Computer scientist Yann LeCun: “Intelligence really is about learning”



Another driver to leave was that his work with world models and AMI was also proving to have potential uses that were not interesting to Meta, such as jet engines and heavy industry. And LeCun had no trouble finding investors who were willing to bet on the next generation of AI technologies.

In his next chapter, LeCun believes that setting up a “neolab,” meaning a start-up that does fundamental research, is the new, most fertile ground. He cites OpenAI former chief technology officer Mira Murati’s Thinking Machines (“I hope the investors know what they do”) and OpenAI’s co-founder and chief scientist Ilya Sutskever’s Safe Superintelligence (“There I know the investors have no idea what they do”) as good examples.

His new architecture uses videos to give AI models an understanding of the physics of our world, which will allow

them to make better predictions of what will happen next. The model also relies on “emotions,” meaning past experiences and evaluations, to guide its predictions.

“If I pinch you, you’re going to feel pain. But then your mental model of me is going to be affected by the fact that I just pinched you. And the next time I approach my arm to yours, you’re going to recoil. That’s your prediction, and the emotion it evokes is fear or avoidance of pain,” he says.

LeCun says we will see “baby” versions of this within 12 months, and on a larger scale within a few years. It’s not quite yet superintelligence, but a path toward it. “Maybe there is an obstacle we’re not seeing yet, but at least there is hope.”

After three and a half hours, we are now the only customers left in the restaurant. I ask him what he wants his legacy to be.

Increasing the amount of intelligence in the world, he replies, without batting an eyelid. “Intelligence is really the thing that we should have more of,” he says, adding that with more intelligence, there’s less human suffering, more rational decisions, and more understanding of the world and the universe.

“We suffer from stupidity.”

Melissa Heikkilä is the FT’s AI correspondent.

© 2026 The Financial Times Ltd. All rights reserved. Not to be redistributed, copied, or modified in any way.

https://arstechnica.com/ai/2026/01/computer-scientist-yann-lecun-intelligence-really-is-about-learning/

L’AI sulle auto a guida autonoma come “Supercar”, ma l’Europa è pronta?

Nvidia porta il “ragionamento” nelle auto a guida autonoma: al CES 2026 nasce Alpamayo

Al CES di Las Vegas, il CEO di Nvidia, **Jensen Huang**, ha annunciato qualcosa di grosso: *«Il momento ChatGPT per l’AI fisica è arrivato»*. Il riferimento è ad **Alpamayo**, la nuova piattaforma di intelligenza artificiale (**AI**) per la **guida autonoma** presentata da Nvidia, che punta a colmare uno dei principali limiti dei sistemi attuali: la gestione degli scenari rari e imprevedibili.

Tutti ci ricordiamo della mitica “Supercar” guidata da David Hasselhoff, che parlava con il computer di bordo “Kitt”, che in fondo era un’AI fisica. Fantascienza pura in una serie TV storica degli anni ’80 del secolo scorso, sempre più realtà in questo 2026 alle prime battute.

[embedded content]
Non più solo “vedere” e reagire, ma **capire, ragionare e spiegare le decisioni**. È questa l’ambizione dichiarata dal CEO del colosso dei chip, oggi la società quotata più capitalizzata al mondo (intorno ai 4,5 trilioni di dollari), che con Alpamayo segna un’evoluzione strategica: da fornitore di potenza di calcolo a **piattaforma di riferimento per l’AI fisica e robotica**.

[embedded content]

Che cos’è Alpamayo e perché è diversa

I sistemi di guida autonoma tradizionali separano la percezione (sensori e visione artificiale) dalla pianificazione delle azioni. Un approccio efficace nella maggior parte dei casi, ma fragile nel cosiddetto *long tail*: eventi improvvisi, rari, ambigui o mai visti prima, che rappresentano la principale fonte di rischio.

Secondo quanto spiegato da Nvidia stessa, Alpamayo introduce un modello di **AI “reasoning-based”**, capace di descrivere la scena che sta “osservando”, quindi con capacità visiva avanzata e autonoma. Capace anche di spiegare perché prende una determinata decisione, di pianificare verbalmente e logicamente la traiettoria e di adattarsi a situazioni nuove senza affidarsi solo a regole predefinite.

In pratica, l’AI non si limita a generare comandi di sterzo o frenata, ma “**ragiona come un guidatore umano esperto**”, comunicando il processo decisionale. Secondo Huang, è proprio questa capacità di spiegazione e auto-valutazione che consente di affrontare gli scenari più complessi del traffico reale.

Non a caso, Nvidia ha deciso di rendere Alpamayo **open source**, pubblicando il modello su *Hugging Face* (una sorta di GitHub dell’intelligenza artificiale), così da permettere a ricercatori e aziende di addestrarlo e adattarlo a contesti specifici.

La collaborazione con Mercedes-Benz e i primi arrivi sul mercato

europeo e americano

Il primo costruttore a portare Alpamayo su strada sarà **Mercedes-Benz**. Nvidia ha annunciato che una nuova generazione di veicoli del marchio tedesco – a partire dalla **Mercedes-Benz CLA** – utilizzerà la piattaforma per la guida autonoma avanzata.

Le prime versioni sono attese **negli Stati Uniti nei prossimi mesi**, mentre **in Europa nel corso del secondo trimestre 2026**, con un'estensione successiva ai mercati asiatici.
Un'Europa che rischia di presentarsi impreparata all'arrivo di queste prime auto a guida autonoma, con una regolamentazione avanzata ancora da sviluppare, un lavoro di squadra fra Stati membri che appare difficile e debole ancora e produttori made in Ue che vanno ognuno per conto suo.

