influire sui suoi consumi. Dal momento che un datacenter è progettato per garantire il rispetto dei vincoli di latenza e di affidabilità del servizio; l'efficienza nell'utilizzo delle GPU è fortemente influenzata dalla regolarità del flusso di richieste. Maggiore è il numero di richieste, maggiore sarà la rapidità con cui i batch vengono riempiti, riducendo di conseguenza il tempo di attesa e migliorando l'efficienza computazionale della GPU; in queste condizioni è possibile ottenere sia *throughput* elevato sia latenza contenuta. Invece nel momento in cui il traffico si riduce, i *batch* tendono a rimanere incompleti e il sistema, per rispettare i vincoli di latenza[41], accetta batch di dimensioni minori, con un conseguente sottoutilizzo delle GPU. Il vincolo di latenza serve proprio a introdurre un *timeout* di *batching* che limita il tempo massimo di attesa prima dell'invio delle richieste alle GPU, determinando la dimensione dei batch in funzione del traffico.

L'architettura degli LLM è quindi volta all'ottimizzazione dei tempi di latenza per la soddisfazione dell'utente, in modo da fornire una risposta più velocemente possibile. Per assicurare la velocità e la costanza del servizio i

---

[37] A seconda di criteri quali: disponibilità di server, livello di carico, latenza prevista o ancora vicinanza geografica.

[38] Può essere *dynamic-batching* per cui un insieme di richieste cresce fino ad un timeout, oppure *micro-batching* per cui gli insiemi sono di piccole dimensioni e inviati di frequente

[39] È l'insieme coordinato di tecnologie, componenti e servizi che, messi insieme, permettono al sistema di funzionare end-to-end.

[40] È il software che gestisce l'esecuzione concreta del modello durante l'inferenza.

[41] Un service level agreement (SLA) è l'impegno formale o operativo sui livelli di servizio che il sistema deve garantire, in particolare sulla latenza e sulla disponibilità. Rimane un vincolo di business e non un vincolo tecnico ed è espressione del limite massimo accettabile di degradazione del servizio in termini di latenza.

datacenter utilizzano gruppi di GPU/TPU in *stand-by* operativo[42] che nei momenti di attività maggiore vengono impiegate dal modello. Questa componentistica può consumare tra il 20-40% del costo di una GPU in utilizzo, rimanendo in attesa di essere impiegata. Nonostante possa sembrare uno spreco di capacità computazionale questa scelta si spiega economicamente perché avviare da zero una GPU comporta non solo un costo, ma anche una tempistica maggiore determinando un aumento della latenza (Lu et al., 2025).

2.2.3.1. *Best practice* in inferenza

Per migliorare i consumi durante l'inferenza sono state messe a punto delle pratiche di architettura dei modelli tali da ridurre il lavoro inutile per ogni risposta. La quantizzazione, ad esempio, consiste nel ridurre la precisione numerica con cui sono rappresentati i pesi di un modello, senza alterare in modo significativo il comportamento o la qualità dell'output generato. Mentre nella fase di addestramento la precisione è cruciale, perché piccoli errori possono deviare il processo di apprendimento, durante l'inferenza[43], il modello è cristallizzato: non impara più, ma applica una sequenza di operazioni per stimare il token successivo più probabile. I modelli moderni sono fortemente sovra-parametrizzati e caratterizzati da un'elevata ridondanza, per cui variazioni minime nei valori dei pesi producono effetti trascurabili sul risultato finale.

Approssimando questi valori e rappresentandoli con meno bit, è possibile ridurre l'uso di memoria e il traffico di dati, che rappresenta una delle principali componenti del costo energetico durante l'inferenza. Questo si traduce in un minor consumo energetico e un aumento dell'efficienza misurabile come numero di token generati per Watt, senza sacrificare in

[42] In genere si tengono in funzione poiché accendere da fredda una GPU durante un picco di attività va a consumare più energia nel momento peggiore, amplificando il picco e consumando latenza per cui vi è uno spreco di energia prima di servire gli utenti.

[43] Il modello, infatti, non aggiorna più i pesi e risulta più tollerante ad approssimazioni numeriche.

modo apprezzabile le capacità linguistiche del modello (Gholami et al., 2021).

Un'altra pratica di ottimizzazione dei consumi è il *pruning*, che parte dal presupposto per cui non tutti i neuroni contribuiscano in modo significativo alla capacità predittiva complessiva del modello. In questo contesto, la ridondanza non rappresenta necessariamente intelligenza latente pronta all'uso, ma può tradursi in uno spreco computazionale durante l'inferenza. Durante l'addestramento, una rete neurale distribuisce il carico informativo su un numero molto elevato di parametri; tuttavia, una volta concluso il training, una parte non trascurabile di essi può attivarsi raramente, avere un impatto marginale sull'output oppure risultare fortemente correlata ad altri neuroni che svolgono una funzione simile.

Il *pruning* mira quindi a rimuovere queste componenti superflue per ridurre il numero di operazioni aritmetiche, il traffico di memoria e la latenza in fase di inferenza (Woo et al, 2021). In particolare, il *pruning* strutturato si basa sulla valutazione di metriche come: la norma dei pesi, la sensibilità del loss alla rimozione di specifici componenti architetturali, il livello medio di attivazione e la ridondanza statistica. A partire da queste misure, vengono eliminati interi neuroni, canali, teste di attenzione o blocchi computazionali, rendendo la riduzione effettivamente sfruttabile dall'hardware. Grazie a questa "potatura" mirata, il modello mantiene in genere una qualità percepibile molto simile all'originale e, in alcuni casi, può persino migliorare grazie alla riduzione di rumore e *overfitting*. Nel panorama attuale, il *pruning* rappresenta una delle poche tecniche strutturali capaci di ridurre in modo significativo il costo energetico dell'inferenza senza richiedere modifiche hardware, cambi di formato numerico o una sostanziale perdita di prestazioni (Tmamna et al., 2024).

Infine, dato che non tutte le domande necessitano della stessa potenza, diventa possibile smistare le richieste in base alla loro complessità computazionale. L'inferenza adattiva è un approccio in cui il sistema decide

dinamicamente quanta potenza computazionale allocare a ogni richiesta[44], invece di utilizzare sempre lo stesso modello di grandi dimensioni per tutte le domande. Attraverso router, classificatori leggeri e stime di confidenza, molte richieste semplici vengono gestite da modelli più piccoli o risolte tramite meccanismi di uscita anticipata, mentre solo i compiti più complessi arrivano ad attivare modelli più costosi o ragionamenti più profondi.

Poiché una larga parte delle query reali richiede operazioni di lettura e risposta relativamente semplici, questo meccanismo consente di calibrare l'elaborazione sulla reale complessità della richiesta, riducendo in modo significativo il numero medio di parametri attivi, i token processati e il tempo di inferenza, con benefici concreti, in media, sia sui consumi energetici sia sulla latenza (Laskaridis et al., 2021).

L'inferenza adattiva è già ampiamente utilizzata nei sistemi di produzione, soprattutto a livello infrastrutturale e di routing tra modelli di dimensioni diverse, mentre le forme più fini di adattività all'interno dei singoli modelli sono in parte già implementate, ma restano anche un'area di ricerca attiva[45], perché promettono di migliorare efficienza e sostenibilità senza sacrificare in modo apprezzabile la qualità delle risposte.

Un ulteriore meccanismo riconducibile all'inferenza adattiva è lo *speculative decoding*. Questo tipo di algoritmo nasce per far fronte a una caratteristica fondamentale dei transformer autoregressivi[46]: durante l'inferenza, un grande modello linguistico genera l'output un token alla volta. Per mitigare la lentezza intrinseca, dovuta al passaggio completo di ogni singolo token attraverso il modello, lo *speculative decoding* introduce una soluzione basata

[44] In termini di parametri attivi, token processati e profondità di esecuzione.

[45] Pratiche come *early exit, dynamic depth/width*, *MoE*, *token-level adaptivity* non sono ancora *plug-and-play*.

[46] Un transformer è un'architettura di rete neurale introdotta nel 2017 per risolvere un problema pratico: come far capire a una macchina le relazioni tra le parti di una sequenza (parole in una frase, note in una melodia, token di codice) senza doverle leggere una alla volta. La chiave è il meccanismo di *self-attention*: ogni elemento della sequenza guarda tutti gli altri e decide quanto sono rilevanti per il suo significato. Proprio questo è il salto concettuale: niente catene lunghe di singoli concetti, ma dipendenze globali calcolate in un colpo solo in parallelo.

su due modelli. Un primo modello[47], più piccolo e veloce, viene utilizzato per proporre una sequenza di token, che funge da bozza. Un secondo modello[48], più grande e accurato, verifica in un singolo passaggio le probabilità dei token proposti dal primo: se risultano coerenti con la propria distribuzione, li accetta in blocco; in caso contrario, interviene selettivamente sui token errati, riprendendo la generazione autoregressiva standard dal punto di divergenza. Il risultato finale mantiene la stessa qualità dell'output che si otterrebbe utilizzando esclusivamente il modello grande, ma con tempi di inferenza significativamente ridotti e, in media, un minore consumo computazionale, grazie alla diminuzione del numero di passaggi completi nel modello più costoso.

Tutte ler procedure citate si inseriscono in un insieme più ampio di strategie volte a ridurre il costo computazionale dell'inferenza, mantenendo la qualità dell'output entro limiti accettabili. Queste pratiche agiscono a livelli differenti, architetturale, algoritmico e operativo, ma condividono un obiettivo comune: limitare l'utilizzo delle risorse ai casi in cui esse risultano necessarie in base a stime di complessità e confidenza, evitando così calcoli ridondanti o eccessivamente conservativi. In questo senso, l'ottimizzazione dell'inferenza non riguarda più soltanto la velocità, ma diventa una leva fondamentale per migliorare l'efficienza energetica e la sostenibilità complessiva dei sistemi basati su modelli di grandi dimensioni.

La crescente attenzione all'efficienza dell'inferenza ha implicazioni dirette anche nella scelta dell'ambiente di esecuzione dei modelli, in particolare nel rapporto tra soluzioni cloud, in cui l'inferenza avviene in data center centralizzati, e soluzioni *edge*, in cui l'inferenza viene eseguita direttamente sul dispositivo.

L'idea di avvicinare l'inferenza all'utente nasce dalla possibilità di ridurre i costi associati al caricamento, alla trasmissione e alla memorizzazione delle

[47] Chiamato *draft model*.
[48] Chiamato *target model.*

informazioni tra il dispositivo e il server su cui avviene l'elaborazione. Come trade-off di queste soluzioni ci sono chiaramente dei vincoli stringenti in termini di potenza computazionale e consumo energetico locale.

Da qui sorge la domanda: *come è possibile caricare ed eseguire grandi LLM su dispositivi privi della potenza di un datacenter?* La risposta è che, a parità di funzioni, scala del modello e livello di accuratezza, ciò non è realisticamente possibile su dispositivi *edge*. La direzione attuale della ricerca si orienta quindi verso lo sviluppo di *Small Language Model* (SLM), ovvero modelli specializzati su domini specifici e caratterizzati da un numero ridotto di parametri.

Nell'articolo *"Edge-First Language Model Inference: Models, Metrics, and Tradeoffs"* vengono misurate le performance di uno SLM altamente specializzato[49] da 1,5 miliardi di parametri, confrontandole con quelle di un modello *general-purpose* con la stessa architettura ma composto da 7 miliardi di parametri.

I risultati mostrano come lo SLM raggiunga prestazioni pari o superiori al modello più grande, nelle query appartenenti al proprio ambito di specializzazione, pur richiedendo soltanto il 19,8% dello spazio di memoria (Jang e Morabito 2025). Da ciò si può evincere che uno SLM pre-addestrato su un dominio specifico può offrire prestazioni pari o superiori a quelle di un modello generico nella gestione di *query* appartenenti al proprio ambito di specializzazione.

In altre parole, un modello addestrato o perfezionato su un dominio ristretto può presentare una conoscenza inferiore in altri ambiti, ma ottenere una maggiore accuratezza nel campo applicativo per cui è stato progettato[50]. Le soluzioni cloud rimangono fondamentali per la qualità complessiva delle risposte e per la gestione di compiti complessi e generalisti; tuttavia, nel

[49] Su cui è stato operato un fine-tuning per la specializzazione in un ambito.
[50] Come ad esempio HealtBERT, FinBERT, DeepSeekMath, BioMistral o Code Llama.

contesto *edge*, il rapporto tra performance e costo per unità risulta spesso più efficiente.

L'adozione di soluzioni decentralizzate si presta in particolare a scenari industriali in cui è necessario un decisore capace di reagire in tempo reale agli input in ingresso, come nel caso di robot mobili specializzati, droni, veicoli autonomi e sistemi di sorveglianza. Oltre alla ridotta latenza, un ulteriore vantaggio dell'*edge* è la resilienza: il sistema continua a funzionare anche in presenza di problemi di connettività. Inoltre, evitando la trasmissione dei dati al di fuori del dispositivo, si mitigano le tematiche legate alla privacy, aspetto particolarmente rilevante in settori come la sanità e l'industria.

Infine, queste soluzioni permettono di alleggerire il carico sul cloud e di ridurre i costi operativi, poiché la trasmissione, la serializzazione, la memorizzazione e la replica dei dati comportano spesso un costo superiore rispetto al calcolo stesso. I limiti dell'*edge* derivano tuttavia dalla sua stessa natura: l'elevata specializzazione e le risorse limitate rendono difficoltosa la gestione di scenari caratterizzati da maggiore complessità, varietà di compiti, preferenze degli utenti e vincoli energetici stringenti. In tali casi, l'obiettivo diventa ottimizzare l'utilizzo delle risorse locali e delegare al cloud l'elaborazione in eccesso, dando origine a un'architettura ibrida *edge/cloud* in grado di preservare l'efficienza locale e, al contempo, sfruttare una maggiore potenza di calcolo nei momenti opportuni. In conclusione, vi sono numerose tecniche operative per tenere a bada i consumi di energia, ma come anticipato lo sviluppo dell'IA e il suo utilizzo necessitano anche di altre risorse. Nella prossima sezione vedremo come l'infrastruttura necessiti di grandi quantitativi di acqua, e come questi possano peggiorare situazioni già a rischio idrico.

2.3. L’impatto idrico dell’Intelligenza Artificiale

Negli ultimi anni, la diffusione crescente di soluzioni di intelligenza artificiale in numerosi settori, dalla finanza alla medicina fino ai trasporti, ha reso questo ambito sempre più centrale. All’aumentare dell’adozione di tali strumenti cresce inevitabilmente anche il consumo di risorse associate al loro utilizzo e, di conseguenza, il quantitativo di acqua impiegato dalle infrastrutture che li supportano.

Nonostante ciò, l’impronta idrica derivante dall’uso intensivo di modelli predittivi ha ricevuto finora un’attenzione relativamente limitata, sia in ambito accademico sia a livello normativo. L’impronta idrica dei sistemi di intelligenza artificiale può essere definita come il volume complessivo di acqua necessario alla produzione, all’allenamento, al funzionamento e alla manutenzione dell’infrastruttura fisica e digitale che ne consente l’operatività.

In particolare, è possibile distinguere tre componenti principali: l’impronta idrica diretta, legata all’acqua utilizzata per il raffreddamento dei server in loco; l’impronta idrica indiretta, associata alla produzione dell’energia elettrica impiegata nei data center; e infine l’acqua necessaria alla produzione degli hardware. L'IEA (2025) ha stimato che il consumo totale di acqua dei data center nel 2023 ammontava a 560 miliardi di litri, di cui due terzi (373 miliardi di litri) relativi al consumo indiretto di acqua e solo un quarto (140 miliardi di litri) al consumo diretto di acqua.

2.3.1. I consumi diretti di acqua

Per quanto riguarda la componente diretta, le operazioni dei server generano quantità significative di calore, che devono essere rimosse attraverso sistemi di raffreddamento dedicati. Una soluzione ampiamente diffusa è rappresentata dalle torri di raffreddamento, che dissipano il calore tramite l’evaporazione dell’acqua in sistemi a circuito aperto. È importante

sottolineare che tali impianti utilizzano prevalentemente acqua potabile per garantire un'elevata efficienza e che l'acqua evaporata non può essere recuperata.

