29. gennaio

Il tallone di Achille della AI: la criticità energetica

Renderò l'elettricità così economica
che solo i ricchi si potranno permettere il lusso di utilizzare le candele.
Thomas A. Edison

Sam Altman, il mega-milionario che gestisce OpenAI in un intervento di 30 minuti a *AI World Economic Forum di Davos,* pochi giorni fa, ha detto molte cose interessanti sul *futuro dell' intelligenza artificiale.* in particolare quando ha annunciato che AI per poter funzionare deve attuare una *"svolta copernicana"* per quanto riguarda la **criticità energetica** che costituisce un inquietante "***tallone di Achille"***

Non è un segreto che l'industria dell'intelligenza artificiale richieda ***una quantità di energia mozzafiato*** . In effetti, l'impatto ecologico di questo business è sufficiente a spingere gli ambientalisti più seri a scappare a gambe levate.

Negli ultimi mesi, le aziende tecnologiche sono state alla ricerca di fonti di energia alternative che siano più rispettose dell'ambiente e, soprattutto, forniscano quantità sempre maggiori di energia per soddisfare le modalità di risucchio dei dati dell'intelligenza artificiale. Di per sè il dato è agghiacciante: entro il *2025,* il consumo di energia dei datacenter, utilizzerà non meno del 10% dell'utilizzo mondiale di elettricità.

Attualmente durante il *tuning,* la fase di apprendimento degli algoritmi un qualsiasi sistema di AI brucia oltre 284 tonnellate di anidride carbonica, un valore che tende a raddoppiare ogni 3-4 mesi. Attualmente *i* data center sono *energivori* e producono calore che deve essere neutralizzato grazie a massivi quantitativi di acqua

La quantità di acqua necessaria per addestrare **ChatGPT-3** alla diagnosi di COVID è la stessa quantità necessaria per produrre **370 auto BMW elettriche** e **320 Tesla.ChatGPT-3** ha bisogno di bere una bottiglia d'acqua di 500 ml per una semplice conversazione di circa 20/50 domande e risposte convenzionali

Questa estate a causa della siccità l'università del Texas di Arlington ha razionata l'acqua per il raffreddamento dei DATA center

L'intelligenza artificiale si candida ad essere la prossima grande minaccia al cambiamento climatico.

Lo *stesso Altman* sta'investendo centinaia di milioni di dollari in progetti di fusione energetica , tra cui Helion Energy , una startup di fusione che ora è di proprietà di Microsoft, uno speciale partner commerciale di OpenAI . Molti esperti sostengono che l'energia nucleare, se gestita correttamente, potrebbe essere una buona cosa per l'America. Detto questo, se il miglior utilizzo di quell'energia nucleare sarebbe quello di creare chatbot e generatori di contenuti più veloci è una questione completamente diversa.

Tra i numerosi dibattiti sui potenziali pericoli dell'intelligenza artificiale, alcuni ricercatori sostengono che si sta trascurando un'importante preoccupazione: ***l'energia*** utilizzata dai computer per addestrare ed eseguire grandi modelli di intelligenza artificiale.

Alex de Vries della *VU Amsterdam School of Business and Economics* avverte che la crescita dell'intelligenza artificiale è pronta a renderla un contributore significativo alle *emissioni globali di carbonio*. Secondo le sue stime, se Google trasferisse l'intera attività di ricerca all'intelligenza artificiale, finirebbe per utilizzare ***29,3 terawattora*** all'anno, equivalenti al consumo di elettricità dell'Irlanda, e quasi il doppio del consumo energetico totale dell'azienda di ***15.4 terawattora*** nel 2020.

Google non lo ha fatto.