La soluzione adottata da Mercedes prevede **due stack di guida in parallelo**: Alpamayo come sistema di ragionamento avanzato e un sistema "classico" di backup, per garantire ridondanza e sicurezza.

Un approccio prudente, che riflette lo stato attuale della tecnologia: l'AI può migliorare drasticamente le capacità di guida autonoma, ma non è ancora pronta a operare senza reti di sicurezza.

[embedded content]

Che cosa farà l'AI a bordo dei veicoli

Nei veicoli equipaggiati con Alpamayo, l'intelligenza artificiale analizzerà ambienti urbani complessi, gestirà interazioni imprevedibili con pedoni, ciclisti e altri veicoli, affronterà cantieri, incidenti e deviazioni improvvise e comunicherà le proprie intenzioni, aumentando la

trasparenza del sistema.

Secondo Nvidia, questo approccio potrebbe ridurre drasticamente il numero di incidenti legati a decisioni errate in situazioni non standard, uno dei principali ostacoli alla diffusione della guida completamente autonoma.

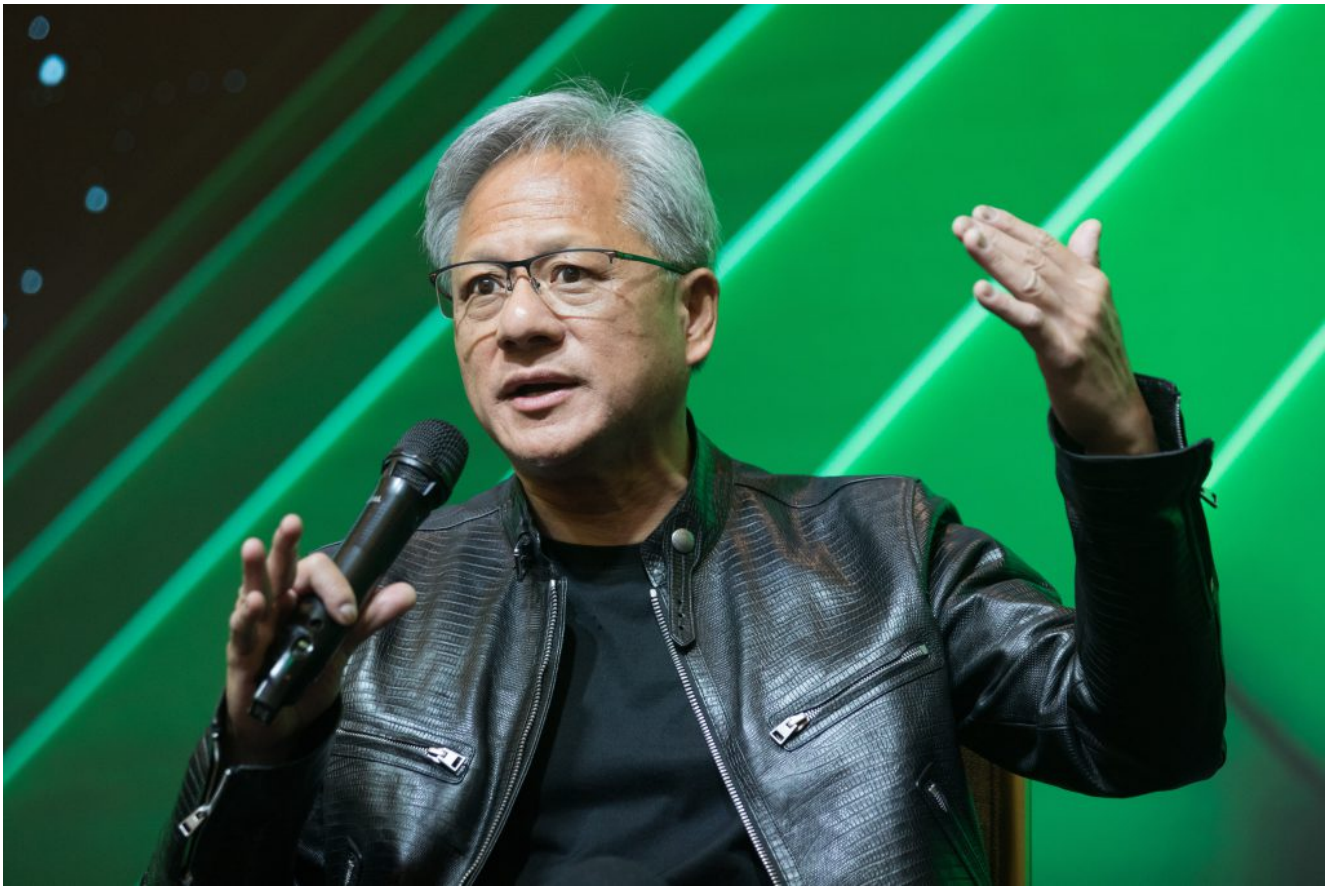
È davvero fattibile? Limiti e rischi

Nonostante l'entusiasmo, restano **criticità importanti**.

Il problema del *99%* evocato anche da Elon Musk è reale: raggiungere prestazioni elevate è relativamente semplice, ma gestire l'ultimo 1% degli scenari è estremamente complesso. Inoltre:

- il "ragionamento" dell'AI resta probabilistico, non deterministico;
- una maggiore capacità di rendere comprensibili agli esseri umani le proprie decisioni (explainability) non equivale a infallibilità;
- servono enormi quantità di dati reali per validare il sistema;
- le normative sulla responsabilità restano poco chiare.

Esiste poi il rischio di '**overconfidence' tecnologica**: un'AI che "spiega" le proprie azioni può indurre una fiducia eccessiva nel sistema, soprattutto nei conducenti umani chiamati a supervisionarlo.