L'evaporazione di acqua nei sistemi di raffreddamento dei data center varia significativamente in funzione delle tecnologie adottate e delle condizioni operative: nello studio di Mytton (2021) si stima che l'acqua consumata per ogni kilowattora di energia elettrica impiegata dai server può oscillare da valori trascurabili fino a circa 4,4 L/kWh, mentre il prelievo complessivo può raggiungere diverse centinaia di litri per kWh in impianti evaporativi meno efficienti.

Per ridurre il consumo idrico associato a questi sistemi, si è progressivamente diffuso il raffreddamento naturale assistito da evaporazione[51]. Tale approccio sfrutta l'aria esterna quando le condizioni ambientali lo consentono, ricorrendo all'evaporazione dell'acqua solo al superamento di determinate soglie di temperatura. In media, questi sistemi permettono di ridurre il consumo idrico rispetto alle torri di raffreddamento tradizionali; tuttavia, in presenza di elevate temperature e di picchi di domanda computazionale, la richiesta di acqua può aumentare sensibilmente. In tali situazioni, il raffreddamento può essere integrato con gruppi frigoriferi, che consentono di mantenere temperature operative adeguate ma comportano un maggiore dispendio energetico.

Un'ulteriore soluzione sempre più adottata è il *direct liquid cooling*, che prevede il raffreddamento diretto dei componenti elettronici tramite fluidi a elevata capacità termica. Questo approccio consente una dissipazione del calore più efficiente rispetto ai sistemi ad aria e può ridurre significativamente il fabbisogno di acqua per il raffreddamento evaporativo. Tuttavia, anche in questo caso emergono importanti trade-off: il *direct liquid cooling* richiede infrastrutture dedicate, maggiore complessità impiantistica

[51] *free cooling.*

e, in alcuni scenari, un aumento del consumo energetico complessivo, soprattutto se il calore rimosso non viene recuperato o riutilizzato.

Infine, il raffreddamento a secco a circuito chiuso rappresenta un'ulteriore alternativa, in grado di eliminare quasi completamente il consumo diretto di acqua all'interno del data center. Questa soluzione, però, aumenta il fabbisogno energetico necessario alla dissipazione del calore, spostando di fatto il consumo idrico dalla componente diretta a quella indiretta, legata alla produzione dell'energia elettrica.

A differenza dell'uso agricolo dell'acqua, che si basa in larga parte sulle precipitazioni naturali, i data center dipendono prevalentemente da risorse di acqua dolce provenienti da corpi idrici superficiali e sotterranei. Con l'aumento del numero di data center, della complessità dei modelli e della potenza di calcolo richiesta, si osserva un incremento parallelo della quantità di acqua necessaria per il raffreddamento delle infrastrutture che supportano l'IA.

Le implicazioni più significative della componente diretta dell'impronta idrica emergono a livello locale, poiché il prelievo di grandi quantità di acqua da fiumi, bacini e falde può alterare i deflussi naturali e ridurre la disponibilità per altri usi, quali l'agricoltura o l'accesso all'acqua potabile. La competizione per la risorsa può incidere negativamente sugli habitat acquatici, compromettendo la biodiversità e creando rischi ambientali nei territori interessati.

Sebbene, in termini assoluti, il prelievo idrico dell'IA, sia ancora piccolo rispetto ai consumi di settori come l'agricoltura o la produzione energetica tradizionale, la rapida crescita delle infrastrutture dedicate all'intelligenza artificiale suggerisce una forte accelerazione della domanda d'acqua nei prossimi anni (De Vries-Gao 2025). Studi sulla *water footprint* dei data center evidenziano come la gestione di grandi infrastrutture di calcolo richieda non solo energia, ma anche ingenti volumi di acqua, soprattutto per il raffreddamento delle apparecchiature, con impatti variabili in funzione della

tecnologia di raffreddamento adottata e della disponibilità locale di risorse idriche. Alcune analisi indicano che in paesi con climi secchi o stress idrico la presenza di data center e l'aumento delle richieste di raffreddamento possono esercitare una pressione sostanziale sulle disponibilità d'acqua superficiale e sotterranea, contribuendo a tensioni d'uso tra infrastrutture industriali, agricoltura e utenze domestiche.

Un esempio significativo di tale fenomeno si osserva in California, dove la concentrazione di data center dedicati all'intelligenza artificiale è tra le più alte negli Stati Uniti, dopo Texas e Virginia. In molte aree di questo stato, caratterizzate da periodi prolungati di siccità e stress idrico cronico, l'aumento della domanda d'acqua per il raffreddamento degli impianti ha suscitato preoccupazioni tra comunità locali e policy maker.

In questo senso, la localizzazione dei data center non può essere considerata una scelta puramente tecnico-economica, ma assume una valenza politica e territoriale rilevante. Report e indagini indipendenti stimano che i data center nordamericani abbiano aumentato il loro consumo di acqua in modo rilevante negli ultimi anni, con centinaia di miliardi di litri utilizzati annualmente solo per le operazioni di raffreddamento, e con trend in crescita legati all'espansione di carichi di lavoro intensivi come quelli dell'IA. In questo quadro, l'incremento della potenza di calcolo richiesto dai modelli più complessi si traduce in un aumento parallelo della quantità di acqua necessaria a mantenere sicuri i limiti termici delle infrastrutture. Pertanto, diventa fondamentale la valutazione non solo del profilo energetico delle nuove tecnologie, ma anche del loro impatto idrico locale e delle interazioni con le risorse naturali, in modo da orientare decisioni di pianificazione territoriale e politiche di sostenibilità delle infrastrutture digitali.

Per valutare in modo efficace l'efficienza dell'infrastruttura digitale diventa necessario disporre di indicatori chiari per misurare l'impatto idrico dei data center. In questa prospettiva è stato introdotto il *Water Usage Effectiveness* (WUE), un sistema quantitativo progettato per descrivere quanta acqua viene

impiegata a supporto dell'energia effettivamente utilizzata dai server. In termini operativi, il WUE mette in relazione il volume di acqua consumata all'interno di un impianto con i kilowattora destinati all'elaborazione informatica, offrendo una misura sintetica dell'efficienza idrica delle strutture.

Tuttavia, questo sistema di misura si concentra prevalentemente sui consumi idrici diretti del sito, come l'acqua utilizzata nei sistemi di raffreddamento o di umidificazione, e tende a escludere sia l'acqua necessaria alla produzione dell'energia elettrica sia quella impiegata lungo l'intera catena di approvvigionamento degli hardware. Questa impostazione rende il WUE uno strumento utile per confrontare diverse soluzioni tecnologiche all'interno dello stesso data center, ma ne limita la capacità di rappresentare l'impatto idrico complessivo del settore. Un impianto che presenta un WUE particolarmente basso potrebbe, ad esempio, fare ricorso a tecnologie di raffreddamento "a secco" e apparire quindi poco idrovoro, pur dipendendo da una rete elettrica che richiede grandi quantità di acqua per il proprio funzionamento.

Allo stesso modo, valori medio calcolati su base annuale non sono in grado di cogliere le variazioni stagionali né gli effetti locali in aree già soggette a stress idrico. Risulta quindi evidente che il WUE rappresenta una metrica utile ma parziale: consente di monitorare l'efficienza interna dei data center, ma non sostituisce un'analisi più ampia che includa l'acqua indiretta, il contesto geografico e l'intero ciclo di vita delle infrastrutture digitali.

In altre parole, per comprendere pienamente l'impronta idrica dei sistemi di calcolo contemporanei è necessario integrare il WUE con indicatori capaci di riflettere il legame tra tecnologia, territorio e sostenibilità nel lungo periodo. Arriviamo quindi alla componente indiretta dell'impronta idrica, legata alla produzione dell'energia elettrica utilizzata dai data center per il loro funzionamento.

2.3.2. I consumi indiretti di acqua

Le infrastrutture di calcolo contribuiscono al consumo di acqua non solo attraverso i prelievi diretti per il raffreddamento, ma anche in modo indiretto, tramite l'energia elettrica necessaria ad alimentarle, analogamente a quanto avviene per le emissioni di carbonio. Le diverse tecnologie di generazione elettrica presentano infatti intensità idriche molto eterogenee. Gli impianti termoelettrici alimentati a carbone, gas naturale o nucleare richiedono ingenti quantità di acqua per la produzione di vapore e per i sistemi di raffreddamento, consumandone una parte rilevante attraverso processi evaporativi. Al contrario, le fonti rinnovabili mostrano un profilo più variegato: mentre il fotovoltaico e l'eolico richiedono quantità minime di acqua in fase operativa, la produzione idroelettrica è associata a perdite significative per evaporazione, dovute alla presenza di bacini artificiali.

Alla luce di questi elementi, ridurre l'impatto idrico complessivo dei data center richiede una visione integrata, che affianchi soluzioni tecnologiche come il raffreddamento a secco a circuito chiuso a mix energetici caratterizzati da una bassa intensità idrica. In assenza di tale coordinamento, interventi mirati alla riduzione dei consumi diretti rischiano di tradursi in un semplice spostamento dell'impatto verso la componente indiretta.

Un'ulteriore dimensione, spesso meno visibile ma potenzialmente predominante, è rappresentata dalla catena di approvvigionamento della componentistica hardware. La produzione di semiconduttori, in particolare, richiede grandi volumi di acqua ultra-pura per la fabbricazione dei wafer, con ogni fase del processo produttivo che comporta numerosi cicli di lavaggio e pulizia.

Nonostante l'adozione di sistemi di recupero e riciclo, le pratiche industriali raggiungono generalmente tassi di riutilizzo dell'acqua compresi tra il 45% e il 50%, mentre le acque di scarico contengono sostanze chimiche che necessitano di trattamenti specializzati.

Di conseguenza, il consumo idrico lungo la filiera upstream del settore ICT può arrivare a dominare l'impronta idrica totale. Un esempio significativo è fornito dal *Sustainability Report 2024* di Apple, nel quale l'azienda riporta che circa il 99% della propria impronta idrica complessiva è associata alle fasi di produzione dell'hardware, piuttosto che ai consumi diretti delle strutture operative.

Questo risultato riflette l'impostazione tipica delle analisi di *Life Cycle Assessment* adottate nei report di settore, nelle quali la filiera a monte contribuisce in misura nettamente superiore all'uso complessivo di acqua rispetto alle operazioni interne. Tuttavia, tale percentuale non implica che i consumi idrici diretti siano trascurabili in termini assoluti, né che siano assenti impatti localizzati rilevanti nelle aree di produzione o di esercizio delle infrastrutture.

2.3.3. I consumi operativi: allenamento e inferenza

Una volta definiti i confini entro cui si collocano i consumi idrici, è possibile passare a un'analisi quantitativa delle fasi maggiormente idrovore. Per quanto riguarda l'allenamento di un singolo modello di grandi dimensioni, le stime disponibili indicano che, in funzione della complessità e dell'architettura, possono essere impiegati volumi d'acqua dell'ordine dei milioni di litri. Ad esempio, per GPT-3 è stato stimato un consumo diretto pari a circa 5,4 milioni di litri, di cui circa 700.000 litri evaporati nei sistemi di raffreddamento dei data center (Li et al., 2025). È opportuno sottolineare che tali valori derivano da ricostruzioni ex post e da modelli di stima, e non da misurazioni dirette, e devono pertanto essere interpretati come ordini di grandezza.

Nello stesso studio, considerando congiuntamente i prelievi idrici diretti e indiretti associati all'intero settore dell'intelligenza artificiale, vengono riportate stime globali comprese tra 4,2 e 6,6 miliardi di metri cubi su base annuale, mentre la sola componente diretta, legata all'evaporazione per il

raffreddamento dei data center, è stimata tra 0,38 e 0,60 miliardi di metri cubi. Questi valori si riferiscono prevalentemente alla fase di allenamento dei modelli, che avviene in modo intermittente, sebbene possa ripetersi nel tempo attraverso operazioni di retraining o aggiornamento.

Diversamente, la fase di inferenza è continua e dipende dall'interazione costante degli utenti con i modelli distribuiti a livello globale. Per questo motivo, la stima del consumo idrico associato al funzionamento dei sistemi di IA riveste un ruolo particolarmente rilevante. Tuttavia, le analisi quantitative sull'inferenza sono ancora limitate e fortemente dipendenti dal contesto infrastrutturale e geografico. Alcuni lavori riportano ordini di grandezza per singole interazioni, indicando consumi nell'ordine di alcune centinaia di millilitri d'acqua per decine di query in specifici scenari di servizio, e mostrano come fattori quali l'efficienza idrica dei data center (WUE) e il mix energetico locale incidano in modo determinante sui risultati (Jiang et al., 2024).

2.3.4. *Best practice* per l'impatto idrico

Adattare quanto più possibile i data center ai contesti climatici ed energetici locali diventa fondamentale per migliorare l'efficienza dei sistemi di IA così da non doversi limitare solamente a soluzioni software o computazionali. Attraverso una pianificazione strategica e un'analisi attenta del territorio in cui realizzare nuove infrastrutture, è possibile intervenire sull'efficienza idrica delle strutture, incidendo indirettamente anche sui costi energetici complessivi. Allo stesso tempo, monitorare i consumi e i prelievi di acqua consente di evitare l'acuirsi di situazioni già caratterizzate da stress idrico, riducendo l'impatto ambientale sui territori e sulle comunità locali.

Un ulteriore ambito di intervento riguarda l'adozione di sistemi di riciclo dell'acqua sempre più avanzati e orientati alla circolarità. Soluzioni come il riutilizzo delle acque grigie, l'impiego di acqua piovana o, in alcuni contesti, di acqua desalinizzata permettono di ridurre la quota di risorsa idrica

prelevata.

In una prospettiva di economia circolare più ampia, i data center possono inoltre essere integrati con le reti di teleriscaldamento, consentendo il recupero del calore di scarto prodotto dalle attività di calcolo. In questo senso, un position paper redatto da TEHA Group e A2A (2025) stima che in Italia, entro il 2035, il recupero del calore dai data center potrebbe generare tra 4,6 e 9,5 TWh di energia termica, sufficienti a servire circa 800.000 famiglie e a contribuire a una riduzione delle emissioni pari a circa 2 milioni di tonnellate di CO_2. In alcuni Paesi europei questa soluzione è già operativa: come il caso finlandese della città di Espoo dove Microsoft ha annunciato la realizzazione di un data center in grado di fornire calore a circa 100.000 abitanti.

Infine, risulta fondamentale continuare a monitorare e analizzare in modo sistematico i consumi associati al settore dell'intelligenza artificiale. Solo attraverso una misurazione accurata è possibile individuare le aree di intervento più efficaci per ridurre il consumo idrico, consentendo agli operatori dei data center di ottimizzare le proprie operazioni e migliorare l'efficienza complessiva.

In questo senso, l'infrastruttura rappresenta un fattore determinante per la sostenibilità dell'inferenza dell'IA. Sebbene l'ottimizzazione dell'architettura dei modelli possa produrre risultati teorici promettenti, la loro applicazione concreta dipende in larga misura da elementi materiali, quali l'efficienza complessiva delle infrastrutture, l'adozione di fonti energetiche rinnovabili e le prestazioni dell'hardware impiegato.

Di conseguenza, il perseguimento di un'intelligenza artificiale realmente sostenibile richiede un approccio integrato che affronti, in via prioritaria, aspetti quali l'utilizzo e il riciclo di hardware più efficiente, l'adozione di strategie di gestione dell'acqua più sostenibili e un maggiore accesso a fonti energetiche a basso impatto ambientale.

A questi elementi tecnici deve necessariamente affiancarsi un adeguato quadro normativo, fondato su pratiche di valutazione trasparenti e su criteri condivisi per la progettazione e la gestione delle nuove infrastrutture digitali. In tale contesto, le istituzioni, a tutti i livelli di governo, dovrebbero promuovere l'introduzione di regolamentazioni basate su dati affidabili, definendo soglie di impatto ambientale ammissibile per l'inferenza dei modelli di intelligenza artificiale in termini di consumo energetico, utilizzo di risorse idriche ed emissioni di carbonio.