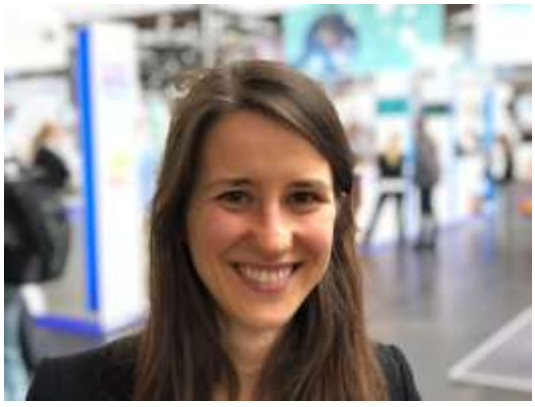
Da un lato ci sono buone ragioni per non farsi prendere dal panico. Realizzare questo tipo di cambiamento è praticamente impossibile, poiché richiederebbe più di ***4 milioni di potenti chip*** per computer noti ***come unità di elaborazione grafica (GPU)*** che sono attualmente molto richiesti con una offerta limitata. Ciò costerebbe ***100 miliardi di dollari***, che anche le profonde tasche di Google avrebbero difficoltà a finanziare.

Nella corsa all'oro dell'intelligenza artificiale, i produttori di chip scarseggiano. Poiché l'ultima generazione di modelli di intelligenza artificiale come ChatGPT sembra destinata a trasformare le nostre vite, l'hardware che rende tutto ciò possibile sta diventando una risorsa strategica, con paesi, aziende e ricercatori che lottano per procurarsi le forniture. Mentre si parla di carenze che dureranno almeno fino al prossimo anno, alcuni aspiranti sviluppatori di intelligenza artificiale rischiano di essere lasciati indietro.Al centro di tutto questo c'è ***Nvidia****, che si è unita alla piccola cricca di aziende da trilioni di dollariall'inizio di quest'anno, affiancandosi ad aziende del calibro di Apple e Google, proprietaria di Alphabet.* ***Nvidia****, con sede in California, si è fatta un nome creando hardware per eseguire giochi per computer, ma queste unità di* ***elaborazione grafica (GPU)*** *si sono rivelate estremamente capaci di elaborare le grandi quantità di dati necessari per addestrare un'intelligenza artificiale. I rapporti suggeriscono che* ***Nvidia*** *ora controlla l'80% del mercato mondiale delle GPU .*

D’altra parte, col tempo, il consumo energetico dell’intelligenza artificiale rappresenterà un vero problema. ***Nvidia*****,** che vende il ***95% delle GPU*** utilizzate per l’intelligenza artificiale, quest’anno spedirà 100.000 dei suoi server A100, che possono consumare complessivamente ***5,7 terrawattora*** all’anno.

Le cose potrebbero, e probabilmente peggioreranno, col tempo, quando nuovi impianti di produzione entreranno in funzione e aumenteranno notevolmente la capacità produttiva. Il produttore di *chip TSMC*, che fornisce ***Nvidia***, sta investendo in nuove fabbriche che potrebbero fornire **1,5 milioni** di server all’anno entro il 2027, e tutto quell’hardware potrebbe consumare ***85,4 terawattora*** di energia all’anno..

Con le aziende che si affrettano a integrare l’intelligenza artificiale in tutti i tipi di prodotti, ***Nvidia*** probabilmente non avrà problemi a smaltire le proprie scorte***. Ma de Vries afferma*** *come sia importante che l’intelligenza artificiale venga utilizzata con parsimonia, dato il suo elevato costo ambientale."Le persone hanno questo nuovo strumento e pensano, 'OK, è fantastico, lo useremo', senza preoccuparsi se ne hanno effettivamente bisogno", afferma. “Si dimenticano di chiedere o di chiedersi se l’utente finale ne abbia bisogno in qualche modo o se questo migliorerà la sua vita. E penso che la disconnessione sia, in definitiva, il vero problema”.*



Sandra Wachter *dell’Università di Oxford* afferma che i consumatori dovrebbero essere consapevoli che giocare con questi modelli ha un costo. ***"È uno degli argomenti che mi tiene sveglia la notte",*** afferma Wachter. *“****Interagiamo semplicemente con la tecnologia e non siamo effettivamente consapevoli di quante risorse – elettricità, acqua, spazio – siano necessarie”. Una legislazione volta a imporre la trasparenza sull’impatto ambientale dei modelli spingerebbe le aziende ad agire in modo più responsabile, afferma.***