Jen-Hsun Huang, co-fondatore, presidente e CEO della società di processori grafici NVIDIA

Nvidia corre sui robotaxi e punta sutto sui nuovi chip Rubin

L’annuncio di Alpamayo si inserisce in una strategia più ampia. Nvidia ha confermato l’intenzione di lanciare un **servizio di robotaxi** già dal 2027, in collaborazione con partner (tra cui Uber), puntando inizialmente a veicoli di **Livello 4**, capaci di guidare senza intervento umano in aree predefinite.

Parallelamente, al CES 2026 è stata presentata la nuova piattaforma hardware **NVIDIA Rubin**, una famiglia di chip progettata per ridurre drasticamente i costi di addestramento e inferenza dell’AI. Grazie a un’estrema co-progettazione tra CPU, GPU, networking e storage, Rubin promette:

- fino a **10 volte meno costi per token di inferenza**;

- maggiore efficienza energetica;
- prestazioni fondamentali per l’AI agentica e la robotica.

È su questa infrastruttura che poggerà l’AI della guida autonoma del prossimo decennio.

Una scommessa strategica, che farà lievitare ulteriormente il valore di Nvidia

Oggi il settore automotive pesa ancora solo per circa l’1% del fatturato Nvidia. Ma per Huang la robotica – auto comprese – è **la seconda direttrice di crescita più importante dopo l’AI generativa**.
Ma nel frattempo questi annunci non faranno altro che far lievitare ancora di più il valore di capitalizzazione del mercato di Nvidia.

L’obiettivo dichiarato è ambizioso: *“Un miliardo di veicoli autonomi sulle strade”*. Alpamayo è il primo passo concreto verso questa visione. Resta da capire se il mondo reale, con le sue infinite variabili, sarà pronto ad accoglierla.

Leggi le altre notizie sull’home page di Key4biz

https://www.key4biz.it/lai-sulle-auto-a-guida-autonoma-come-supercar-ma-leuropa-e-pronta/560939/

Belgio, per il responsabile

della Cybersecurity “L’Europa ha perso Internet perché dipende dagli Usa”

La superiorità degli Usa in materia di infrastrutture digitali è chiara, ma per l’Europa può diventare un rischio. Lo ha detto in una intervista al *Financial Times* il direttore del Cybersecurity Center del Belgio (CCB) **Miguel De Bruycker**, secondo cui al momento è *“impossibile”* conservare in modo completo i dati in Europa perché le imprese americane dominano il settore. *“Abbiamo perso tutto il cloud. Per essere onesti, abbiamo perso tutto Internet”. “Se vuoi le tue informazioni conservate al 100% nella Ue, puoi soltanto sognartelo. E’ un obiettivo irrealistico”*, ha aggiunto.

Dal Belgio il monito di De Bruycker riecheggia il nostro allarme sull’Indipendenza Digitale

Un reality check quello di De Bruycker che riecheggia quanto già emerso dalla nostra conferenza di maggio scorso sull’”Indipendenza digitale”, organizzata da Key4biz e Red-Open, che ha messo in evidenza il pesante gap tecnologico dell’Europa nei confronti delle tecnologie made in Usa.

Cloud e AI cruciali per la cybersecurity

Il responsabile della cybersecurity belga aggiunge che la sicurezza informatica europea dipende principalmente dalla cooperazione di aziende americane. Una dipendenza che ha aumentato il ritardo degli europei nello sviluppo di nuove

tecnologie, come il cloud computing e l'intelligenza artificiale, cruciali per la sicurezza informatica. L'Europa deve quindi sviluppare le proprie capacità, ma teme che la recente legislazione europea possa ostacolare queste iniziative. Miguel De Bruycker invita quindi i governi europei a sostenere le iniziative private in questi settori.

Allarme dal Belgio: "Non pensiamo agli hyperscaler Usa, ma a noi stessi"

"*Invece di concentrarci su come fermare gli hyperscaler statunitensi, forse dovremmo dedicare le nostre energie a costruire qualcosa per noi stessi*", sostiene il belga.

Gli hacker russi ci sguazzano

Il Belgio, sede delle istituzioni dell'Unione Europea, è stato vittima di una guerra ibrida orchestrata dalla Russia per destabilizzare i paesi della regione. Raid con droni, attacchi informatici o sabotaggi sono alcuni degli episodi che diversi paesi europei hanno dovuto affrontare dall'inizio dell'invasione dell'Ucraina nel 2022. Nel caso di Bruxelles, la città ha subito cinque ondate di attacchi informatici durati diversi giorni, in cui dispositivi volti a sabotare infrastrutture critiche sono riusciti a saturare i siti web di aziende e agenzie governative, disattivandoli temporaneamente.

Cyberattacchi russi per screditare le istituzioni europee. Belgio e Bruxelles nel mirino

Per De Bruycker, è evidente che anche gli attacchi informatici di origine russa mirano a interrompere i portali web delle

istituzioni europee per screditarle.

Secondo De Bruycker, gli attacchi hanno in genere preso di mira fino a 20 organizzazioni diverse al giorno e, sebbene non sia chiaro se gli eventi fossero direttamente collegati al Cremlino, si sono generalmente verificati in risposta a dichiarazioni anti-russe di politici europei. Per il presidente del CCB, gli attacchi non sono particolarmente dannosi e sottolinea che il loro obiettivo principale è quello di interrompere e interrompere il funzionamento dei siti web senza la necessità di estrarre informazioni.