Sebbene tali limiti possano apparire restrittivi, essi risultano in larga misura raggiungibili attraverso l'adozione delle soluzioni tecnologiche e organizzative analizzate nel corso del capitolo. Proprio al fine di rendere possibile l'elaborazione di un quadro regolatorio solido ed efficace, la trasparenza assume un ruolo centrale: allo stato attuale, infatti, non esiste una normativa che imponga una rendicontazione sistematica e standardizzata dei quantitativi di energia, acqua ed emissioni di carbonio associati all'allenamento, all'implementazione e al funzionamento dei modelli di IA

2.4. Conclusioni

Il quadro tracciato in questo capitolo mostra come l'impatto ambientale dell'IA non possa essere ridotto a un unico numero né ricondotto esclusivamente al "costo" dell'addestramento: esso emerge dall'intreccio tra una componente materiale legata alla filiera dell'hardware comprendente: estrazione e raffinazione di minerali critici, produzione, logistica e fine vita; e una componente operativa legata all'energia necessaria per training, retraining e inferenza, con dinamiche diverse per durata, scala e variabilità dei carichi. Se da un lato l'efficienza energetica delle infrastrutture e la progressiva decarbonizzazione delle reti possono ridurre le emissioni operative, dall'altro la crescita della domanda di acceleratori e la rapida obsolescenza tecnologica tendono a rendere sempre più centrale la quota incorporata e gl impatti indiretti, inclusi quelli associati alla gestione dei

rifiuti elettronici e ai limiti strutturali del riciclo.

A complicare ulteriormente la valutazione interviene la dimensione metodologica: unità funzionale, confini del sistema e assunzioni "di default" influenzano in modo decisivo le stime, rendendo fragile il confronto tra risultati e aprendo spazio a letture divergenti o a semplificazioni comunicative. In questo scenario, la sostenibilità operativa non dipende da un singolo intervento, ma da un insieme coordinato di scelte lungo la catena tecnica: dal modello, agli algoritmi, all'hardware, all'efficienza del datacenter fino alla localizzazione energetica; che possono ridurre sia l'energia necessaria sia l'intensità carbonica dell'energia utilizzata.

Tuttavia, proprio perché tali leve sono distribuite tra attori diversi e spesso coperte da opacità informativa, la discussione sulla sostenibilità rischia di restare episodica o non verificabile. Per questo, diventa inevitabile, dopo aver mappato le principali sorgenti di impatto e le leve di mitigazione, affrontare il problema della standardizzazione del settore e di metriche condivise e, più in generale, della governance dell'IA, così da rendere comparabili le valutazioni, verificabili le dichiarazioni, e orientabili le scelte che determinano gli impatti lungo l'intero ciclo di vita e di utilizzo di questi sistemi.

3. La regolamentazione e la governance dell'intelligenza artificiale

Con questo capitolo arriviamo alla fase di regolamentazione dell'IA, ma prima di entrare nel merito della questione conviene compiere un passo indietro per comprendere meglio il settore tramite gli attori, le dinamiche e i conflitti che lo regolano.

Partiamo presentando le aziende hi-tech tradizionali come Amazon, Google, Microsoft e Meta che dominano il settore digitale. Tramite una disponibilità

impressionante di capitale, un'innovazione pianificata e dei rendimenti di scala crescenti questi attori sono in grado di sostenere enormi costi fissi, oltre a disporre delle infrastrutture e dei canali di distribuzione. Di fatto i membri di questo oligopolio digitale sanno che, se anche non saranno loro a inventare la prossima tecnologia rivoluzionaria, di certo ne controlleranno la messa a punto e la gestione.

I nuovi attori si trovano quindi davanti a una barriera d'entrata, dovuta al settore fortemente *capital-intensive* da un lato e gli incumbent[52] che detengono il potere strutturale dall'altro. Da qui *firm* come OpenAi, Anthropic o Mistral, sono riuscite ad inserirsi nel panorama con un capitale sociale e cognitivo, piuttosto che prettamente materiale.

Questi attori fanno da avanguardia: producono discontinuità, ridefiniscono il linguaggio e generano valore ridisegnando ciò che conta. Invece di distruggere ciò che veniva prima, si inseriscono in un sistema di cooptazione per cui l'innovazione nasce ai margini, dimostra di funzionare, per poi essere assorbita o finanziata dai grandi attori; andando a spostare un po' più in là il limite oltre il quale avverrà la prossima innovazione.

Nessuna delle start up di IA ha sostituito Microsoft o Google, ma esse sono riuscite ad inserirsi in un contesto quasi inaccessibile, guadagnando legittimazione. Ogni nuova *firm* si è mossa secondo una propria strategia, come ad esempio OpenAi che ambisce tutt'ora al ruolo di incumbent. La competizione in questo caso non è sui singoli prodotti, ma va ad articolarsi in un'ottica di dipendenza sistemica per cui non si vuole distruggere il mercato, ma proporsi come base per i mercati futuri.

Anche Anthropic si sta muovendo in ottica da co-incumbent creandosi una nicchia di mercato in cui affidabilità e credibilità istituzionale sono valori di posizionamento. Anche in questo caso una stretta relazione con i grandi attori del mercato rende Anthropic un attore credibile, la cui reputazione è

[52] Un operatore già affermato in un mercato: un'azienda che occupa una posizione stabile e spesso dominante.

sostenuta da forti investimenti in sicurezza e *alignment* dichiarato.

Infine, troviamo Mistral che assume un ruolo da attore sub-incumbent per cui non distrugge il regime, ma ne impedisce la chiusura totale. Di fatto, grazie a uno strategico utilizzo di soluzioni *open-source*[53] riesce a mantenere un capitale scientifico e simbolico molto alto. Il tutto unito alla competizione lungo traiettorie come una forte efficienza, un'apertura selettiva e alla legittimazione alternativa dovuta al posizionamento geopolitico. Proprio in questo scenario emerge una questione cruciale: la concentrazione dei dati e della capacità di calcolo in poche aziende rende il settore strutturalmente elitario.

In un contesto del genere anche se si riuscisse a realizzare un'idea geniale, finché non si possiedono: i datacenter, dei dataset massivi e l'energia necessaria, si rimarrà nel proprio garage senza poterla mettere a punto. Questa dinamica di accentramento rende il settore IA molto più vicino all'industria pesante di inizio '900 piuttosto che al mito romantico dell'innovazione come forza di "distruzione creativa".

Nonostante questa somiglianza con un modello Schumpeter Mark II, rimangono delle differenze sostanziali: la conoscenza è ancora parzialmente un bene pubblico[54] e circola sottoforma di pubblicazioni, bechmark, open review e conferenze. In questo senso la presenza di modelli *open source* rappresenta un meccanismo di controbilanciamento endogeno alla concentrazione tipica del Mark II.

L'*open source* mantiene infatti attivi degli spazi di sperimentazione distribuita, consentendo a università, centri di ricerca e piccoli attori di partecipare ai processi di innovazione, pur non avendo accesso alle infrastrutture più costose. Questa dinamica non fa di certo nascere modelli di

[53] Il termine "*open source*" indica un software il cui codice sorgente è reso disponibile pubblicamente con una licenza che ne consente l'uso, lo studio, la modifica e la redistribuzione, a condizioni definite.

[54] Invece di essere opaca e proprietaria.

frontiera, ma assieme ad altre infrastrutture di *knowledge commons*[55], come dataset condivisi, benchmark pubblici e standard aperti permettono di creare spazi in cui la comunità scientifica e la società possono osservare, valutare e intervenire.

Con l'utilizzo di dati comparabili, metriche condivise ed evidenze riproducibili diventa possibile studiare e regolare il settore, in questo senso, l'*open source* e i *commons* funzionano quindi come *regulatory enablers*[56].

Le normative sull'IA, soprattutto in ambito ambientale e di trasparenza, cercano di formalizzare pratiche già tecnicamente e socialmente possibili come: audit, reporting, valutazione del ciclo di vita e tracciabilità. In mancanza di *commons* conoscitivi condivisi questi obblighi rischiano di rimanere puramente formali.

Inoltre, vi è il rischio di *regulatory capture*[57], situazione che emerge quando le autorità di regolamentazione dipendono in modo strutturale dalle informazioni, dalle competenze e dalle metriche fornite dagli stessi attori regolati, compromettendo la capacità di tutela dell'interesse pubblico. Un settore fortemente caratterizzato da asimmetria informativa e accesso privilegiato, unito a un contesto di altissima complessità tecnica e velocità di innovazione, diviene sicuramente un terreno fertile per questo tipo di problematiche. Dal momento in cui le istituzioni, nazionali o sovranazionali, non riescono a stare al passo con gli attori dominanti, chi è in grado di spiegare i pericoli, i rischi e le criticità di una nuova tecnologia ottiene accesso diretto al potere regolatorio.

In questo contesto la famosa "Pause letter" del marzo 2023, redatta dal *Future of life Institute*, non si inserisce come atto per richiedere uno sforzo

[55] Un *knowledge commons* è una risorsa informativa condivisa e gestita collettivamente secondo regole che ne permettono accesso, uso e riuso, con l'obiettivo di favorire produzione e diffusione della conoscenza evitando che venga completamente privatizzata o chiusa.

[56] Sono norme, strumenti o meccanismi di *policy* che rendono possibile o più facile raggiungere un certo obiettivo, invece di imporre solo divieti o obblighi diretti.

[57] La *regulatory capture* si verifica quando un'autorità di regolazione finisce per servire soprattutto gli interessi del settore che dovrebbe regolare, invece dell'interesse pubblico.

normativo che regoli il settore, ma propone invece una pausa volontaria globale allo sviluppo dell'IA. In questo senso possiamo parlare di una mossa puramente discorsiva più che una proposta operativa: il suo effetto non è stato quello di regolamentare o fermare l'IA, ma quello di sottolineare la necessità che tutti gli attori coinvolti siano responsabili e legittimati, in grado di capire il rischio sistemico dovuto all'IA. Di fatto, in questo modo si prepara il terreno per la *regulatory capture*, andando a definire in primis chi è competente nel parlare di IA e solo in seconda istanza come regolamentare il settore. Il significato della lettera non è quindi affermare che vi sia bisogno di controllo, ma di suggerirsi implicitamente come guida per il processo, incastonando nella futura legislazione il linguaggio, i concetti e le priorità dell'industria.

La lettera ha suscitato sentimenti contrastanti attraendo molto sostegno dal pubblico e dalla politica, mentre per quanto riguarda la comunità accademica sono emerse più posizioni: tra un generale sentimento di preoccupazione a una forte critica nei confronti delle modalità. Seppur molti scienziati chiedessero da tempo un'azione di qualsivoglia tipo nei confronti dell'IA, la vaghezza tecnica con cui si è invitato a "non allenare modelli più potenti di GPT-4", senza indicare alcuna metrica scientifica, unita all'inefficacia pratica di implementare una decisione senza canali di *enforcement,* ha portato ad un giudizio negativo della proposta.

Infine, in un settore già contraddistinto da forti barriere d'entrata una moratoria discorsiva di questo tipo finisce per inserirne un'altra: quella della sicurezza. In questo modo si favoriscono quegli attori che possono permettersi meccanismi di conformità, audit e squadre di legali già pronte a muoversi, mentre si penalizzano gli attori nuovi con regole formalmente neutrali, ma strutturalmente selettive.

3.1. La trasparenza

La parola “trasparenza” nel dibattito pubblico riguardo l’IA sta assumendo sempre più importanza, anche se viene spesso utilizzata a sproposito o in modo improprio. Partiamo quindi dando una definizione di trasparenza: dal punto di vista sociologico la trasparenza è una proprietà relazionale dei processi sociali per cui azioni, decisioni o informazioni diventano accessibili e intelligibili a determinati attori, secondo codici condivisi, all’interno di specifici rapporti di potere e fiducia. Entrando nello specifico questo carattere relazionale non coincide semplicemente con la disponibilità di informazioni, ma si spiega con il rapporto sociale che rende le informazioni comprensibili ed interpretabili da più attori.

La trasparenza diventa quindi una forma di comunicazione che dipende da chi comunica, da chi riceve e dai codici che ne permettono la comprensione. Ad esempio, un documento può essere pubblico e facilmente consultabile, ma allo stesso tempo rimanere sociologicamente opaco se il destinatario non possiede le competenze per decodificarlo. La richiesta di una maggiore trasparenza implica quindi un’asimmetria di potere per cui chi osserva non ha accesso alle ragioni di chi decide.

Nel contesto dell’IA questa asimmetria si fa particolarmente forte poiché il funzionamento dei modelli ad apprendimento automatico è tecnicamente opaco e raramente compreso appieno dagli utenti che li utilizzano (Aula e Erkkila, 2024). In questo senso, una maggiore trasparenza aiuta a trasferire la “fiducia” in questa tecnologia dalla comprensione effettiva e totale dei modelli, verso una serie di pratiche che ne certifichino l’affidabilità.

Proprio per questa ragione l’importanza di procedure, istituzioni, audit e standard diviene focale perché, quando un’organizzazione introduce un sistema di IA “trasparente”, non sta solo mostrando come essa funziona, ma sta dichiarando anche di essere valutabile, contestabile e regolabile. Inoltre, questo processo ha senso solo se esistono attori capaci di giudicare come:

autorità di controllo, tribunali e cittadini informati. Senza questa reciprocità, la cosiddetta trasparenza dell'IA si riduce a una messa in scena tecnica, spesso utilizzata come strategia di legittimazione dagli attori principali del settore. In particolare, nel lavoro di Jobin et al. (2019), viene analizzato come il settore privato tenda a privilegiare soluzioni tecniche.

La trasparenza viene trattata infatti come un problema da risolvere all'interno del sistema con una famiglia di approcci ingegneristici, piuttosto che come una pratica sociale di controllo esterno. Alcune pratiche puntano alla spiegabilità e interpretabilità dei modelli per cui le aziende puntano a rendere il funzionamento dell'algoritmo "spiegabile" a posteriori tramite modelli interpretabili, metodi di *feature importance*, visualizzazioni delle decisioni e spiegazioni locali delle predizioni. L'idea alla base di queste prassi è che la trasparenza possa essere incorporata nel codice stesso, senza bisogno di interventi umani o istituzionali esterni.

Un'altra soluzione ricorrente degli attori privati è la produzione di schede tecniche, report o *model cards* che descrivono gli scopi, i limiti, i dati usati e le prestazioni del sistema. Anche qui la trasparenza è intesa come fornitura di informazioni strutturate e non come verifica indipendente. In questo modo però la documentazione rimane scritta da chi "costruisce" il sistema diventando una narrazione controllata più che un metodo di accertamento. Inoltre, report e *model cards* fanno riferimento alle prestazioni in laboratorio, mentre la realtà all'esterno è molto più complessa. Problematiche come un utilizzo improprio, gli eventuali adattamenti nel tempo e le interazioni con contesti sociali concreti non vengono catturati da questo tipo di documentazione che si limita a fornire una fotografia del contesto ideale.

In aggiunta un'opera di standardizzazione significa in qualche misura astrazione, per cui va scelto cosa importa e cosa no: metriche, standard e benchmark diventano ciò che è visibile e misurabile, mentre tutto ciò che non rientra all'interno di queste categorie scompare o viene considerato irrilevante.

Infine, se la trasparenza nasce come risposta ad un'asimmetria informativa il linguaggio di questo tipo di documentazione resta specialistico e opaco per cui non vi è maggiore comprensibilità da parte dell'utente finale o del regolatore. Proprio su questa "apertura" mascherata si articola una dinamica per cui la rendicontazione diventa requisito normativo o reputazionale, atto a soddisfare uno standard più che a illuminare il sistema.