Un portavoce di **OpenAI**, lo sviluppatore di **ChatGPT**, *ha dichiarato “Riconosciamo che l’addestramento di modelli di grandi dimensioni può essere ad alta intensità energetica ed è uno dei motivi per cui lavoriamo costantemente per migliorare l’efficienza. Riflettiamo attentamente su come utilizzare al meglio la nostra potenza di calcolo”.*

Thomas Wolf co-fondatore della società di intelligenza artificiale **Hugging Face**

Hugging Face

ricorda che ci sono segnali che i *modelli di intelligenza artificiale più piccoli* si stanno ora avvicinando alle capacità di quelli più grandi, il che potrebbe portare a notevoli risparmi energetici,

Mistral 7B e e Meta's Llama 2 sono da 10 a 100 volte più piccoli di GPT4, l'intelligenza artificiale dietro ChatGPT, e possono fare molte delle stesse cose, dice. "*Non tutti hanno bisogno della GPT4 per tutto, così come non serve una Ferrari per andare al lavoro."*

Un portavoce di ***Nvidia*** afferma che l'esecuzione dell'intelligenza artificiale sulle sue GPU è più efficiente dal punto di vista energetico rispetto a un tipo alternativo di chip chiamato CPU. *"Il calcolo accelerato sulla tecnologia Nvidia è il modello di elaborazione più efficiente dal punto di vista energetico per l'intelligenza artificiale e altri carichi di lavoro dei data center", affermano. "I nostri prodotti sono più performanti ed efficienti dal punto di vista energetico con ogni nuova generazione."*

La maggior parte delle IA viene eseguita su server realizzati da Nvidia, che sono pieni di chip GPU assetati di energia

Poiché la domanda globale di servizi Internet e di intelligenza artificiale continua a crescere, i data center che li supportano potrebbero raddoppiare il consumo di elettricità in soli due anni.
Ciò significa che nel *2026* divoreranno tanta energia quanta ne consuma oggi l'intera nazione del Giappone, con una popolazione di 125 milioni di abitanti.

Un rapporto **dell'Agenzia internazionale per l'energia** , che stima che il consumo globale di elettricità dei data center, compresi i servizi di intelligenza artificiale e le criptovalute, potrebbe aumentare da **460 terawattora nel 2022** a oltre ***1.000 terawattora*** entro il 2026. quantità di energia equivalente alla produzione di 1 trilione di watt per 1 ora.
Sarà "*fondamentale moderare l'impennata del consumo energetico dei data center*", scrive l'agenzia nel rapporto.
Per fare ciò, l'IEA raccomanda di stabilire nuove normative e apportare miglioramenti tecnologici per aumentare l'efficienza energetica dei data center. Misure di mitigazione come queste potrebbero determinare se, entro il 2026, i data center finiranno per consumare solo ***620 terawattora o fino a 10***

I **data center** ad alto consumo energetico potrebbero funzionare altrettanto bene con meno raffreddamento
Permettere ai data center di funzionare a temperature più elevate potrebbe ridurre la quantità di energia utilizzata per raffreddarli del **56%,** senza incidere sulle prestazioni computazionali.
I server di fascia alta generano molto calore durante il funzionamento, quindi le apparecchiature all'interno dei data center vengono normalmente mantenute a temperature più basse, in genere facendo scorrere acqua fredda in tutti i locali per raffreddare l'aria calda.
Normalmente gli operatori mirano a mantenere i data center a una temperatura compresa tra ***20°C e 25°C (68°F e 77°F),*** per prevenire il surriscaldamento e potenziali danni alle apparecchiature.
Questo raffreddamento ha un costo, che rappresenta un terzo del consumo energetico totale di un tipico data center.
L'aumento dei servizi di *streaming video* e di *intelligenza artificiale* ,, che richiedono entrambi enormi data center, ha portato maggiore attenzione a questo utilizzo di energia, con sforzi in corso per rendere i sistemi di raffreddamento più efficienti.

Shengwei Wang ed il su team del *politecnico Universitario di di Hong Kong* suggeriscono un approccio alternativo: basta far funzionare i server a temperature più elevate.

I ricercatori hanno sviluppato una simulazione al computer che ha modellato le prestazioni delle apparecchiature IT a diverse temperature in *57 città in tutto il mondo*, scelte per essere rappresentative del quadro globale.