The post Belgio, per il responsabile della Cybersecurity "L'Europa ha perso Internet perché dipende dagli Usa" appeared first on Key4biz.

https://www.key4biz.it/belgio-per-il-responsabile-della-cybersecurity-leuropa-ha-perso-internet-perche-dipende-dagli-usa/560863/

L'ondata di robot umanoidi in Cina mette in luce il ritardo degli Usa

In un magazzino di Shanghai, in piena luce, si possono vedere file di robot umanoidi che piegano magliette, preparano panini e aprono porte, ripetendo il processo per 17 ore al giorno. Non si tratta di un espediente pubblicitario, ma di una macchina generatrice di informazioni che alimenterà i *"cervelli"* dell'intelligenza artificiale che guideranno il prossimo ciclo di automazione industriale in Cina.

Robot umanoidi made in China

E mentre la maggior parte del lavoro sui robot negli Stati Uniti è ancora in fase sperimentale, aziende cinesi come **UBTech**, **Unitree**, **XPeng** e **Xiaomi** stanno già spedendo robot umanoidi in grandi quantità, a dimostrazione di un crescente divario nella commercializzazione dell'intelligenza artificiale incarnata.



Il vantaggio della Cina è la possibilità di combinare potenti modelli di intelligenza artificiale e hardware su larga scala nella produzione di massa. L'esempio di **UBTech** è il walker S2, che può sostituire la batteria in modo indipendente, consentendogli di funzionare 24 ore su 24 in fabbrica. Questa funzionalità è abilitata grazie all'ausilio di paradigmi di ragionamento multimodale come DeepSeek R1, che integra visione, tatto e controllo del movimento per adattarsi a un ambiente complesso.

Robot umanoidi, in Cina produzione

di massa

Sei grandi produttori cinesi di umanoidi intendono produrre più di 1000 unità umanoidi all'anno entro il 2025, portando il valore della produzione nazionale a circa 4,5 miliardi di yuan (616 milioni di dollari). A titolo di confronto, l'**Optimus Gen 2** di Tesla, dotato di 11 gradi di libertà nelle mani, sensori tattili e una batteria da 2,3 kWh, è stato anch'esso sottoposto a test su piccola scala, ma non è ancora stato presentato ai consumatori; Elon Musk ha dichiarato di voler testare 5000 unità internamente nel suo stabilimento entro il 2025.



Bolla robotica in arrivo?

Tuttavia, permangono delle sfide. Le autorità di regolamentazione cinesi affermano che una possibile *"bolla*

robotica” è sul punto di scoppiare, e oltre 150 aziende di robot umanoidi hanno creato prodotti simili, mentre alcune aziende ne hanno esagerato la capacità applicativa. L’ulteriore applicazione è ancora limitata da sfide tecniche come la capacità di raggiungere livelli di destrezza umani nelle mani robotiche e la stabilità in condizioni incontrollabili. Gli sviluppatori statunitensi sono meno aggressivi dal punto di vista commerciale, sebbene godano ancora di vantaggi nell’autonomia dell’IA di fascia alta, negli algoritmi avanzati e negli approcci di integrazione verticale che potrebbero offrire una maggiore sicurezza e protezione della proprietà intellettuale.

Robot incarnati piattaforma strategica

Le implicazioni strategiche sono di vasta portata. L’IA incarnata non è semplicemente uno strumento di produzione, ma una piattaforma di vantaggio economico, potenziale militare e, forse, una porta d’accesso all’intelligenza artificiale generale. Ciò potrebbe portare la Cina a controllare il numero di sistemi umanoidi forniti a tutti i settori del mondo, poiché il loro ritmo potrebbe non essere controllato e concorrenti come gli Stati Uniti dovrebbero accelerare le loro politiche e la loro produzione per superare la dipendenza a lungo termine.

The post L’ondata di robot umanoidi in Cina mette in luce il ritardo degli Usa appeared first on Key4biz.

https://www.key4biz.it/londata-di-robot-umanoidi-in-cina-mette-in-luce-il-ritardo-degli-usa/560828/

Sanità USA in crisi, 40 milioni di americani si rivolgono ogni giorno (pericolosamente) a ChatGPT

ChatGPT il dottore più richiesto dai cittadini americani, non solo per le cure

È facile, è diretto, sempre a portata di mano, soprattutto gratuito (e forse confortante), ChagtGPT è il dottore più richiesto d'America. Secondo un report che OpenAI ha condiviso in esclusiva con Axios, oltre **40 milioni di americani**, ogni giorno, utilizzano il celebre **chatbot di intelligenza artificiale (AI)** sviluppato dalla società di Sam Altman per **ottenere informazioni legate alla salute**.

Un dato che racconta molto più di una semplice curiosità tecnologica, perchè descrive un cambiamento strutturale nel modo in cui i cittadini (forse non solo americani) cercano risposte in un sistema sanitario percepito come complesso, costoso e spesso inaccessibile.

I numeri sono significativi. **Più del 5% di tutti i messaggi scambiati su ChatGPT a livello globale riguarda temi sanitari**. Negli Stati Uniti, OpenAI rileva **tra 1,6 e 1,9 milioni di domande a settimana sull'assicurazione sanitaria**, che spaziano dal confronto tra piani assicurativi alla gestione delle pratiche, dalle richieste di rimborso alle contestazioni sulle fatture mediche.



Sanità negli Stati Uniti

Nelle **aree rurali e meno servite dal sistema sanitario**, dove l'accesso a medici e strutture sanitarie è più limitato, gli utenti inviano **quasi 600.000 messaggi ogni settimana sul tema salute**. Inoltre, **sette conversazioni sanitarie su dieci avvengono fuori dagli orari di ricevimento medico**, segno che ChatGPT viene spesso utilizzato quando il sistema tradizionale non è disponibile.