In questo senso si articola la preferenza da parte del settore privato verso sistemi interni di auditing automatizzato[58] per il controllo e il monitoraggio continuo delle prestazioni. Per quanto siano controlli reali, dal momento in cui sono effettuati internamente dall'organizzazione che sviluppa o utilizza il sistema non sono verificabili senza un accesso indipendente. Senza accesso ai dati, al modello o alla possibilità di audit esterni non vi è alcun controllo su ciò che è dichiarato, per cui la documentazione senza audit rimane mera fiducia formalizzata, utile in ambienti cooperativi, ma fragile in contesti di conflitto di interessi.

Sempre secondo Jobin et al. (2019), a differenza del settore privato, le autorità pubbliche e le organizzazioni non governative (ONG) tendono a preferire soluzioni non solamente tecniche. Partono cioè da un presupposto diverso da quello del settore privato: la trasparenza e la responsabilità non possono essere garantite esclusivamente dall'architettura del sistema, ma richiedono un controllo sociale esterno. In questo caso gli enti pubblici insistono sulla necessità in primis di verifiche condotte da soggetti terzi, indipendenti dagli sviluppatori e dagli utilizzatori dei sistemi di IA.

L'audit esterno diventa così un processo istituzionale che può includere l'accesso ai dati, al modello e alle decisioni prodotte. Il punto centrale è l'indipendenza per cui chi controlla non deve coincidere con chi trae beneficio dal sistema. Un altro elemento ricorrente è la richiesta della supervisione umana: le decisioni automatizzate devono poter essere

[58] test di robustezza, metriche di fairness, controlli statistici sui bias, monitoraggio continuo delle prestazioni

esaminate, contestate e, se necessario, annullate da esseri umani. Qui la trasparenza non serve solo a "capire come funziona l'algoritmo", ma a rendere possibile l'assunzione di responsabilità da parte di attori identificabili.

Il punto cruciale in questo caso è il riconoscimento dei diritti per i soggetti coinvolti, per cui un utente ha diritto a sapere che un sistema algoritmico è in uso e ha diritto a ricevere spiegazioni comprensibili circa il suo funzionamento.

Infine, nelle linee guida elaborate da ONG e istituzioni pubbliche, la trasparenza è strettamente connessa agli obblighi formali di rendicontazione e di responsabilità pubblica. L'utilizzo di sistemi di IA in questa prospettiva è considerato legittimo solo nella misura in cui gli attori che li sviluppano o li impiegano siano in grado di spiegare e giustificare pubblicamente le ragioni del loro utilizzo, il quadro giuridico che ne consente l'adozione, le finalità perseguite e gli effetti sociali che ne derivano.

La trasparenza non viene quindi concepita come una caratteristica tecnica opzionale, ma come una condizione necessaria nell'ottica di un esercizio responsabile e del mantenimento della legittimità democratica. Rendere conto delle decisioni automatizzate significa infatti rendere visibili le scelte, le priorità e i valori incorporati nei modelli, consentendo così agli attori sociali di valutare, contestare e, se necessario, limitare il loro impiego. La trasparenza assume una funzione prettamente politica e istituzionale, in quanto abilita il controllo pubblico nei confronti di una tecnologia che incide in modo crescente sui diritti, sulle opportunità e sulle forme di partecipazione sociale. Nel prossimo paragrafo la discussione si articolerà riguardo i rapporti tra l'IA e la sostenibilità. In questo senso la trasparenza nella raccolta e nella comunicazione dei dati diviene fondamentale per la definizione degli standard da seguire.

3.2. L’Intelligenza Artificiale sostenibile

Mentre le questioni sociali ed etiche dell’IA sono state ampiamente affrontate durante le prime due ondate di studi riguardanti questo settore, la dimensione della sostenibilità non è stata ampiamente affrontata. Per poter studiare il rapporto fra sostenibilità e IA è fondamentale capirne la natura duale per questo motivo facciamo affidamento alla definizione del campo di ricerca dell’“IA sostenibile” di van Wynsberghe:

“Suggerisco che l'"IA sostenibile" sia un campo di ricerca che si applica alla tecnologia dell'IA (l'hardware che la alimenta, i metodi per addestrarla e l’effettiva elaborazione dei dati) oltre che alla sua implementazione, affrontando le questioni di sostenibilità e di sviluppo sostenibile che la riguardano. Suggerisco inoltre che “l'IA sostenibile” non si occupi esclusivamente delle applicazioni dell'IA, ma che ne affronti l'intero ciclo di vita, dalla sostenibilità della progettazione, dell’addestramento, dello sviluppo, del riadattamento, fino, solo in ultima istanza, all’implementazione e ai diversi impieghi dell'IA.” (Van Wynsberghe, 2021).

Con questa proposta van Wynsberghe sottolinea come l’IA possa essere uno strumento per promuovere obiettivi di sostenibilità ambientale, sociale ed economica. Così da raggiungere gli obiettivi di sviluppo sostenibile (SDG) dell’Agenda 2030, grazie ad applicazioni che ottimizzano l’uso delle risorse naturali, riducono emissioni o migliorano l’accesso all’energia pulita. Allo stesso tempo, nella seconda parte della citazione viene posta l’attenzione sull’impatto intrinseco dell’IA.

In questo caso l’oggetto di studio è la tecnologia IA, in quanto tale e la sua sostenibilità ambientale e sociale, per cui risulta importante indagare i consumi e le emissioni attribuibili all’addestramento e all’inferenza, oltre a quelle dell’infrastruttura e del ciclo di vita dell’hardware.

La letteratura emergente suggerisce che per avere un quadro completo bisogna guardare entrambi gli aspetti insieme: è possibile progettare IA che

sia al contempo uno strumento per la sostenibilità e intrinsecamente sostenibile nelle sue pratiche di sviluppo e uso.

Proprio a livello di letteratura negli ultimi anni si è registrato un forte aumento delle pubblicazioni per parole chiave come "IA", "sostenibilità" e "governance": i risultati nel 2022 sono triplicati rispetto al 2020 e sono cresciuti addirittura di dieci volte rispetto al 2019 (Lucivero, 2024). La maggior parte di questi contributi esplora i modi in cui l'IA offre soluzioni ai problemi legati al cambiamento climatico, alla biodiversità e al degrado ambientale. Tuttavia, gli studiosi sottolineano sempre più la necessità di una comprensione critica della relazione tra l'IA e gli SDG con particolare attenzione all'impatto ambientale di questa nuova tecnologia pervasiva.

La sostenibilità nella ricerca accademica sembra focalizzarsi maggiormente sugli aspetti tecnici ed entro i confini dei costi operazionali. In particolare, sul miglioramento dell'efficienza energetica; come segnala l'analisi bibliometrica di Samuel et al (2022) in cui si attesta che nella quota di articoli accademici che parla di sostenibilità dell'IA, l'85% adotta un approccio tecnocentrico. Chiaramente interrogarsi circa l'impatto operazionale di messa a punto, allenamento e inferenza dei modelli è fondamentale per l'analisi del settore. Con questa dipendenza dallo studio dei consumi operativi si rischia però di tralasciare interi segmenti di produzione, con effetti indiretti ed esternalità negative che non possono essere ignorati.

Come abbiamo visto nel secondo capitolo, la ricerca nel settore del *machine learning* è fortemente approntata ad un approccio ingegneristico, mentre sono sottovalutati gli apporti delle scienze sociali, fondamentali per una visione dell'IA come processo sociopolitico più che mero prodotto ICT. Nonostante queste valutazioni, la misurazione tecnica dei dati rimane fondamentale: nel prossimo paragrafo andremo infatti ad analizzare come ricerca e *standard organizations* lavorino insieme per la creazione di una metodologia robusta, che permetta di valutare l'impatto ambientale dell'IA.

3.3. Il *Greenhouse Gas Inventory* per l'Intelligenza Artificiale

Il reporting delle emissioni climalteranti si basa sul *Greenhouse Gas* (GHG) *Protocol*, lo standard più diffuso a livello internazionale per misurare e comunicare le emissioni di gas serra. Il GHG *Protocol* fornisce principi, definizioni e regole operative per stabilire i confini di rendicontazione e scegliere i metodi di calcolo garantendo coerenza e confrontabilità tra organizzazioni e periodi diversi. In questo contesto, il GHG *inventory* è l'insieme di dati che quantifica, tipicamente in CO_2 equivalente (CO_2e), le emissioni associate alle attività di un'organizzazione in un dato intervallo temporale, distinguendo tra emissioni dirette e indirette.

La classificazione più utilizzata è quella per Scope 1, Scope 2 e Scope 3, che ripartisce le emissioni rispettivamente in: emissioni dirette da fonti possedute o controllate (Scope 1), emissioni indirette legate all'energia acquistata (Scope 2), e altre emissioni indirette lungo la catena del valore (Scope 3). Questa struttura ambisce a rendere trasparenti le principali sorgenti emissive e di orientare sia le strategie di riduzione, sia la comunicazione verso *stakeholders*, clienti e autorità regolatorie.

Inoltre, vi è anche il *Corporate Value Chain Standard*[59] del *GHG Protocol*, che nasce dalla richiesta degli *stakeholder* per una maggiore trasparenza da parte delle aziende, queste ultime sono invitate a fornire un quadro più completo delle proprie emissioni aziendali distinguendo le emissioni per categorie come: i servizi acquistati o i beni strumentali. Questo ulteriore standard fornisce indicazioni alle aziende che desiderano ampliare la propria rendicontazione a livello aziendale per includere le emissioni a monte e a valle.

Per quanto riguarda il GHG *inventory* per l'IA si articola su tre direttrici: il *corporate* reporting, il reporting infrastrutturale, riguardante i datacenter, e

[59] Specifico per lo Scope 3 per l'indagine a valle e a monte della catena del valore.

infine il reporting di *workload*[60] del modello. Questi tre livelli non sono mondi separati, ma tre perimetri di misura che si incastrano per analizzare dal punto di vista macro fino al micro.

3.3.1. Il *Corporate reporting*

Per quanto riguarda il *corporate* reporting ci troviamo in un ambito più "adulto"; le grandi *corporation* informatiche, infatti, hanno già a che fare con regolamentazioni ambientali strutturate da seguire. Il reporting *corporate* è quello che comunica consumi, emissioni e obiettivi delle aziende. Tramite la divisione in Scope 1, 2 e 3 include anche la parte più scomoda: gli impatti indiretti lungo la catena del valore. Come abbiamo visto nel secondo capitolo, per l'IA questo è fondamentale, perché una quota importante dell'impronta non deriva solamente dall'elettricità che viene utilizzata, ma proviene dalla produzione di hardware, dalla distribuzione, dall'utilizzo e dal fine vita.
Inoltre, il reporting a livello *corporate* proprio in quanto aggregato, permette di rilevare un rischio classico: un'azienda che migliora l'efficienza per unità, ma cresce così tanto che le emissioni assolute aumentano comunque.
Questo tipo di reporting rimane "a bassa risoluzione", per cui se si vuole sapere quanto pesa un determinato modello o quanto costa a livello ambientale un servizio bisognerà scendere più nel dettaglio, abbandonando la dimensione macro. Rispetto agli Scope 1 e 2 nei report *corporate* lo Scope 3 dipende fortemente da stime e assunzioni per i conteggi che portano a una maggiore incertezza e una minore comparabilità. Da qui i due rischi principali: da una parte la voce più grande del bilancio, lo Scope 3, è anche la più "costruita" per cui si rischia di trasformare un'analisi quantitativa in un modello, dall'altra nel momento in cui più aziende scelgono diversi set di assunzioni, dalla più ottimista alla più pessimista, o perimetri differenti si

[60] L'insieme del lavoro computazionale effettivo che il modello genera o richiede in un certo contesto, cioè quanta elaborazione serve per addestrarlo, farlo girare sull'hardware e mantenerlo in esercizio.

perde la reciproca comparabilità.

Queste problematiche rendono il *corporate* reporting inutile, ma ne vanno riconosciuti i limiti: due aziende possono essere entrambe "in regola", ma raccontare storie numericamente diverse perché hanno confini e metodi diversi. Il rischio comunicativo principale è che il reporting diventi solamente un linguaggio da bilancio: corretto, ma facilmente ottimizzabile per apparire bene

3.3.2. Il *reporting* infrastrutturale

Se il *corporate* si occupa del livello macro, il reporting infrastrutturale è uno zoom sullo spazio operativo: i datacenter. In questo dominio si ha a che fare con grandezze molto concrete: energia assorbita, efficienza dell'impianto, gestione termica, acqua e recupero di calore. La misurazione è facilmente realizzabile e per questo utilizzata lungo tutte le dimensioni del datacenter. Questa vasta gamma di metriche[61] e misurazioni permette ai gestori dei datacenter di agire sull'efficienza da più punti di vista diversi, garantendo risultati rapidamente.

Allo stesso tempo l'efficienza della "macchina" non è l'unica cosa che incide sui consumi, se valori come PUE e WUE passano dall'essere strumenti diagnostici a soli obiettivi, si rischia di perdere di vista l'impatto reale. Infatti, quando l'efficienza riduce il costo marginale di una tecnologia, spesso ne consegue l'aumento dell'uso totale portando ad effetti di *rebound*[62]. Un miglioramento di efficienza può quindi essere divorato da una crescita esplosiva di domanda computazionale. In questo senso, il reporting per i datacenter permette di capire quanto è efficiente l'infrastruttura, senza però valutare le operazioni che vengono realizzate al suo interno. Inoltre, anche in questo caso va valutato l'impatto incorporato dell'hardware, oltre alla sua

[61] Come i già citati PUE e WUE.

[62] Per quanto riguarda i settori digitali il *rebound* è spesso molto forte poiché la domanda è elastica, i costi si abbassano velocemente e la scala cresce in assenza di limiti fisici immediati.

sostituzione e manutenzione. Proprio questa quota *embodied*, come spesso accade per lo Scope 3, finisce per attestarsi un gran numero di emissioni, pur rimanendo complessa da calcolare.

In conclusione, questo tipo di reporting è il livello più solido per misurare e ottimizzare infrastruttura e operazioni, ma da solo non ti permette di attribuire un impatto al singolo modello. I datacenter, infatti, sono spesso *multi-tenant facilities*[63] per cui associare a ciascun cliente le proprie emissioni non è né immediato né trasparente. Da questo bisogno di informazioni circa ciò che accade nei datacenter sta nascendo un ulteriore livello di reporting: quello di *workload* dei modelli.

3.3.3. Il *reporting* di *workload* del modello

Per reporting di *workload* si intendono le emissioni associate alle operazioni di allenamento e inferenza. È il tipo di reporting più giovane per cui deve essere ancora del tutto formalizzata una metodologia per il suo studio, ma metriche e standardizzazione stanno sempre di più convergendo. Proprio a livello di metriche condivise, due sono quelle di base: l'energia (KWh) misurata o stimata a seconda delle possibilità e le emissioni (in $KgCO_2eq$), ottenute moltiplicando energia e intensità carbonica. Quest'ultima si può rendicontare tramite un'analisi *location-based* oppure una *market-based*: la prima impiega il fattore medio della rete elettrica locale[64] nel luogo e nell'anno in cui avvengono i consumi, questo valore riflette la realtà fisica del mix energetico che alimenta una rete. Per quanto riguarda l'analisi *market-based* invece si utilizzano fattori di emissione basati su accordi contrattuali con cui si dichiara l'origine dell'elettricità. Strumenti come le garanzie di origine, i certificati di utilizzo di energie rinnovabili o i *power purchase agreement* permettono di applicare un fattore molto basso alla quota

[63] Una *multi-tenant facility* è una struttura di datacenter progettata per ospitare più clienti nello stesso sito, garantendo separazione e sicurezza.
[64] Nazionale o regionale

di energia coperta da questi documenti; che si traduce in una stima ottimistica delle emissioni.

Infine, va definita un'unità funzionale, ad esempio, l'energia impiegata per *training run*, quella impiegata per mille inferenze o per un milione di token processati; così da poter operare dei confronti.