Hanno scoperto che i ***data center*** potrebbero risparmiare tra il **13 e il 56%** dell'energia utilizzata per il raffreddamento funzionando a **41°C (106°F)** rispetto a quelli funzionanti a **22°C (72°F),** senza un significativo degrado delle prestazioni del server.

I luoghi con una temperatura ambiente più elevata, come il Brasile o l'Africa occidentale, normalmente richiedono il massimo raffreddamento e quindi consentirebbero il maggiore risparmio energetico.

Markkula Center
for Applied Ethics
at Santa Clara University

Irina Raicu *dell'Università di Santa Clara,* punto di riferimento per gli approfondimenti etici dell'intelligenza artificiale ritiene che ***La ricerca incentrata sulla riduzione del consumo energetico dei data center è assolutamente necessaria, così come lo è la copertura mediatica dell'impatto ambientale dei data center, ma a i miglioramenti tecnici spesso comportano dei compromessi e, come riconoscono gli autori di questo articolo, richiedono un attento studio preliminare all'implementazione".***

Il mondo attualmente fa affidamento su circa **8.000 data center**, di cui gli Stati Uniti detengono circa il **33%** della quota totale, seguiti dal ***16% in Europa*** e dal ***10% in Cina***.

Si prevede che i data center in tutte queste regioni registreranno una crescita significativa.

Negli Stati Uniti, ad esempio, si prevede che oltre un terzo dell'aumento della domanda di elettricità da qui al 2026 deriverà dall'espansione dei data center. Ciò significa che questi impianti passerebbero dal consumo del *4% di tutta l'elettricità negli Stati Uniti nel 2022 al 6% entro il 2026.*

L'Unione Europea può aspettarsi che il consumo di elettricità dei data center aumenti da poco meno di ***100 terawattora nel 2022 a quasi 150 terawattora entro il 2026.***

L'Irlanda da sola potrebbe vedere un aumento del consumo di elettricità dei data center dal **17%** del totale nazionale nel 2022 – già equivalente all'elettricità consumata da tutti gli edifici residenziali urbani al **32%** del totale nazionale entro il 2026.

Nel frattempo, il settore dei ***data center cinese*** potrebbe consumare ***circa 300 terawattora entro il 2026, rispetto ai circa 200 terawattora attuali.***

La "buona notizia" è che l'AIE prevede che l'aumento complessivo della domanda di elettricità a livello mondiale sarà coperto da un'impennata record nella produzione di elettricità da fonti a basse emissioni, compresa **l'energia nucleare** e le fonti rinnovabili **come solare eolica ed idroelettrica**

Ma soddisfare la crescente domanda di elettricità potrebbe essere più difficile per le singole regioni e comunità.

Viatico:
La Quarta Illuminazione rivela che gli esseri umani sono da sempre *a corto di energia* e hanno quindi cercato di controllarsi a vicenda per impossessarsi di quella che scorre fra le persone. La Quinta ci rivela che esiste una fonte alternativa, ma noi non potremo mantenere il contatto con essa finché non saremo consapevoli del modo in cui cerchiamo di controllare gli altri e smetteremo di farlo perché ogni volta che ricadiamo in questa abitudine il collegamento si interrompe.
(James Redfield . La profezia di Celestino)

SCIENCE WEEK

Dal 14 al 28 gennaio

Ma che succede al Dana Farber Cancer Institstitute ?

Decine di documenti del direttore del *Dana-Farber Cancer Institute* e di tre ricercatori senior del DFCI devono essere ritirati o corretti, ha rivelato l'istituto la settimana scorsa. L'annuncio è arrivato a seguito di un post sul blog del 2 gennaio dell'investigatore freelance di dati Sholto David in cui si affermava ***la manipolazione di dati e immagini in 57 articol***i su aspetti fondamentali della biologia del cancro scritti tra il 1997 e il 2017 dal presidente e CEO di DFCI *Laurie Glimcher, Chief Operating Officer William Hahn, Senior Vice La presidente Irene Ghobrial e il direttore del centro Kenneth Anderson.* La DFCI ha richiesto la ***ritrattazione di sei articoli e la correzione di altri 31***, per i quali gli autori "hanno la responsabilità primaria per i potenziali errori nei dati", afferma Barrett Rollins, responsabile dell'integrità della ricerca dell'istituto. Secondo quanto riferito, DFCI ha iniziato le sue indagini un anno fa e Rollins afferma che ***sta ancora indagando su altri 16 documenti*** contenenti dati provenienti da altri laboratori DFCI segnalati da David.