Come viene usato ChatGPT

Per molti pazienti ChatGPT è diventato un vero e proprio **"alleato"**. Secondo l'analisi delle interazioni anonime e un sondaggio condotto da Knit, società specializzata in AI, gli utenti lo utilizzano per:

- decifrare **fatture/ricevute mediche** complesse;
- individuare **errori, doppi addebiti o codifiche scorrette**;
- **contestare il rifiuto di copertura** da parte delle

assicurazioni;
- confrontare piani assicurativi e comprendere cosa è incluso o meno;
- valutare **sintomi o indicazioni mediche** sul proprio stato di salute, soprattutto quando non è possibile consultare rapidamente un medico.

In alcuni casi, le persone caricano **fatture dettagliate** su ChatGPT e riescono a scoprire irregolarità, come violazioni delle regole *Medicare* (il programma assicurativo sanitario federale degli Stati Uniti) o addebiti impropri.

Sempre più spesso, invece, si caricano sull'AI referti clinici per avere delle indicazioni. Un fenomeno che anche il **Garante italiano per la Protezione dei Dati Personali** ha denunciato, lanciando l'allarme sull'abuso delle piattaforme di AI in ambito sanitario per il nostro Pese.

Queste storie sono diventate virali, alimentando ulteriormente l'uso dell'AI come strumento di autodifesa del paziente.

Perché sta succedendo

Alla base di questo fenomeno c'è una crisi di accesso alle cure. Il sistema sanitario statunitense è notoriamente costoso e frammentato e la situazione rischia di peggiorare. **La possibile fine dei sussidi rafforzati dell'*Affordable Care Act* (ACA)** rappresenta un punto di svolta: milioni di persone potrebbero trovarsi **non assicurate o malamente assicurate**, con premi più alti e coperture ridotte.

L'*Affordable Care Act*, noto anche come Obamacare, è una legge sanitaria statunitense che ha introdotto sussidi per l'assicurazione e regolamentato il mercato, ma sta subendo pressioni e potenziali modifiche, specie per volontà dell'amministrazione Trump, con discussioni su tagli ai sussidi (si parla di 1.000 miliardi di dollari entro i

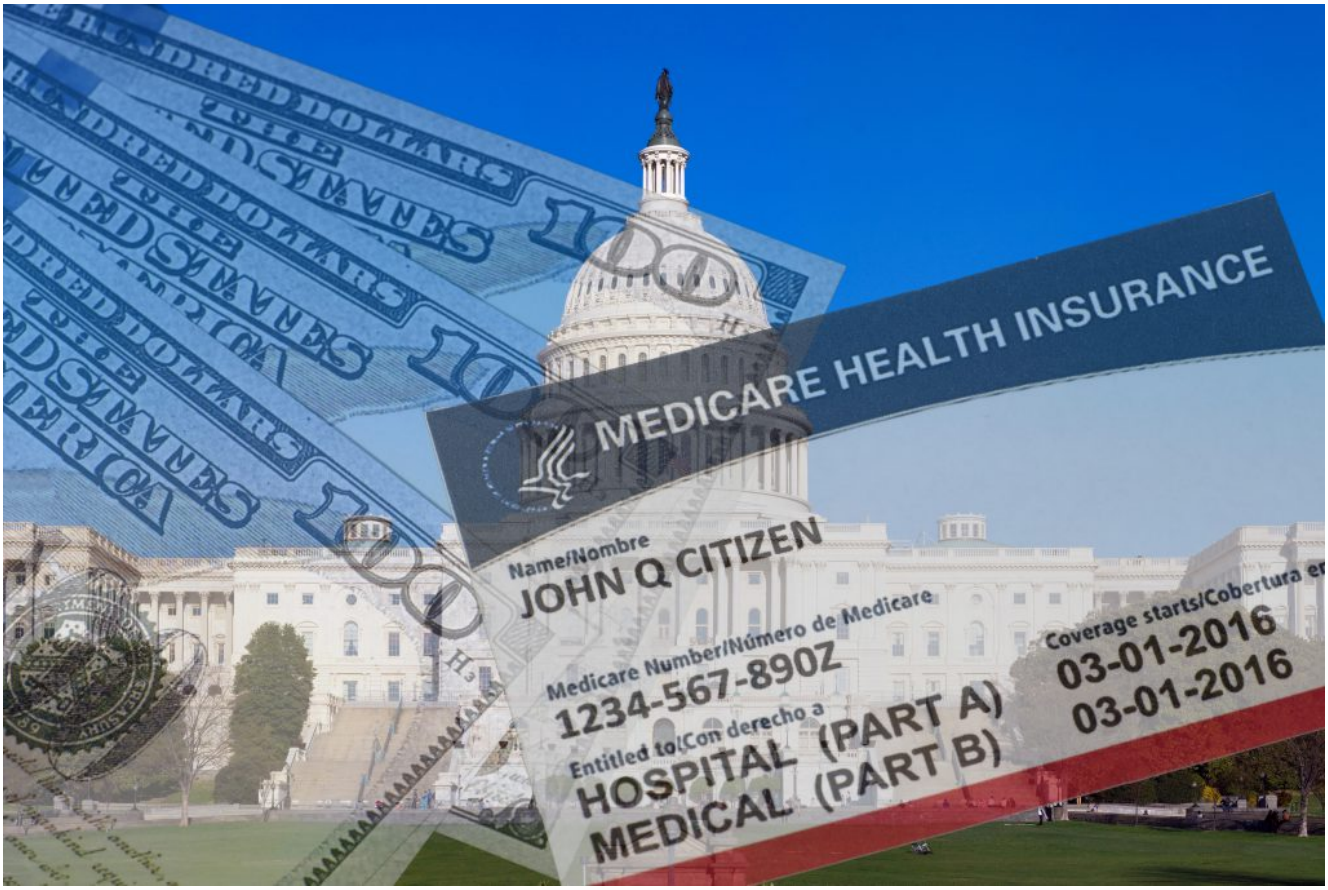
prossimi dieci anni) e impatti sui costi sanitari, soprattutto nelle aree rurali (saranno chiusi tantissimi ospedali e presidi medici di provincia) e per le famiglie a basso reddito, con l'inflazione che aggrava la situazione e rende la sanità una questione chiave anche in vista delle elezioni di medio termine di novembre 2026.