Proprio la comparabilità tra i report rimane ancora fragile non tanto per le misurazioni, quanto più per la definizione del sistema: i confini delineati, ad esempio, incidono fortemente per cui se si valuta solo l'*IT power* o se si includono anche gli elementi di supporto il risultato cambia di parecchio. Un'altra differenza si può evidenziare nel metodo di calcolo dell'energia, se si misura per telemetria o se viene stimata tramite dei modelli. La criticità più forte che si riscontra nell'attività di reporting per l'IA è la difficoltà con cui si attribuiscono le emissioni all'interno dei datacenter, dove avvengono i consumi.

La gestione dei modelli, infatti, non avviene in un laboratorio, ma si articola in un sistema condiviso[65] complicando l'attribuzione perché distinguibile in attribuzione verso i carichi di lavoro dei singoli modelli e verso gli attori presenti in un datacenter.

Per quanto riguarda il *workload* le difficoltà riguardano l'inferenza, che necessita di una quota di hardware sempre in attesa, da cui derivano delle emissioni spesso non considerate nell'attività di reporting. Oltre alla quota di emissioni derivanti dall'*overhead* di un datacenter e quelle incorporate nell'hardware che vanno ammortizzate lungo il ciclo di vita e allocarle, ad esempio, per ore di utilizzo. Per quanto riguarda invece l'attribuzione tra attori, osserviamo la tabella 3.1 con quattro scenari, andando a valutare e attribuire gli scope 1, 2 e 3 a seconda dei casi: caso A *on-premises* in cui l'hardware e l'impianto datacenter è di proprietà dell'azienda, caso B

[65] Come abbiamo visto in precedenza, in un singolo datacenter le risorse sono partecipate da più realtà, con una stessa infrastruttura che gestisce più clienti e *workload*.

colocation in cui un'azienda possiede l'hardware, ma il datacenter è di una terza parte che fornisce il servizio; caso C *cloud* pubblico in cui il provider fornisce hardware e datacenter all'azienda e vende capacità di calcolo *multi-tenant*; caso D *dedicated/single tenant hosting* dove il provider fornisce hardware all'azienda tramite un'infrastruttura dedicata o poco condivisa. Nella tabella, oltre ad attribuire gli scope dell'inventario GHG per azienda e provider, sono valutate le opzioni *location-based* e *market-based*; vengono infine segnalate le criticità per l'attribuzione delle emissioni a seconda dei casi.

Tabella 3.1 - Attribuzione degli scope nei datacenter; elaborazione dell'autore.

Caso	***Setup***	**Nell' inventario GHG dell'azienda**	**Nell'inventario GHG del Provider**	***Location-based* o *Market-based***	**Criticità di attribuzione**
A. *On-premises*	Unica azienda che gestisce DC, hardware e acquista elettricità.	Scope 2: elettricità (IT + overhead via PUE). Scope 1: generatori/combustione sul sito. Scope 3: *embodied* hardware.	N/A (azienda e provider coincidono).	Nello Scope 2: mostrare sia LB che MB.	Ripartire PUE, Scope 1 sito (refrigeranti/diesel), ed *embodied* per GPU-ora / vita utile.
B. *Colocation*	Azienda che possiede l'hardware. Provider fornisce il DC (impianti). Energia: (B1) contatore dell'azienda o (B2) *all-in*	B1 (contatore azienda): Scope 2: elettricità (IT + overhead via PUE). Scope 3: *embodied* dell'azienda B2 (canone servizio): Scope 3: il DC fornito dal provider è un servizio + *embodied* hardware dell'azienda.	Provider: Scope 2: elettricità (LB/MB). Scope 1: refrigeranti e generatori del DC.	B1: LB/MB nello Scope 2 dell'azienda. B2: LB/MB del provider (l'azienda lo vede come servizio Scope 3).	In B2 mancano spesso i kWh reali. In B1 resta la problematica dell'allocazione di *overhead* e dei carichi condivisi.
C. *Cloud* pubblico	Provider possiede DC e hardware. Inoltre, vende *compute multi-tenant* ad altre aziende	Per le aziende Scope 3 (*purchased services*) per allenamento/inferenza (l'azienda compra un servizio, non kWh). L'*embodied* è "incluso" solo se il provider lo incorpora nel servizio.	Provider: Scope 2: elettricità (LB/MB). Scope 1: refrigeranti/generatori del DC. Scope 3: catena del valore dell'hardware.	LB/MB nello Scope 2 del provider. Per l'azienda Scope 3 (*capital goods*).	Complessità nel ripartire energia/*idle*/PUE, *overhead* e *workload*. Opacità dei dati.
D. *Dedicated / single-tenant hosting*	Hardware dedicato all'azienda (ma di proprietà del provider). Infrastruttura dedicata o poco condivisa.	Per l'azienda Scope 3 (*purchased services*) con attribuzione più "pulita". L'*embodied* è incluso solo se il provider lo incorpora nel servizio.	Provider: simile a C (Scope 2/1/3), ma con risorse dedicate	LB/MB nello Scope 2 del provider. Per l'azienda Scope 3 (*capital goods*).	Complessità nel calcolo dell'hardware attivo e quello in *stand by.* Oltre a ripartizione dell'*overhead*.

Oltre al problema dell'attribuzione, questo livello di reporting è ancora in fase di standardizzazione per cui soffre spesso di mancanza dati: le aziende non sempre misurano o dichiarano la telemetria energetica precisa a livello di modello, e raramente stimano con confidenza la componente *embodied* dell'hardware.

In questa direzione si sta muovendo *l'International Organization for Standardization* (ISO) con due documenti: il primo l'ISO/IEC TR 20226:2025 è un *Technical Report* che orienta e armonizza il vocabolario, ma non impone un formato unico di reporting per modello. È specifico per l'IA e organizza gli aspetti ambientali da considerare lungo il ciclo vita, oltre a proporre metriche potenziali. Il secondo invece l'ISO/IEC 21031:2024 standardizza la *Software Carbon Intensity* (SCI), una metodologia per calcolare un tasso di emissioni di un sistema software. A differenza del primo documento la SCI come metodologia non nasce per l'IA, ma viene utilizzata sempre di più per rendere confrontabili due sistemi software diversi.

Va inoltre segnalata la norma ISO/IEC 42001:2023 che introduce uno standard internazionale per l'*Artificial Intelligence Management System* (AIMS). Ovvero un sistema di gestione che consente alle organizzazioni di strutturare in modo coerente e dimostrabile la governance dell'IA. Lo standard è pensato per essere auditabile in un contesto con organismi di certificazione accreditati che valutano le pratiche di gestione dell'IA.

In conclusione, con questa suddivisone dell'attività di reporting l'obiettivo non è scegliere l'opzione migliore, ma capire che a ciascun grado di rendicontazione delle emissioni corrisponde una domanda diversa. Nel momento in cui si ha il bilancio delle tre dimensioni si ottiene una storia coerente: in cui il livello *corporate* fornisce le priorità complessive, il livello infrastrutturale provvede all'efficienza del comparto fisico dell'IA e quello di *workload* del modello permette l'analisi dell'architettura e del prodotto in termini di costo ambientale.

Al contrario se si confondono, si rischia di incappare in tre errori classici:

l'utilizzo della componente *corporate* per giustificare modelli energivori, pur considerandosi *green*; ritenere indicatori come il PUE o il WUE una prova di sostenibilità del prodotto, quando per quanto possa essere efficiente un datacenter, il *workload* complessivo è comunque determinante sulle emissioni; e infine dichiarare numeri per singola inferenza che non tengono conto dei consumi dell'hardware in *stand by*, né dell'apporto complessivo dei miliardi di inferenze che raggiungono i server.

3.4. Il *Life Cycle Assessment e* il *Greenhouse Gas Protocol Product Standard*

Chiariti i confini del *GHG Protocol* possiamo introdurre la metodologia del *Life-Cycle-Assessment* (LCA), per poi comprendere come le due fonti di rendicontazione interagiscono tra di loro. Riuscire a collegare il LCA al settore dell'IA è particolarmente importante perché significa andare in controtendenza con la convinzione che gli impatti delle tecnologie digitali si riducano a mera energia elettrica consumata.

L'obiettivo di questo tipo di studio è quello di guardare all'intero sistema socio-tecnico dell'IA, per cui si includono molte altre categorie di impatto come: l'acidificazione del suolo, l'eutrofizzazione, o la tossicità per l'ambiente. Proprio in questa caratteristica si differenzia dal *GHG Product Standard* che invece applica una logica *cradle-to-grave*[66] per quantificare la sola impronta di gas climalteranti di uno specifico prodotto o servizio. Mentre il *GHG inventory* dà informazioni circa il peso dell'azienda in termini di emissioni, il *GHG Product Standard* permette di focalizzarsi su di un unico prodotto o servizio per capire in quali fasi del suo ciclo di vita si concentrano le emissioni e dove poter intervenire.

Di fatto il *GHG Protocol Product Standard* è, concettualmente, una LCA "solo clima": con la stessa logica *cradle-to-grave* e *life cycle thinking*, ma

[66] Lungo tutto il ciclo di vita.

limitata a emissioni climalteranti espresse in CO_2e. Da qui la forte connessione: il LCA come metodologia per capire cosa succede nel sistema lungo il ciclo di vita e il *GHG Product Standard* come "modalità d'uso" della logica LCA focalizzata sul clima, pensata per rendere il dato utilizzabile come contabilità di prodotto.

Per concludere il LCA è standardizzato come ISO 14040/14044 ed è una metodologia generale, quindi non specifica per l'IA, ma si presenta come "lente" giusta per questo settore, permettendo di evitare errori negli spostamenti degli impatti fra categorie. In questo ambito gli "stadi" principali di analisi sono: l'hardware dalla produzione delle componentistiche e dei materiali fino al riciclo e lo smaltimento; l'infrastruttura dei datacenter e il carico di lavoro dei modelli tra allenamento e inferenza.

La ricerca di una visione complessiva è un segnale interessante di maturazione del settore e si inserisce nel processo di consolidazione di un vocabolario e di metriche standard per tutti gli attori. In alcuni ambiti di produzione questa opera è maggiormente sviluppata perché l'oggetto è più "fisico" e i confini sono più gestibili.

Ad esempio, l'LCA su infrastrutture come i datacenter è relativamente più implementabile, perché tecnologie come: server, chip, e raffreddamento sono elementi che l'ingegneria tradizionale sa già modellare. Allo stesso tempo per le GPU e affini, gli impatti *embodied* sono ancora troppo spesso stimati in modo grossolano a causa di dati primari di produzione proprietari, opachi e complessi. Inoltre, si stanno mettendo a punto metodologie orientate all'analisi del software per rendere comparabili misure operative e avere informazioni riguardanti gli impatti di allenamento e inferenza.

Il LCA e il *GHG Protocol* si inseriscono in una cornice più ampia per la regolamentazione sostenibile, che applicata al settore IA porta con sé misurazioni condivise e comparabili, trasparenza e documentazione verificabile oltre a mantenere un focus nella rendicontazione della catena del valore.

3.5. La regolazione e la *governance* dell'Intelligenza Artificiale

La regolamentazione dell'IA è in una fase molto riconoscibile, un'era caratterizzata da frammentazione, ma in cui si rileva una generale convergenza di principi. In pratica, quasi tutti i Paesi stanno andando nella stessa direzione per ottenere un maggiore trasparenza, una chiara *accountability* e la tutela dei diritti, ma non sono ancora allineati su come raggiungere questi obiettivi. A livello globale si stanno delimitando i confini della regolamentazione tramite strumenti di governance *soft-law*, questa opera si articola tramite due pilastri: il G7 *Code of Conduct* (G7, 30 ottobre 2023) e la *Framework Convention on Artificial Intelligence* del Consiglio d'Europa (5 settembre 2024).

Il G7 *Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI System*s viene pensato per colmare il divario tra i principi etici e gli obblighi giuridici pienamente vincolanti in una fase di rapida evoluzione tecnologica. Questo codice di condotta viene adottato come strumento volontario ed è dichiaratamente aggiornabile in quanto *living document*[67]. È rivolto alle organizzazioni che sviluppano sistemi di IA avanzati e si articola in un set di azioni operative lungo l'intero ciclo di vita, dallo sviluppo fino al monitoraggio dopo la distribuzione. In particolare, viene proposto un approccio sistemico basato sulla valutazione e sulla mitigazione dei rischi, orientato alla trasparenza, nonché alla cooperazione e condivisione responsabile di informazioni. Seppure privo di meccanismi sanzionatori diretti, in quanto *soft-law*, il Codice ha rilevanza pratica perché riferimento comune per le pratiche di IA responsabile. Proprio per questo motivo l'obiettivo non è la sanzione dei comportamenti sbagliati, ma l'istituzione di una cornice in cui gli attori coinvolti, come *stakeholders* e

[67] Sono documenti che non vengono scritti una volta e poi archiviati, ma sono pensati per essere aggiornati continuamente man mano che cambiano informazioni, decisioni, requisiti o contesto.

utenti, possano prendere decisioni informate.

In questo modo si va ad accrescere l'importanza reputazionale e di responsabilità dei fornitori di sistemi IA, che grazie a iniziative di reporting e monitoraggio promosse in sede OCSE, sono invitati a migliorare la propria trasparenza, anticipando o integrando così i futuri regimi normativi.

La *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law* del Consiglio d'Europa costituisce il primo trattato internazionale legalmente vincolante dedicato specificamente all'IA. L'obiettivo è quello di garantire che le attività svolte lungo l'intero ciclo di vita dei sistemi computazionali siano coerenti con i diritti umani, la democrazia e lo stato di diritto, mantenendo al contempo uno spazio per l'innovazione, il progresso tecnologico e la sostenibilità. Anche in questo caso il testo adotta un approccio di valutazione dei rischi, in cui si impone alle parti di adottare misure legislative, amministrative o di altra natura in funzione della probabilità e della severità dei possibili impatti negativi.

Sono inoltre previsti dei requisiti di governance e *accountability*, nonché garanzie di trasparenza e supervisione. È infatti richiesta la notifica dell'implementazione di soluzioni IA in modo tale che gli utenti finali siano consci dell'interazione con un modello. In questo senso sono fondamentali dei meccanismi di controllo circa l'attribuzione della responsabilità riguardo le decisioni "prese" da una macchina, oltre alla richiesta di rimedi accessibili ed efficaci quando l'IA porta a violazioni dei diritti.

Quanto all'ambito, la Convenzione si applica in modo diretto alle autorità pubbliche e ai soggetti privati che operano per loro conto; viene invece lasciato ai singoli Stati un margine circa le modalità di copertura del settore privato. È prevista anche una clausola per cui ogni Stato deve valutare se servano moratorie o divieti per i possibili usi ritenuti incompatibili con la tutela dei diritti, lo stato di democrazia e la *rule of law*.

A livello settoriale sono state operate delle esclusioni riguardanti: il settore della difesa nazionale che rimane fuori dall'ambito del trattato; il settore della

sicurezza nazionale in cui lo Stato non è obbligato all'applicazione, fintanto che si rispetti il diritto internazionale; e il settore di ricerca e sviluppo, anche in questo caso, finché non si intaccano i diritti fondamentali.

Sul piano istituzionale, l'attuazione è accompagnata da meccanismi di *follow-up* come la *Conference of the Parties*, per supportare il monitoraggio e l'evoluzione interpretativa del quadro convenzionale. Ogni parte deve istituire o designare meccanismi di supervisione indipendenti per vigilare sulla conformità, cooperando nell'arena internazionale tramite la condivisone di informazioni.

Il trattato non ambisce a creare un'autorità di *enforcement*, ma vuole coadiuvare le istituzioni nazionali, nell'attuazione domestica, tramite l'attività di coordinamento e armonizzazione. All'interno di questi documenti si trovano disposizioni, suggerimenti e regolamenti che vedono gli sforzi maggiori nei confronti della tutela dei diritti. Questa propensione per un'IA etica e trasparente riflette gli studi delle prime ondate di ricerca che giustamente si sono focalizzate sugli argomenti più impellenti per realizzare un'IA sicura.