Ruolo di ZEB2 nella patogenesi dell' auto immunità

Le cellule B associate all'età (ABC) sono un sottoinsieme distinto di cellule B che si accumulano con l'avanzare dell'età e durante alcune infezioni croniche. Gli ABC contribuiscono anche alla patogenesi di alcune malattie autoimmuni come il lupus eritematoso sistemico e la sclerosi multipla. Il team di *Dai Dai del Shanghai Institute of Rheumatology, Shanghai Renji Hospital*, riportano che il ***fattore di trascrizione ZEB2*** è fondamentale nel guidare la formazione dell'ABC e la patogenicità sia nei topi che nell'uomo. ZEB2 promuove la firma genetica, il fenotipo (ad esempio, l'espressione di CD11c) e la funzione (ad esempio, la capacità fagocitaria) degli ABC. Inoltre, ZEB2 reprime MEF2B, un fattore di trascrizione che istruisce lo sviluppo del centro germinale, che di conseguenza indirizza gli ABC verso una risposta extrafollicolare. La regolazione degli ABC da parte di ZEB2 richiede anche la segnalazione JAK-STAT, il che suggerisce che prendere di mira questo percorso può ridurre gli ABC nelle malattie autoimmuni.

Dai D et alThe transcription factor ZEB2 drives the formation of age-associated B cells. Science. 2024 Jan 26;383(6681):413-421.

Organizzazione della vulnerabilità

I neuroni sensoriali periferici, come il sistema nervoso centrale, sono protetti dagli agenti patogeni da barriere anatomiche e cellule immunitarie. Il team di *Harald Lund del Department of Physiology and Pharmacology, Center for Molecular Medicine, Karolinska Institutet* ha analizzato utilizzando imaging combinato con analisi trascrizionali di singole cellule per caratterizzare le cellule endoteliali vascolari e i macrofagi associati ai gangli della radice dorsale (DRG) nei topi. I vasi sanguigni del DRG mostravano una zonazione molecolare, strutturale e funzionale lungo l'asse artero-venoso. Macrofagi che esprimono il recettore scavenger CD163 localizzato specificamente nelle regioni del sistema vascolare DRG con la più alta permeabilità del sangue. Questi macrofagi fagocitavano molecole circolanti nel sangue e si attivavano in risposta all'infiammazione sistemica indotta dal lipopolisaccaride. Una popolazione simile di macrofagi è stata identificata nei tessuti umani.

Lund H et al. CD163+ macrophages monitor enhanced permeability at the blood-dorsal root ganglion barrier. J Exp Med. 2024 Feb 5;221(2):e20230675.

La proteina MerTK nel repair miocardico

Un infarto miocardico, o "attacco cardiaco", si verifica quando il flusso sanguigno, e quindi l'ossigeno, verso una parte del cuore viene bloccato, causando ischemia. Un trattamento efficace consiste nel ripristinare il flusso sanguigno con vari mezzi, ma la riperfusione può stimolare l'infiltrazione immunitaria e il danno tissutale. Tuttavia, non tutte queste cellule immunitarie sono dannose: i macrofagi che esprimono una proteina chiamata MerTK rimuovono le cellule morenti e altri detriti, aiutando il tessuto a guarire. Yihui Shao del Beijing Anzhen Hospital of Capital Medical University Beijing ha

identificato un fattore di trascrizione chiamato ATF3 che aiuta a proteggere il cuore prevenendo la perdita di macrofagi MerTK + . Hanno anche scoperto un possibile modo per stimolare questo percorso e potenzialmente migliorare la riparazione cardiaca in ambito clinico.

Shao, Y et al ATF3 coordinates the survival and proliferation of cardiac macrophages and protects against ischemia–reperfusion injury. Nat Cardiovasc Res 3, 28–45 (2024). https://doi.org/10.1038/s44161-023-00392-x