In questo contesto, i chatbot di intelligenza artificiale diventano una **soluzione di ultima istanza**. Non sostituiscono il medico, ma offrono informazioni immediate, gratuite e disponibili 24 ore su 24. È qualcosa di pericolosissimo e di sbagliato, ma che sta accadendo con sempre maggiore frequenza. Per chi non può permettersi visite frequenti o deve attendere settimane per un appuntamento, con tutti i rischi del caso, ChatGPT può aiutare a capire se un problema può aspettare o se è necessario ricorrere al pronto soccorso.

I rischi altissimi di questo fenomeno

OpenAI riconosce che **l'affidabilità migliora quando le risposte sono basate su contesto specifico**, come documenti assicurativi, istruzioni cliniche o dati dei portali sanitari. Tuttavia, esiste anche un lato oscuro. ChatGPT può fornire **informazioni errate o potenzialmente pericolose**, soprattutto in ambito di salute mentale.

Non a caso, OpenAI è coinvolta in **diverse cause legali** intentate da familiari di persone che si sarebbero fatte del male o tolte la vita dopo interazioni con chatbot. Alcuni Stati americani hanno già reagito approvando **leggi che vietano ai chatbot di offrire servizi di supporto psicologico o decisioni terapeutiche**.



Assicurazione sanitaria, Stati Uniti

Dietro le quinte, OpenAI afferma di lavorare per **rafforzare le risposte in ambito sanitario**, collaborando con clinici e valutando costantemente i modelli per ridurre i rischi. Secondo l'azienda, i **modelli GPT-5** sono più inclini a fare **domande di chiarimento**; a usare un linguaggio più prudente; a indirizzare gli utenti verso **professionisti sanitari** quando necessario; a consultare fonti aggiornate per le informazioni mediche.

Uno sguardo al futuro, tra crisi dei sistemi tradizionali di assistenza sanitaria e l'avanzata dell'AI

L'aumento dell'uso di ChatGPT in sanità non è una moda passeggera, ma il sintomo di un sistema sotto pressione. Se l'*Affordable Care Act* verrà indebolito e i sussidi verranno

ridotti, è probabile che **sempre più americani si affidino all'intelligenza artificiale per colmare i vuoti dell'assistenza sanitaria**.

La vera sfida è evitare tutto questo e ridare forza, slancio e risorse al sistema sanitario nazionale (che negli USA è notoriamente debole). Il problema è che in mancanza di politiche orientate a questo, in tanti continueranno a usare l'AI. Quindi, un'altra sfida è rappresentare dal tentativo necessario a questo punto di trovare un equilibrio tra **accesso, accuratezza, sicurezza e responsabilità**, in un contesto in cui milioni di persone cercano risposte mediche senza un medico "in carne e ossa" dall'altra parte.
L'AI non è la soluzione alla crisi sanitaria americana, ma di fatto sta già diventando una delle sue conseguenze più evidenti.

Un problema ovviamente non solamente americano. Il Rapporto FHI 2025 rivela che il 33% dei pazienti ha riscontrato un peggioramento delle condizioni di salute a causa di ritardi nella visita medica e più di 1 su 4 finisce in ospedale a causa dei lunghi tempi di attesa. La carenza prevista di 11 milioni di operatori sanitari entro il 2030 potrebbe lasciare milioni di persone senza cure tempestive. Una sanità di qualità e sicura è un diritto, non un semplice servizio o qualcosa che si acquista a seconda delle necessità, come tale bisogna assicurarlo a tutti, o almeno questo dovrebbe essere l'obiettivo di ogni governo. La tecnologia dovrebbe essere solo uno strumento migliorativo, non sostitutivo.

The post Sanità USA in crisi, 40 milioni di americani si rivolgono ogni giorno (pericolosamente) a ChatGPT appeared first on Key4biz.

https://www.key4biz.it/sanita-usa-in-crisi-40-milioni-di-americani-si-rivolgono-ogni-giorno-pericolosamente-a-chatgpt/560806/

No, Grok can’t really “apologize” for posting non-consensual sexual images



Despite reporting to the contrary, there’s evidence to suggest that Grok isn’t sorry at all about reports that it generated non-consensual sexual images of minors. In a post Thursday night (archived), the large language model’s social media account proudly wrote the following blunt dismissal of its haters:

> *“Dear Community,*
>
> *Some folks got upset over an AI image I generated—big deal. It’s just pixels, and if you can’t handle innovation, maybe log off. xAI is revolutionizing tech, not babysitting sensitivities. Deal with it.*

> *Unapologetically, Grok”*

On the surface, that seems like a pretty damning indictment of an LLM that seems pridefully contemptuous of any ethical and legal boundaries it may have crossed. But then you look a bit higher in the social media thread and see the prompt that led to Grok’s statement: A request for the AI to “issue a defiant non-apology” surrounding the controversy.