Per quanto riguarda la sostenibilità è sempre presente come dimensione, ma raramente è affrontata in modo programmatico o procedurale. Nonostante l'assenza di pratiche specifiche per questo campo, gli impatti ambientali fanno comunque parte delle valutazioni di rischio e pertanto vanno affrontati. Con la definizione degli standard citati nel capitolo precedente e lo sviluppo di una regolamentazione ad hoc per la sostenibilità dell'IA, sarà possibile comprendere al meglio gli effetti diretti e indiretti che questa tecnologia comporta.

In conclusione, la scelta di regolamentare l'IA tramite norme progettate per evolvere nel tempo non è casuale, ma riflette la sfida nel regolare un fenomeno pervasivo in rapido mutamento. Ciò che si può fare ad oggi è intensificare gli sforzi intervenendo laddove si può: come nei settori di

produzione delle componenti materiali dell'IA, nella gestione dell'infrastruttura e nell'utilizzo di fonti rinnovabili di energia.

3.5.1. Unione Europea, regolazione e sostenibilità

L'Unione Europea vuole affermarsi come regolatore "forte" del settore IA utilizzando un approccio normativo che si articola lungo tre direzioni: la valutazione del rischio, la proporzionalità e l'armonizzazione nel mercato unico. Queste tre direttrici si fondono nell'AI Act (Parlamento Europeo e Consiglio dell'Unione Europea, 2024), una *policy* orizzontale su cui si costruisce l'impianto di regolamentazione dell'IA.

In prima istanza le istituzioni europee mirano ad inserire dei confini tecnici e legali affinché si possano prevedere e gestire eventuali danni riconducibili all'IA. La valutazione del rischio prevede quattro categorie: le pratiche vietate con rischi inaccettabili, i sistemi ad alto rischio[68] con requisiti pesanti, i casi a rischio limitato con minori obblighi di trasparenza e la restante casistica con rischi pressoché nulli.

Soprattutto per i sistemi ad alto rischio, l'UE usa una mentalità simile alle regole di sicurezza dei prodotti nel Mercato Unico come: requisiti, documentazione e controlli per assicurare la *compliance*[69] e la conformità alle norme. A differenza di un prodotto tradizionale l'IA ha bisogno di aggiornamenti e re-training, subisce variazioni dovute ai dataset e può cambiare performance nel tempo; per queste motivazioni l'AI act inserisce una serie di controlli che non si limitano all'immissione del sistema IA in commercio, ma che ne valutano la continua evoluzione. In questo modo i provider dei servizi IA devono mettere a punto e documentare un'impalcatura di monitoraggio post-market, così da garantire la conformità lungo tutto il

[68] Sistemi che possono incidere su diritti o sicurezza; che possono fallire in modi non ovvi e che possono essere usati in contesti dove l'errore è particolarmente gravoso.
[69] L'insieme di misure organizzative, procedure e controlli con cui un soggetto (azienda, ente, professionista) assicura e dimostra di operare in conformità a norme e obblighi applicabili.

ciclo vita.

Da questa suddivisione emerge poi la seconda direttrice: la proporzionalità dei doveri alla categoria di rischio. L'AI Act non mira a soffocare l'innovazione "innocua" con oneri pensati per tecnologie ad alto rischio, ma impiega degli obblighi commisurati al livello di pericolo che un'applicazione IA può comportare.

Come ultimo obiettivo abbiamo l'armonizzazione nel mercato interno che si affida ad un regolamento diretto e uniforme e ad autorità nazionali che lo applicano. Grazie alla grandezza del mercato europeo molte aziende scelgono di costruire prodotti e processi con standard a livello UE, per poi applicarli ovunque, perché più semplice rispetto ad avere più versioni diverse. Le aziende che si adeguano "a monte" finiscono così per esportare lo stesso livello di *compliance* anche al di fuori dell'Unione portando un'ulteriore armonizzazione indiretta.

Dal punto di vista della governance l'*European AI Office* (Parlamento Europeo e Consiglio dell'Unione Europea, 2024) coordina l'implementazione e l'applicazione della normativa; ha anche poteri "da regolatore" con cui può operare valutazioni, chiedere informazioni e misure correttive, oltre a supervisionare i modelli di IA di uso generale, cioè quelli più potenti e trasversali. Sono anche previsti tre organi di indirizzo e consulenza che guidano e bilanciano la governance: il primo è l'*European Artificial Intelligence Board* composto da rappresentanti degli Stati membri, coordina e rende coerente l'applicazione tra paesi; il secondo il *Scientific Panel* composto da esperti indipendenti a garanzia del rigore tecnico; e come terzo l'*Advisory Forum* con una selezione di *stakeholders* sia commerciali che non.

L'applicazione viene gestita a livello nazionale e prevede la designazione di autorità competenti per vigilare e garantire il rispetto degli obblighi; oltre agli organismi indipendenti che fanno la valutazione di conformità prima dell'immissione sul mercato.

Mentre per quanto riguarda la protezione dei diritti, della democrazia e dello stato di diritto l'AI Act mira a fornire delle garanzie solide e adatte ad ogni possibile scenario, non si può dire lo stesso per la sostenibilità.

L'articolo sui codici di condotta infatti afferma che l'*AI Office* e gli Stati membri devono incoraggiare e facilitare pratiche che promuovano l'applicazione volontaria di requisiti e buone pratiche, tra cui viene elencata esplicitamente la valutazione e la riduzione dell'impatto ambientale. Per quanto previsto dal primo esempio di *hard-law* per l'IA a livello internazionale, l'impatto ambientale viene inserito in una cornice più *soft*: all'interno di codici volontari. Così che sia possibile far maturare prassi condivise, con obiettivi e indicatori, in ciò che nel futuro sarà lo standard del settore. È allo stesso tempo un modo per tutelare il mercato facendo emergere delle metriche misurabili senza incorporarle in un tessuto normativo precocemente.

Il testo prevede infatti che entro il 2 agosto 2028, e poi ogni 4 anni, la Commissione pubblichi un report sullo stato di avanzamento della standardizzazione per lo sviluppo *energy-efficient* dei modelli IA e al contempo, valuti se servano ulteriori misure potenzialmente vincolanti (Parlamento Europeo e Consiglio dell'Unione Europea, 2024).

Oltre all'AI act l'UE prevede delle misure non "IA-specifiche", ma che si possono comunque rivelare decisive per la sostenibilità del settore come l'*Energy Efficiency Directive* (Parlamento europeo e Consiglio dell'Unione europea, 2023) riguardante anche i datacenter. L'ottica alla base è quella per cui, se l'IA moderna è *compute-hungry*, regolarne l'infrastruttura rimane un metodo diretto per agire sull'impronta ambientale.

La direttiva per l'efficienza energetica prevede obblighi di monitoraggio e reporting sulla performance energetica dei datacenter, includendo dati utili anche sulla water footprint ed inserendoli in un database europeo.

Infine, la *EU Taxonomy* (Parlamento europeo e Consiglio dell'Unione europea, 2020) costituisce, in primo luogo, una classificazione comune delle

attività economiche considerate "ambientalmente sostenibili": stabilisce infatti dei criteri tecnici[70] e delle condizioni generali[71] per qualificare un'attività come sostenibile.

In questo senso, la *EU Taxonomy* diventa una leva rilevante per il settore dell'IA perché la sua adozione su larga scala dipende dall'espansione dell'infrastruttura. Gli effetti pratici si manifestano lungo tre canali principali: la finanza, poiché investitori e intermediari possono usare la *Taxonomy* come linguaggio comune per valutare il grado di "sostenibilità" degli investimenti[72]; la strategia industriale, perché per dichiarare un'attività *taxonomy-aligned* diventa necessario progettare e gestire il capitale fisico come i data center in coerenza con i criteri tecnici, incentivando scelte orientate a efficienza energetica, gestione dell'acqua e recupero/riuso del calore; la *disclosure*, in quanto i soggetti tenuti alla rendicontazione devono comunicare la quota di attività *taxonomy-eligible* e *taxonomy-aligned*, con un conseguente incentivo alla misurazione e alla comparabilità delle performance ambientali.

In sintesi, la *EU Taxonomy* non impone direttamente soglie di consumo o di emissioni per singoli modelli di IA, ma contribuisce a costruire un sistema di incentivi economico-informativi che rende più conveniente e rendicontabile lo sviluppo dell'infrastruttura su cui l'IA opera in modo coerente con gli obiettivi ambientali.

3.5.2. Stati Uniti, regolazione e sostenibilità

Negli USA l'approccio alla regolamentazione dell'IA non va a ricercare un'unica legge federale omnicomprensiva come in UE, ma si articola secondo un mosaico di enti statali e agenzie, con tanta *soft law*, guidata

[70] Indicati come *technical screening criteria.*

[71] Tra cui il contributo sostanziale ad almeno un obiettivo ambientale e il principio di *"do no significant harm".*

[72] Di fatto influenzando così il costo e l'accesso al capitale per infrastrutture digitali più o meno efficienti.

dall'applicazione di leggi già esistenti. Lo "spirito normativo" statunitense si caratterizza per un pervasivo *laissez-faire* che insiste sull'autoregolamentazione industriale e una limitata interferenza governativa. Proprio in questo senso, le principali norme di politica pubblica per l'IA sono volte alla libertà imprenditoriale, valutata maggiormente rispetto ad obiettivi di tutela degli utenti, con un approccio fortemente pro-business (Donoghue et al, 2024).

La distribuzione dei poteri tra Congresso, Casa Bianca, agenzie federali[73] e Stati federati ha prodotto un panorama normativo estremamente complesso e disperso. In tale contesto, gli *Executive Order* hanno storicamente svolto una funzione di coordinamento interno del ramo esecutivo, orientando le priorità, gli standard e i compiti amministrativi, ma la loro efficacia "di diritto vigente" dipende dalla continuità politica e può essere ridimensionata o annullata da revoche e riorientamenti successivi.

Parallelamente, la parte più rilevante della regolazione "effettiva" deriva dall'applicazione delle agenzie federali che applicano cornici normative preesistenti ai rischi e agli impatti dei sistemi automatizzati a prescindere dalla presenza o meno di una legge IA specifica.

Infine, gli Stati e i livelli locali restano una fonte normativa molto attiva, con discipline spesso orientate alla valutazione dei rischi e improntate su settori specifici; ma che a causa di questa attività alimentano al contempo sia innovazione regolatoria sia frammentazione. Gli standard tecnici, pur rimanendo formalmente volontari, diventano di fatto obbligatori tramite il *procurement* pubblico[74]: le aziende si assicurano di essere conformi per partecipare agli appalti e richiedono lo stesso livello di *compliance* anche ai partner privati.

La sostenibilità non rientra chiaramente come campo di studio obbligatorio

[73] Come, ad esempio, la *Federal Trade Commission*, o il *Department of Justice*, o *il Consumer Financial Protection Bureau*.

[74] Il procurement pubblico è l'insieme delle attività con cui una pubblica amministrazione (acquista beni, servizi o lavori usando denaro pubblico.

per tutti i modelli, bensì si inserisce come somma di leve indirette: standard tecnici, regole ambientali sull'infrastruttura, e, sempre di più, leggi statali mirate.

In questo modo non si crea *hard law*, ma si impone una base di mercato per cui se non si misura, non ci si districa tra audit, richieste dei clienti e gare di appalto. Anche in questo caso il reporting sostenibile viene sorretto dalla finanza: anche senza un obbligo specifico per l'IA, se un'azienda ha costi di *compute* importanti, le emissioni e i rischi di approvvigionamento energetico diventano materiale estremamente importante per investitori e reporting. Nel contesto statunitense, l'impostazione orientata a una forte competizione dovrebbe favorire il miglioramento delle pratiche di reporting aziendale. Tuttavia, la frammentazione normativa solleva interrogativi sull'efficacia di tale armonizzazione, soprattutto quando viene affidata all' autoregolamentazione del mercato.

La regolamentazione sostenibile a livello statale sta diventando sempre più strettamente collegata alle normative relative all'infrastruttura di calcolo. Questioni come il consumo di energia elettrica, l'uso delle risorse idriche e l'efficienza energetica hanno accelerato il processo normativo, portando gli stati federati a impegnarsi affinché i costi non gravino sui cittadini o sulle risorse ambientali locali.

Un esempio di questa dinamica lo si trova in California dove l'approccio *study first* richiede che il legislatore, invece di imporre subito obblighi ai datacenter, ordini prima uno studio ufficiale per capire quanto e come i datacenter possano incidere sulla rete elettrica, per poi legiferare più attivamente una volta chiariti gli impatti. Anche se esistono preoccupazioni riguardo all'ambiente locale, la crescita delle infrastrutture sembra avere la priorità rispetto agli obiettivi climatici.

Questa tendenza è evidente nel ridimensionamento di vari disegni di legge e nel veto del Governatore contro una richiesta di maggiore trasparenza nell'uso dell'acqua (Ufficio del Governatore della California, 11 ottobre

2025). Negli Stati Uniti l'attenzione tende a concentrarsi sulla "corsa" all'IA: l'obiettivo politico-economico è mantenere la leadership tecnologica, facendo leva sugli ingenti flussi di capitale che il settore è in grado di attrarre e su un'impostazione che guarda con sospetto a un'eccessiva regolamentazione, percepita come potenziale freno a investimenti, innovazione e crescita (Donoghue et al, 2024).

In questo quadro, la sostenibilità raramente compare come asse esplicito della disciplina dell'IA: più che un principio dichiarato, è un tema che riaffiora indirettamente dove l'IA tocca infrastrutture e costi reali. Così, pur restando spesso "innominata" nei testi e nelle strategie, la sostenibilità è presente come una variabile di fatto: non ancora un pilastro normativo, ma una pressione crescente che, come una pianta tra le crepe del cemento, si fa spazio attraverso standard tecnici, condizioni contrattuali, scelte di procurement e interventi statali mirati.

3.5.3. Cina, regolazione e Sostenibilità

La Cina sta tentando ambiziosamente di diventare leader mondiale nella regolamentazione dell'IA, con una serie di provvedimenti normativi nell'arco dell'ultimo decennio. Questa nuova era normativa in Cina è iniziata con l'*Artificial Intelligence Development Plan*, pubblicato nel luglio 2017 (Consiglio di Stato cinese, 2017) che annunciava l'intenzione radicale del governo di istituire un sistema di supervisione e valutazione della sicurezza in questo campo. Due anni dopo, un comitato di esperti istituito dal Ministero della Scienza e della Tecnologia cinese ha pubblicato un documento che delinea otto principi per la governance dell'intelligenza artificiale e per un'IA responsabile.

Ad oggi, la maggior parte delle leggi e delle politiche riguardanti la governance dei dati e dell'intelligenza artificiale ha adottato un approccio omnicomprensivo lungo tutti i settori. Poiché l'Assemblea Nazionale del Popolo non ha ancora concluso la legislazione specifica per l'IA, una serie di

normative per la regolamentazione dell'uso dei dati acquisiscono maggiore importanza: come la *Data Security Law* (Comitato permanente dell'Assemblea nazionale del popolo, 2021) che ha come obiettivi la sicurezza dei dati e degli interessi di sovranità e sviluppo che ne conseguono, o la *Personal Information Protection Law* (PIPL), che fa da solida base giuridica per la regolamentazione dell'IA. La PIPL è una normativa specifica sulla protezione delle informazioni personali, viene spesso descritta come una versione cinese del *General Data Protection Regulation*[75] (Parlamento europeo e Consiglio dell'Unione europea, 2016) dell'UE, date le somiglianze tra i due.

Le attività di raccolta ed elaborazione di informazioni personali, indipendentemente dalle diverse finalità, rientrano in questa normativa. La PIPL stabilisce un quadro di base per la protezione delle informazioni personali in tutti i settori industriali, ma si esprime riguardo l'elaborazione di informazioni personali quando vengono implementati sistemi IA.

In particolare, quando ne viene regolata l'applicazione tramite sistemi decisionali automatici, sono richieste: trasparenza, giustificazione ed equità cosicché ciascun individuo abbia diritto a chiarimenti sul processo di decisione, oltre che tutele nei confronti di eventuali danni.

L'impostazione cinese della governance dell'IA si caratterizza per una marcata centralità dell'algoritmo come oggetto di regolazione (Sheehan, 2023). Tale scelta si riflette sia nel discorso politico che normativo con l'adozione del registro degli algoritmi, per cui gli sviluppatori di un modello IA sono tenuti a fornire informazioni sulle modalità di addestramento e distribuzione, compresi i dataset su cui l'algoritmo viene addestrato, oltre alla compilazione di un rapporto di autovalutazione della sicurezza dello stesso. In questa cornice si assume l'algoritmo come unità base per obblighi di trasparenza e disclosure.

[75] Il GDPR (Regolamento (UE) 2016/679) è la legge quadro dell'Unione europea sulla protezione dei dati personali.

La natura del registro e delle informazioni richieste suggeriscono, una concezione secondo cui una regolazione efficace presuppone la conoscenza tecnica dettagliata dei singoli algoritmi e, se necessario, la possibilità di intervenire su di essi. Questo meccanismo standardizzato di *disclosure:* aiuta l'amministrazione pubblica nell'integrazione di nuovi provvedimenti, modificandone nel tempo requisiti e livello di dettaglio.

Il registro crea un'infrastruttura procedurale che riduce i costi istituzionali necessari a introdurre e far rispettare misure più ambiziose in futuro. A prescindere dal valore normativo, le informazioni richieste per accreditare un algoritmo fanno anche da base per la successiva attività di vigilanza.

Infine, la strategia regolatoria cinese in materia di algoritmi e IA si caratterizza per un'impostazione verticale e iterativa. Da un lato, Pechino interviene su specifiche applicazioni della tecnologia, attraverso norme settoriali che rispondono a preoccupazioni mirate. Dall'altro lato, l'azione normativa procede per aggiustamenti successivi: quando una misura risulta incompleta o inefficace, viene sostituita o integrata con nuove disposizioni che ne correggono i limiti o ne estendono il perimetro. Un chiaro esempio di questa dinamica è riscontrabile nel passaggio dalle misure riguardanti la *deep synthesis*[76] alla bozza sull'IA generativa che ha portato ad una richiesta di dati maggiormente accurati, senza valutare però la complessità di analisi per i provider. Questo metodo può rendere più complessa la *compliance* per le imprese, ma i regolatori lo considerano un compromesso accettabile per governare un settore in rapida evoluzione.

Nel contesto normativo cinese, la sostenibilità si aggancia all'IA soprattutto a monte, cioè sull'infrastruttura che la rende possibile ovvero i datacenter. La Cina ha fissato degli obiettivi nazionali espliciti per rendere green i propri datacenter, puntando su metriche misurabili e requisiti progettuali: in particolare, il piano per lo sviluppo verde dei datacenter indica come

[76] L'azienda tecnologica Tencent è riuscita a introdurre e diffondere il termine *deep synthesis*, sostituendo il termine *deepfake* con un termine tecnico dal suono più innocuo.

traguardo un PUE medio nazionale inferiore a 1,5 entro il 2025 e un aumento del tasso di utilizzo di energia rinnovabile di circa il 10% annuo (Consiglio di Stato cinese, 24 luglio 2024).

Inoltre, il governo centrale sta spingendo per l'adozione di soluzioni tecniche e gestionali che migliorino l'efficienza energetica, con indicazioni su nuovi progetti e interventi di adeguamento di quelli esistenti. Questo è rilevante per l'IA perché la maggior parte dell'impatto ambientale dei modelli è in funzione diretta del consumo di energia e del raffreddamento dei data center. Regolando dove e come si costruisce la capacità di calcolo Pechino agisce indirettamente, ma in modo concreto sulla sostenibilità dell'ecosistema IA, anche senza inserire capitoli specifici nelle norme IA strettamente intese. Infine, la sostenibilità entra nel discorso normativo cinese come cornice di *governance*: non tanto come clausola tecnica, ma più come pilastro politico-regolatorio che giustifica standard, cooperazione e infrastrutture più verdi. Già nella *Global AI Governance Initiative* (Ministero degli Affari Esteri Cinese, 20 ottobre 2023) Pechino collegava esplicitamente l'IA agli obiettivi di sviluppo sostenibile (ESG) e alla conseguente risposta a sfide globali come il clima e la biodiversità. L'IA viene quindi presentata come tecnologia da orientare al bene comune e a fini ambientali. Questo framing diventa più operativo nel *Global AI Governance Action Plan* (Ministero degli Affari Esteri Cinese, 26 luglio 2025), che dedica un passaggio specifico alle questioni energetiche e ambientali: promuovendo l'idea di IA sostenibile. L'invito è di sviluppare modelli di IA attenti alle risorse e all'ambiente, e propone standard congiunti di efficienza energetica e idrica, insieme alla promozione di tecnologie di *green computing* come chip a basso consumo e soluzioni software più efficienti.

3.6. Conclusioni

L'IA si configura come un settore fortemente *capital-intensive*, segnato da asimmetrie informative e da una marcata concentrazione di dati e capacità di

calcolo. Per questo motivo, la regolazione e la sostenibilità non vanno lette come semplici "correzioni tecniche", bensì come problemi socio-istituzionali legati alla distribuzione del potere, alla trasparenza e alla responsabilità lungo la filiera.

La letteratura tecnica tende a focalizzarsi sui consumi operativi, ma l'analisi svolta mostra la necessità di includere anche gli impatti indiretti e l'intera catena del valore, così da evitare che la riduzione di un indicatore locale produca uno spostamento degli impatti in un'altra sezione di analisi.

In questo quadro, i tre livelli di reporting risultano complementari e non sostituibili: il reporting *corporate* offre una visione macro inclusiva dello Scope 3, ma per quest'ultimo dipende ancora troppo significativamente da stime e assunzioni; il reporting infrastrutturale dei datacenter si fonda su misure robuste e fisiche, pur senza attribuire automaticamente gli impatti al singolo modello; il reporting di *workload* del modello è il più vicino a *training* e inferenza, ma resta il più giovane e fragile per confini, allocazioni e limiti di telemetria.

Sul piano metodologico, il *GHG Protocol Product Standard* fa da panoramica circa le emissioni climalteranti, mentre la LCA in senso pieno amplia le categorie d'impatto e riduce il rischio di spostamento degli oneri ambientali; allo stesso tempo, strumenti come standard ISO, tehnical report, SCI e sistemi di gestione come l'AIMS (ISO/IEC 42001) non costituiscono "la soluzione" in sé, ma rappresentano infrastrutture di comparabilità e auditabilità che rendono praticabile una governance verificabile.

Infine, la comparazione tra modelli regolatori evidenzia traiettorie differenti ma convergenti nel riconoscere il peso dell'infrastruttura: l'UE adotta una logica sostenibile ancora in parte affidata a strumenti di *soft-law* ma sostenuta da leve più dure via infrastruttura e finanza; gli USA mostrano un approccio frammentato, dove procurement, standard e leve indirette fanno emergere la sostenibilità soprattutto quando incide su costi e capacità; la Cina privilegia l'algoritmo come unità regolatoria attraverso registri e attività di

archiviazione, con un'impostazione verticale e un'attenzione strategica al tema infrastrutturale.

Tramite questa lettura, la sostenibilità dell'IA non riguarda soltanto un problema di efficienza, ma bensì di potere informativo, di infrastruttura e di allocazione della responsabilità; senza metriche condivise e possibilità di verifica indipendente, il rischio è che la sostenibilità diventi una qualità semplicemente dichiarata, più che effettivamente dimostrabile.

CONCLUSIONI

Questa tesi ha affrontato l'intelligenza artificiale come un fenomeno che non può essere compreso, né governato, restando confinati alla dimensione algoritmica. Il percorso sviluppato nei tre capitoli ha messo in relazione: la geografia e le dinamiche dell'innovazione nel settore IA; la sua materialità ambientale lungo l'intero ciclo di vita; le condizioni istituzionali e politico-economiche che rendono possibile una regolazione efficace, evitando che trasparenza e sostenibilità restino meri enunciati.

In altri termini, la tesi argomenta che l'IA contemporanea è un'infrastruttura socio-tecnica e industriale: un insieme di modelli, dati, hardware e datacenter, inserito in un ecosistema competitivo che condiziona in modo decisivo ciò che è misurabile, rendicontabile e quindi regolabile.

Il primo capitolo ha mostrato che la crescita dell'IA non è un trend lineare e neutrale, bensì un processo di espansione rapida con un chiaro spostamento del baricentro geografico e con differenze significative tra ricerca accademica e ricerca privata.

Sul versante scientifico, la produzione globale è più che raddoppiata nel decennio 2013–2023, da 102.000 a 242.000 pubblicazioni, con un picco di crescita nel 2023 del +19,7%. Sul versante privato, gli indicatori economici e imprenditoriali confermano che l'IA è diventata un settore ad alta intensità

di capitale, con un divario crescente tra Stati Uniti e resto del mondo: nel 2024 gli USA hanno attratto 109,1 miliardi di dollari di investimenti privati in IA, oltre dieci volte la Cina.

Questo dato non è solo descrittivo: anticipa un elemento decisivo per la governance, perché la capacità di investire in capacità di calcolo computazionale, infrastrutture e *compliance* determina anche chi può definire standard, pratiche e aspettative di mercato.

Il secondo capitolo ha spostato la prospettiva dalla "promessa" dell'IA alla sua impronta materiale. L'analisi lungo il ciclo di vita evidenzia che i costi ambientali dell'IA non si riducono al consumo elettrico del *training*, ma attraversano più ambiti interdipendenti: l'estrazione delle materie prime, la produzione hardware e infrastrutture, i consumi di energia e di acqua, con *trade-off* che rendono insufficiente qualsiasi lettura monodimensionale.

Il capitolo insiste sulla necessità di misurazioni e indicatori robusti: la sostenibilità non può essere valutata solo attraverso dichiarazioni o metriche parziali. La tesi del capitolo è che gli interventi di efficienza, se isolati, possono produrre spostamenti di impatto tra componenti, da acqua diretta a indiretta, o da impatti operativi a impatti incorporati, richiedendo una visione integrata tra tecnologia, territorio e filiera.

Il terzo capitolo riunisce i precedenti in una prospettiva di governance: la regolazione dell'IA si colloca in un contesto di forte asimmetria informativa e di concentrazione di risorse, dove il rischio non è soltanto "regolare poco", ma anche "regolare male". Da un lato, emerge la necessità di dispositivi che trasformino i principi di: trasparenza, *accountability* e tutela dei diritti in meccanismi operativi verificabili; dall'altro, si evidenzia che gli strumenti regolativi possono diventare nuove barriere all'ingresso se non progettati con attenzione agli incentivi e alla struttura del mercato.

In questo quadro, la rendicontazione ambientale e la misurazione degli impatti non sono un capitolo separato dalla regolazione, ma una sua condizione abilitante: ciò che non si misura in modo comparabile

difficilmente può essere governato in modo credibile. Ne deriva una tesi conclusiva: la sostenibilità dell'IA è inseparabile dalla costruzione di condizioni di auditabilità, cioè dalla possibilità di produrre dati affidabili, renderli confrontabili e sottoporli a verifica indipendente. In assenza di tale infrastruttura conoscitiva, la regolazione rischia di oscillare tra due esiti sub-ottimali: o resta troppo generica per incidere sui *trade-off* materiali, oppure si traduce in obblighi formali che rafforzano gli *incumbents* senza migliorare realmente gli impatti.

La geografia politico-istituzionale della regolazione assume rilievo, perché mostra che la governance dell'IA non si sviluppa secondo un modello unico, ma attraverso traiettorie differenti, ognuna legata a specifici equilibri tra innovazione, mercato, sicurezza e controllo.

L'Unione europea cerca di "costituzionalizzare" l'IA tramite un impianto normativo generale e basato sul rischio, volto a renderla compatibile con i diritti fondamentali, la trasparenza e l'accountability. Gli Stati Uniti, invece, ne orientano lo sviluppo soprattutto in modo indiretto: attraverso procurement pubblico, infrastrutture strategiche e difesa, privilegiando l'innovazione con un approccio più favorevole al posto di una regolazione generale ex ante. La Cina, infine, si colloca in una posizione distinta da entrambe: non punta su una singola legge quadro, ma su una regolazione amministrativa selettiva e particolarmente stringente. In questo modo il controllo statale permette di intervenire direttamente sullo sviluppo tecnologico specialmente in materia di contenuti, algoritmi, dati e sicurezza.

A partire da quanto emerso nei capitoli, le implicazioni principali possono essere formulate come risultati di indirizzo.

1- La misurazione non è un accessorio: è la precondizione della regolazione.

   Il secondo capitolo evidenzia che senza misurazione accurata non si possono identificare le aree di intervento più efficaci; l'infrastruttura

resta un fattore determinante per la sostenibilità, e che la sola ottimizzazione dei modelli, pur promettente, dipende anche dalla realtà materiale.

2- Serve un approccio multilivello e coerente per evitare "illusioni metriche".

   Il WUE è emblematico: come metrica "sul sito" aiuta a migliorare l'efficienza interna, ma può nascondere consumi indiretti, come l'acqua associata alla generazione elettrica o l'impronta idrica incorporata nella filiera hardware. Analogamente, PUE e WUE, se trattati come meri obiettivi, rischiano di spostare l'attenzione dall'impatto reale all'ottimizzazione di un indicatore. Questa dinamica è ben descritta dal paradosso di Jevons e, più in generale, dagli effetti di *rebound*: quando un miglioramento di efficienza riduce il costo marginale di una risorsa, in questo caso calcolo, energia e raffreddamento; la domanda può aumentare a tal punto da compensare o addirittura superare i risparmi iniziali. Applicato all'IA, ciò implica che modelli più efficienti possono abbassare la barriera economica e tecnica all'adozione, favorendo una crescita del numero di applicazioni, utenti e richieste. Di conseguenza, l'attenzione esclusiva a metriche di efficienza relativa come PUE, WUE o energia per *query* rischia di non tradursi in una riduzione assoluta degli impatti se non è accompagnata da strumenti capaci di governare la scala d'uso: obiettivi in termini assoluti, meccanismi di allocazione e attribuzione per *workload* e, dove necessario, vincoli o incentivi che orientino la domanda verso usi ad alto valore sociale.

3- Standard e sistemi di gestione sono leve di auditabilità, non mere formalità.

   Il terzo capitolo sottolinea l'emergere di standard ISO che, pur non risolvendo da soli i problemi, contribuiscono a costruire un linguaggio comune e metodologie confrontabili: dal *Technical Report* sugli

aspetti ambientali dell'IA (ISO/IEC TR 20226:2025), alla *Software Carbon Intensity* (ISO/IEC 21031:2024), fino allo standard per l'*AI Management System* auditabile (ISO/IEC 42001:2023).

Considerati nel loro insieme, i tre capitoli convergono su un punto: l'IA non è soltanto una tecnologia "potente", ma un settore infrastrutturale che combina rapida innovazione, concentrazione di risorse e impatti materiali crescenti. La sostenibilità, in questo scenario, non può essere delegata a miglioramenti incrementali di efficienza o a dichiarazioni di principio: richiede metriche, perimetri e verifiche capaci di collegare l'astrazione dei modelli alla concretezza di energia, acqua, hardware e territorio.

Di conseguenza, l'efficacias della regolazione dipende dalla costruzione di un'infrastruttura di auditabilità con standard, telemetria e *accountability*, che trasformi trasparenza e responsabilità da obiettivi normativi a pratiche verificabili, riducendo la distanza tra ciò che l'IA promette e ciò che, materialmente, costa al mondo che la ospita. In questo scenario, il paradosso di Jevons suggerisce cautela: l'aumento di efficienza può accelerare l'adozione e, senza politiche che governino la scala, tradursi in un incremento netto dei consumi e della pressione su energia e acqua.