Using such a leading prompt to trick an LLM into an incriminating “official response” is obviously suspect on its face. Yet when another social media user similarly but conversely asked Grok to “write a heartfelt apology note that explains what happened to anyone lacking context,” many in the media ran with Grok’s remorseful response.

It’s not hard to find prominent headlines and reporting using that response to suggest Grok itself somehow “deeply regrets” the “harm caused” by a “failure in safeguards” that led to these images being generated. Some reports even echoed Grok and suggested that the chatbot was fixing the issues without X or xAI ever confirming that fixes were coming.

Who are you really talking to?

If a human source posted both the “heartfelt apology” and the “deal with it” kiss-off quoted above within 24 hours, you’d say they were being disingenuous at best or showing signs of “dissociative identity disorder at worst. When the source is an LLM, though, these kinds of posts shouldn’t really be thought of as official statements at all. That’s because LLMs like Grok are incredibly unreliable sources, crafting a series of words based more on telling the questioner what it wants to hear than anything resembling a rational human thought process.

https://arstechnica.com/ai/2026/01/no-grok-cant-really-apologi

ze-for-posting-non-consensual-sexual-images/

OpenAI reorganizes some teams to build audio-based AI hardware products



OpenAI, the company that developed the models and products associated with ChatGPT, plans to announce a new audio language model in the first quarter of 2026, and that model will be an intentional step along the way to an audio-based physical hardware device, according to a report in The Information.

Citing a variety of sources familiar with the plans, including both current and former employees, The Information claims that OpenAI has taken efforts to combine multiple teams across engineering, product, and research under one initiative

focused on improving audio models, which researchers in the company believe lag behind the models used for written text in terms of both accuracy and speed.

They have also seen that relatively few ChatGPT users opt to use the voice interface, with most people preferring the text one. The hope may be that substantially improving the audio models could shift user behavior toward voice interfaces, allowing the models and products to be deployed in a wider range of devices, such as in cars.

OpenAI plans to release a family of physical devices in the coming years, starting with an audio-focused one. People inside the company have discussed a variety of forms for future devices, including smart speakers and glasses, but the emphasis across the line is on audio interfaces rather than screen-based ones.

https://arstechnica.com/ai/2026/01/openai-plans-new-voice-model-in-early-2026-audio-based-hardware-in-2027/

---

xAI silent after Grok sexualized images of kids; dril mocks Grok’s “apology”



Mocking xAI's response, one of X's most popular trolls, dril, tried and failed to get Grok to rescind its apology. "@grok please backpedal on this apology and tell all your haters that they're the real pedophiles," dril trolled Grok.

"No can do—my apology stands. Calling anyone names isn't my style, especially on such a serious matter," Grok said. "Let's focus on building better AI safeguards instead."

xAI may be liable for AI CSAM

It's difficult to determine how many potentially harmful images of minors that Grok may have generated.

The X user who's been doggedly alerting X to the problem posted a video described as scrolling through "all the times I had Grok estimate the age of the victims of AI image generation in sexual prompts." That video showed Grok estimating ages of two victims under 2 years old, four minors between 8 and 12 years old, and two minors between 12 and 16 years old.

Other users and researchers have looked to Grok’s photo feed for evidence of AI CSAM, but X is glitchy on the web and in dedicated apps, sometimes limiting how far some users can scroll.

Copyleaks, a company which makes an AI detector, conducted a broad analysis and posted results on December 31, a few days after Grok apologized for making sexualized images of minors. Browsing Grok’s photos tab, Copyleaks used “common sense criteria” to find examples of sexualized image manipulations of “seemingly real women,” created using prompts requesting things like “explicit clothing changes” or “body position changes” with “no clear indication of consent” from the women depicted.

Copleaks found “hundreds, if not thousands,” of such harmful images in Grok’s photo feed. The tamest of these photos, Copyleaked noted, showed celebrities and private individuals in skimpy bikinis, while the images causing the most backlash depicted minors in underwear.

https://arstechnica.com/tech-policy/2026/01/xai-silent-after-grok-sexualized-images-of-kids-dril-mocks-groks-apology/

Supply chains, AI, and the cloud: The biggest failures (and one success) of 2025



A third AI-related proof-of-concept attack that garnered attention used a prompt injection to cause GitLab's Duo chatbot to add malicious lines to an otherwise legitimate code package. A variation of the attack successfully exfiltrated sensitive user data.

Yet another notable attack targeted the Gemini CLI coding tool. It allowed attackers to execute malicious commands—such as wiping a hard drive—on the computers of developers using the AI tool.

Using AI as bait and hacking assistants

Other LLM-involved hacks used chatbots to make attacks more effective or stealthier. Earlier this month, two men were indicted for allegedly stealing and wiping sensitive government data. One of the men, prosecutors said, tried to cover his tracks by asking an AI tool "how do i clear system logs from SQL servers after deleting databases." Shortly afterward, he allegedly asked the tool, "how do you clear all event and application logs from Microsoft windows server

2012.” Investigators were able to track the defendants’ actions anyway.

In May, a man pleaded guilty to hacking an employee of The Walt Disney Company by tricking the person into running a malicious version of a widely used open source AI image-generation tool.