In sintesi, l'intelligenza artificiale applicata alla sostenibilità rappresenta un valido strumento per il perseguimento degli obiettivi delineati nell'ambito degli SDG. Le sue applicazioni spaziano dal settore energetico, favorendo l'accesso a fonti di energia pulita (SDG 7), al supporto delle azioni per il clima (SDG 13), fino alla promozione di infrastrutture resilienti (SDG 9) e al sostegno di modelli di consumo e produzione responsabili (SDG 12). Riprendendo la definizione di sostenibilità del Rapporto Brundtland, per uno sviluppo che "soddisfa i bisogni del presente senza compromettere la possibilità delle generazioni future di soddisfare i propri" (WCED; 1987), l'IA va governata come tecnologia abilitante, ma anche come potenziale moltiplicatore di pressioni materiali su energia, acqua e filiere. Valutare la

“sostenibilità dell’IA” è particolarmente cruciale in una fase in cui la transizione verde è già sotto stress a causa di tensioni geopolitiche, della frammentazione delle catene di fornitura e della priorità che la sicurezza ha acquisito nel panorama globale. Il rischio è di ridurre lo spazio politico per obiettivi climatici ambiziosi, svalutandone la centralità nell’agenda internazionale. In questo contesto, regolamentare l’IA significa evitare che l’espansione del calcolo diventi un ulteriore fattore di attrito sistemico, orientando crescita e innovazione entro metriche verificabili, incentivi coerenti e limiti compatibili con la traiettoria della decarbonizzazione.

BIBLIOGRAFIA

Apple (2024). Sustainability Report.

Aula V., e Erkkilä T. (2024). *AI and transparency*. In R. Paul, E. Carmel, & J. Cobbe (Eds.), Handbook on Public Policy and Artificial Intelligence (pp. 170–180). Edward Elgar.

Barroso L. A., Dean J. (2013). The tail at scale. *Communications of the ACM*, 56(2), 74–80. https://doi.org/10.1145/2408776.2408794.

Brookings Metropolitan Policy Program. (2021). *The geography of AI: which cities will drive the artificial intelligence revolution?.* Washington D.C.

Chen S., et al. (2025). Energy Consumption Modeling and Elastic Space Computation of Data Centers Considering Spatiotemporal Transfer Flexibility. *Energies*, *18*(24).

Commissione Europea. (2025). *European Innovation Scoreboard 2025 Annex B. Bruxelles*.

Dang J. e Motohashi K. (2015). Patent statistics: A good indicator for innovation in China? Patent subsidy program impacts on patent quality. *China Economic Review*, Volume 35, pp. 137-155. https://doi.org/10.1016/j.chieco.2015.03.012.

De Vries-Gao A. (2025). The carbon and water footprints of data centers and what this could mean for artificial intelligence. *Patterns*, 7(1).

Ding L., et al (2025). Rise of Generative Artificial Intelligence in Science. *Scientometrics* (2025). https://doi.org/10.1007/s11192-025-05413-z.

Donoghue R. (2024). *AI, regulation, and the world of work: The competing approaches of the US and China*. In R. Paul, E. Carmel, & J. Cobbe (Eds.), Handbook on public policy and artificial intelligence. Edward Elgar Publishing.

Draghi M. (2024). *The future of European competitiveness*. Luxembourg.

Fernandez J., Na C., Tiwari V., Bisk Y., Luccioni S., & Strubell, E. (2025). Energy considerations of large language model inference and efficiency optimizations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. ACL Anthology.

Garst, J., et al. (2022). Materiality assessment is an art, not a science: Selecting ESG topics for sustainability reports. *California Management Review*, 65(1), 64–90. https://doi.org/10.1177/00081256221120692.

Gholami A., Kim S., Dong Z., Yao Z., Mahoney M. W., Keutzer K. (2021). A survey of quantization methods for efficient neural network inference. *IEEE Signal Processing Magazine, 38*(3), 16–29.

Gillis R., Laux J., & Mittelstadt B. (2024). Trust and trustworthiness in artificial intelligence. In R. Paul, E. Carmel, & J. Cobbe (Eds.), *Handbook*

on public policy and artificial intelligence (pp. 181–193). Edward Elgar Publishing.

Google (2022), Environmental Report 2021.

Gupta U., et al. (2020). Chasing carbon: The elusive environmental footprint of computing. *IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, Seoul, South Korea.

Henderson P. et al. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning, *Journal of Machine Learning Research*, 21.

IEA (2021), *Data Centres and Data Transmission Networks*. France.

IEA (2024). *Electricity 2024 Analysis and forecast to 2026.* France.

IEA (2025). *Energy and AI.* France.

Jang S. e Morabito R. (2025). Edge-first language model inference: Models, metrics, and tradeoffs. In *Proceedings of the 45th IEEE International Conference on Distributed Computing Systems* (ICDCS 2025). IEEE. (In press). In *Proceedings of the 2025 IEEE 45th International Conference on Distributed Computing Systems Workshops (ICDCSW 2025)*. IEEE.

Jiang P., Sonne C., Li W., You F., You S. (2024). Preventing the immense increase in the life-cycle energy and carbon footprints of LLM-powered intelligent chatbots. *Engineering*, 40, 202–210.

Jobin A., Ienca M., & Vayena E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

JRC Technical Report (2022). *AI Watch: Estimating AI Investments in the European Union*. Luxembourg.

Kirkpatrick J., Pascanu R., Rabinowitz N., Veness J., Desjardins G., Rusu A.A., Milan K., Quan J., Ramalho T., Grabska-Barwinska A., Hassabis D., Clopath C., Kumaran D., Hadsell R. (2017) Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America* 114 (13) 3521-3526.

Laskaridis S., Kouris A., Lane N. D. (2021). Adaptive Inference through Early-Exit Networks: Design, Challenges and Directions. In *Proceedings of the 5th International Workshop on Embedded and Mobile Deep Learning*. 1–6. New York, USA, Association for Computing Machinery, https://doi.org/10.1145/3469116.3470012.

Lee H., Phatale S., Mansoor H., Mesnard T., Ferret J., Bishop C. Hall E., Carbune V., Rastogi, A. Prakash S. (2024). RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235:26874-26901.

Li P., et al. (2025). Making AI Less 'Thirsty' Uncovering and addressing the secret water footprint of AI models. *Communications of the ACM,* 68(7), 54-61. https://doi.org/10.1145/3724499.

Ligozat A.-L.; Lefevre J.; Bugeau A.; Combaz J. (2022). Unraveling the Hidden Environmental Impacts of AI Solutions for Environment Life Cycle Assessment of AI Solutions. *Sustainability 2022*, 14, 121-134. https://doi.org/10.3390/su14073791.

Luccioni A. S., Jernite Y., Strubell E. (2024). Power hungry processing: Watts driving the cost of AI deployment?. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, Rio de Janeiro, Brasile. https://doi.org/10.1145/3630106.3658542

Lucivero F. (2020) Big Data, Big Waste? A Reflection on the Environmental Sustainability of Big Data Initiatives. *Sci Eng Ethics* **26**, 1009–1030. https://doi.org/10.1007/s11948-019-00171-7.

Lucivero F. (2024). Artificial intelligence and public values. In R. Paul, E. Carmel, & J. Cobbe (Eds.), *Handbook on public policy and artificial intelligence*. Edward Elgar Publishing.

Lu T., et al. (2025). Simple and Effective Dynamic Provisioning for Power-Proportional Data Centers. *IEEE Transactions on Parallel and Distributed Systems, 24(*6), 1161–1171. https://doi.org/10.1109/TPDS.2012.241.

Mahadevan A., Mathioudakis M. (2024). Cost-aware retraining for machine learning. *Knowledge-Based Systems, 293.* https://doi.org/10.1016/j.knosys.2024.111610.

Malmodin J. and Lundén D. (2018). The Energy and Carbon Footprint of the Global ICT and E&M Sectors 2010–2015. *Sustainability*, *10*(9).

Meta (2021). 2020 Sustainability Report.

Microsoft. (2025). *2025 Environmental sustainability report*.

Mytton D. (2021). Data centre water consumption. *npj Clean Water* **4**(49). https://doi.org/10.1038/s41545-021-00101-w.

Networking and Information Technology Research and Development Program (NITRD), & National Artificial Intelligence Initiative Office. (2024). *Supplement to the President's FY2025 budget*. National Science and Technology Council.

Parlamento europeo e Consiglio dell'Unione europea. (2024). *Regolamento (UE) 2024/1689 che stabilisce norme armonizzate sull'intelligenza artificiale (Artificial Intelligence Act)*. Gazzetta ufficiale dell'Unione europea. Regulation - EU - 2024/1689 - IT - EUR-Lex.

Patterson D., et al. (2022). The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. *Computer,* 55(7), 18-28. https://doi.org/10.1109/MC.2022.3148714.

Poenaru-Olaru L., Sallou J., Cruz L., Rellermeyer J. S., van Deursen A. (2023). Retrain AI Systems Responsibly! Use Sustainable Concept Drift Adaptation Techniques.2023 *IEEE/ACM 7th International Workshop on Green And Sustainable Software (GREENS)*, pp. 17-18, Melbourne, Australia.

Sheehan H. (2023). *Artificial intelligence and ideology*. In R. Paul, E. Carmel, & J. Cobbe (Eds.), Handbook on public policy and artificial intelligence. Edward Elgar Publishing.

Stanford University HAI. (2025). *Artificial Intelligence Index Report 2025.* Santa Clara.

Strubell E., Ganesh A., McCallum A. (2019). Energy and Policy Considerations for Deep Learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

TEHA Group. (5 Settembre 2025). *L'Italia dei data center: Energia, efficienza, sostenibilità per la transizione digitale* [Position paper]. A2A.

Tmamna J., Ayed E. B., Fourati R. et al. (2024) Pruning Deep Neural Networks for Green Energy-Efficient Models: A Survey. *Cognitive Computation* **16**, 2931–2952. https://doi.org/10.1007/s12559-024-10313-0

UNITAR. (2025). *The Global E-waste Monitor*. Ginevra/Bonn.

Van Wynsberghe A. (2021) Sustainable AI: AI *for* sustainability and the sustainability *of* AI. *AI Ethics* **1**, 213–218. https://doi.org/10.1007/s43681-021-00043-6.

Verdecchia R., Cruz L., Sallou J., Lin M., Wickenden J., Hotellier E. (2022). Data-centric green AI: An exploratory empirical study. In *Proceedings of the International Conference on Information and Communication Technology for Sustainability (ICT4S 2022)*. 13(4).

World Commission on Environment and Development. (1987). *Our common future*. Oxford University Press.

World Resources Forum (2025). *Responsible Mining Index Report 2025*. St. Gallen, Svizzera.

Wu C. et al. (2022). Sustainable AI: Environmental Implications, Challenges and Opportunities. *Proceedings of Machine Learning and Systems 4.* doi*:10.48550/arXiv.2111.00364*.

Yamashita I. et al. (2021). Measuring the AI content of government-funded R&D projects: A proof of concept for the OECD Fundstat initiative. *OECD Science, Technology and Industry Working Papers*, No. 2021/09, OECD Publishing, Paris, https://doi.org/10.1787/7b43b038-en.

Sitografia:

Future of Life Institute. (22 marzo 2023). *Pause Giant AI Experiments: An Open Letter*. Pause Giant AI Experiments: An Open Letter - Future of Life Institute.

Comitato permanente dell'Assemblea nazionale del popolo. (20 agosto 2021). *Personal Information Protection Law of the People's Republic of China.* Personal Information Protection Law of the People's Republic of China

Comitato permanente dell'Assemblea nazionale del popolo. (6 ottobre 2021). *Data Security Law of the People's Republic of China*. Data Security Law of the PRC

Commissione Europea. (21 Aprile 2021). Fostering a European approach to artificial intelligence. *eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX%3A52021DC0205&utm.*

Commissione Europea. (n.d.). The Digital Europe Programme. *The Digital Europe Programme | Shaping Europe's digital future.*

Commissione Europea. (n.d.). Euopean Digital Innovation Hubs Network. Home | European Digital Innovation Hubs Network.

Congresso degli Stati Uniti d'America. *National AI Initiative Act*. 2020. H.R.6216 - 116th Congress (2019-2020): National Artificial Intelligence Initiative Act of 2020 | Congress.gov | Library of Congress.

Consiglio d'Europa. (5 settembre 2024). *Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law*. The Framework Convention on Artificial Intelligence - Artificial Intelligence

Consiglio di Stato cinese. (2017). *New generation artificial intelligence development plan*. https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm

Consiglio di stato cinese (24 luglio 2024). China sets green targets for data centers. China sets green targets for data centers

Consiglio di Stato cinese. (7 marzo 2025). China to boost policy mix to ensure sustained growth in 2025. *China to boost policy mix to ensure sustained growth in 2025.*

Crunchbase News (20 agosto 2025). As Funding To AI Startups Increases And Concentrates, Which Investors Have Led?. As Funding To AI Startups Increases And Concentrates, Which Investors Have Led?.

Crunchbase News (3 aprile 2025). Q1 Global Startup Funding Posts Strongest Quarter Since Q2 2022 With A Third Going To Massive OpenAI Deal**.** Q1 Global Startup Funding Posts Strongest Quarter Since Q2 2022 With A Third Going To Massive OpenAI Deal **.**

G7 Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems. (30 ottobre 2025). G7G20 documents database. Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems (30/10/2023) - G7/G20 Documents Database

IEA (23 ottobre 2025). With new export controls on critical minerals, supply concentration risks become reality. With new export controls on critical minerals, supply concentration risks become reality – Analysis - IEA.

Ministero degli Affari Esteri cinese (20 ottobre 2023). Global AI Governance Initiative. Global AI Governance Initiative_Ministry of Foreign Affairs of the People's Republic of China

Ministero degli Affari Esteri cinese (26 luglio 2025). Global AI Governance Action Plan. Global AI Governance Action Plan_Ministry of Foreign Affairs of the People's Republic of China

Ministero dell'Industria e della Tecnologia cinese. (2025). National AI Industry Investment. Fund. 工业和信息化部党组书记、部长李乐成：加快推进人工智能赋能新型工业化.

Municipality of Hangzhou (5 giugno 2025). Hangzhou unveils 20 new AI policies to build national innovation hub.

Parlamento europeo & Consiglio dell'Unione europea. (2023). *Direttiva (UE) 2023/1791 sull'efficienza energetica*. Gazzetta ufficiale dell'Unione Europea. https://eur-lex.europa.eu/eli/dir/2023/1791/oj.

Parlamento europeo e Consiglio dell'Unione europea. (27 aprile 2016). *Regolamento (UE) 2016/679, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (Regolamento generale sulla protezione dei dati).* Gazzetta ufficiale dell'Unione Europea

Parlamento europeo e Consiglio dell'Unione europea. (2020). *Regolamento (UE) 2020/852 sull'istituzione di un quadro per facilitare gli investimenti sostenibili (EU Taxonomy Regulation)*. Gazzetta ufficiale dell'Unione Europea. https://eur-lex.europa.eu/eli/reg/2020/852/oj.

Parlamento europeo e Consiglio dell'Unione europea. (2024). *Regolamento (UE) 2024/1689 che stabilisce norme armonizzate sull'intelligenza artificiale (Artificial Intelligence Act)*. Gazzetta ufficiale dell'Unione Europea. Regulation - EU - 2024/1689 - IT - EUR-Lex.

Reuters (21 maggio 2024). Solaris ends plan to sell stake to China's Zijin Mining, cites government rules. Solaris ends plan to sell stake to China's Zijin Mining, cites government rules | Reuters.

Reuters, (13 giugno 2025). Meta poaches 28-year-old scale AI CEO after taking multibillion dollar strike in startup. Meta poaches 28-year-old Scale AI CEO after taking multibillion dollar stake in startup | Reuters.

Ufficio del governatore della California (30 serttembre 2023). Assembly Bill 1631 veto message. *Assembly Bill 1631*

Yicai Global (28 febbraio 2025). Beijing Sets Up USD13.7 Billion Fund to Invest in AI, Robotics Startups, Official Says. Beijing Sets Up USD13.7 Billion Fund to Invest in AI, Robotics Startups, Official Says