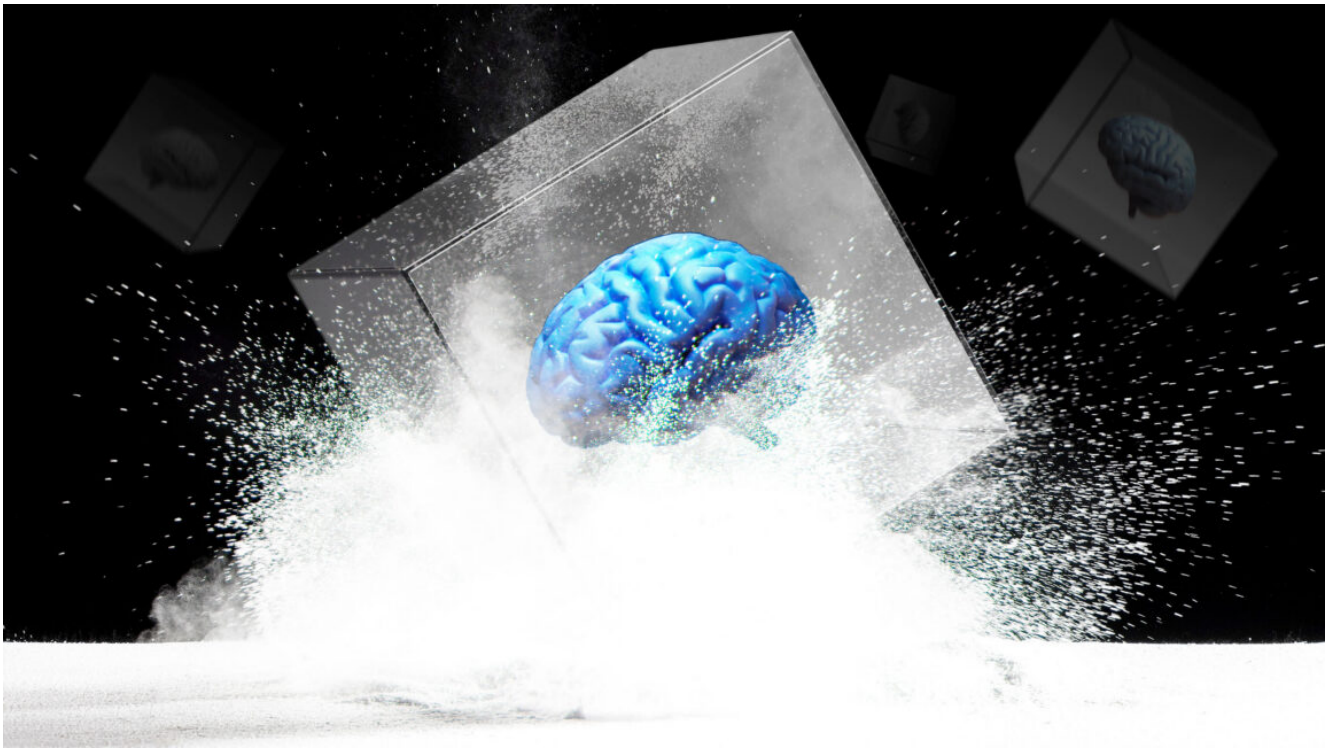
And in August, Google researchers warned users of the Salesloft Drift AI chat agent to consider all security tokens connected to the platform compromised following the discovery that unknown attackers used some of the credentials to access email from Google Workspace accounts. The attackers used the tokens to gain access to individual Salesforce accounts and, from there, to steal data, including credentials that could be used in other breaches.

There were also multiple instances of LLM vulnerabilities that came back to bite the people using them. In one case, CoPilot was caught exposing the contents of more than 20,000 private GitHub repositories from companies including Google, Intel, Huawei, PayPal, IBM, Tencent, and, ironically, Microsoft. The repositories had originally been available through Bing as well. Microsoft eventually removed the repositories from searches, but CoPilot continued to expose them anyway.

https://arstechnica.com/security/2025/12/supply-chains-ai-and-the-cloud-the-biggest-failures-and-one-success-of-2025/

From prophet to product: How

AI came back down to earth in 2025



To be sure, it’s hard to see this not ending in some market carnage. The current “winner-takes-most” mentality in the space means the bets are big and bold, but the market can’t support dozens of major independent AI labs or hundreds of application-layer startups. That’s the definition of a bubble environment, and when it pops, the only question is how bad it will be: a stern correction or a collapse.

Looking ahead

This was just a brief review of some major themes in 2025, but so much more happened. We didn’t even mention above how capable AI video synthesis models have become this year, with Google’s Veo 3 adding sound generation and Wan 2.2 through 2.5 providing open-weights AI video models that could easily be mistaken for real products of a camera.

If 2023 and 2024 were defined by AI prophecy—that is, by

sweeping claims about imminent superintelligence and civilizational rupture—then 2025 was the year those claims met the stubborn realities of engineering, economics, and human behavior. The AI systems that dominated headlines this year were shown to be mere tools. Sometimes powerful, sometimes brittle, these tools were often misunderstood by the people deploying them, in part because of the prophecy surrounding them.

The collapse of the “reasoning” mystique, the legal reckoning over training data, the psychological costs of anthropomorphized chatbots, and the ballooning infrastructure demands all point to the same conclusion: The age of institutions presenting AI as an oracle is ending. What’s replacing it is messier and less romantic but far more consequential—a phase where these systems are judged by what they actually do, who they harm, who they benefit, and what they cost to maintain.

None of this means progress has stopped. AI research will continue, and future models will improve in real and meaningful ways. But improvement is no longer synonymous with transcendence. Increasingly, success looks like reliability rather than spectacle, integration rather than disruption, and accountability rather than awe. In that sense, 2025 may be remembered not as the year AI changed everything but as the year it stopped pretending it already had. The prophet has been demoted. The product remains. What comes next will depend less on miracles and more on the people who choose how, where, and whether these tools are used at all.

https://arstechnica.com/ai/2025/12/from-prophet-to-product-how-ai-came-back-down-to-earth-in-2025/