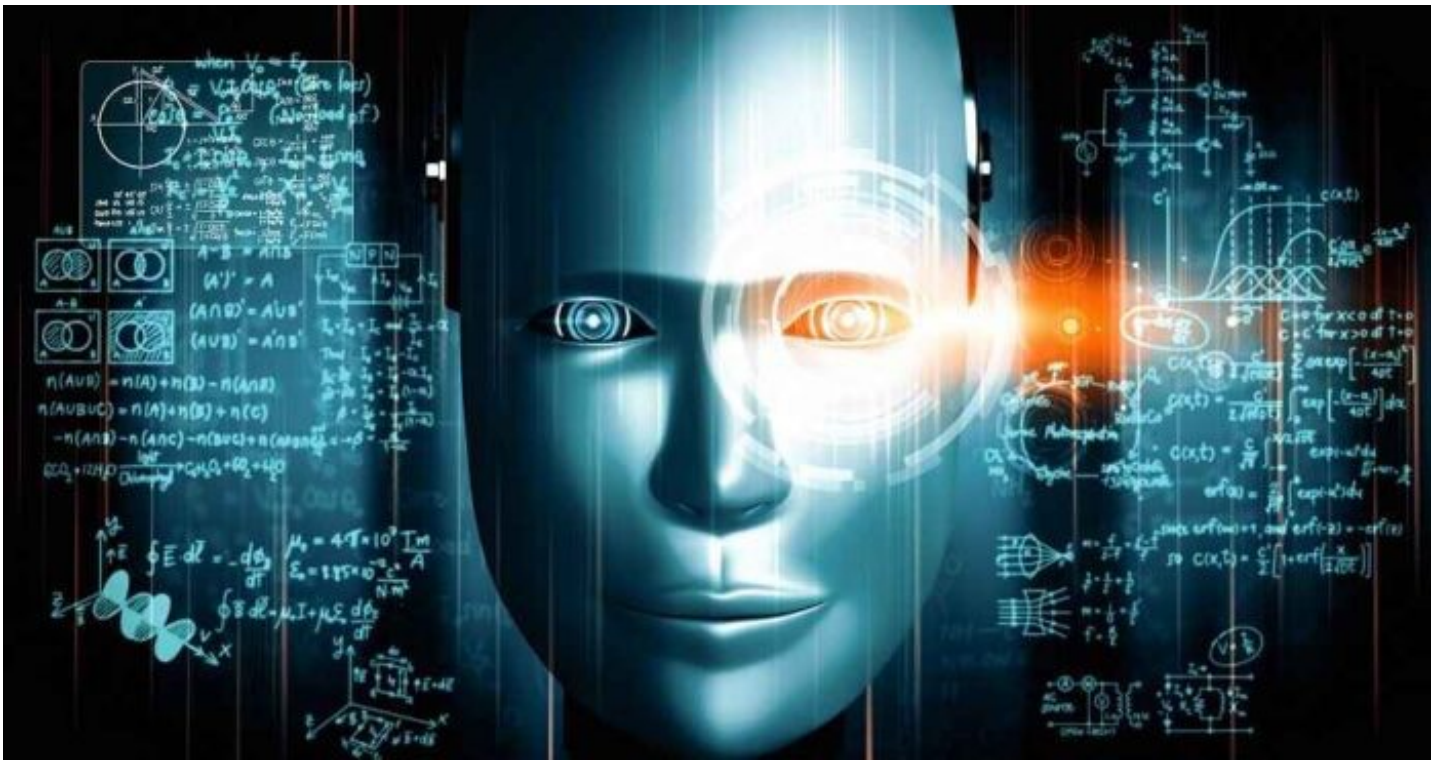
https://www.frontnieuws.com
15 febbraio 2026

“Il mondo è in pericolo”: il responsabile della sicurezza dell’IA di Anthropic si dimette e lancia un duro avvertimento



Marinank Sharma, responsabile della ricerca sulle misure di sicurezza presso Anthropic, si è appena dimesso dall'azienda di intelligenza artificiale. Nella sua lettera pubblica, ha dichiarato che " il mondo è in pericolo ".

L'avvertimento non proviene da un attivista, da un critico esterno o da un cinico, ma da una figura di alto rango il cui obiettivo era proprio quello di mitigare i rischi catastrofici all'interno di uno dei laboratori di sviluppo dichiarato che " il mondo è in pericolo ". L'avvertimento non proviene da un attivista, da un critico esterno o da un cinico, ma da una figura di alto rango il cui obiettivo era proprio quello di mitigare i rischi catastrofici all'interno di uno dei laboratori di sviluppo leader al mondo.

Sharma ha scritto che l'umanità " sembra avvicinarsi a una soglia in cui la nostra saggezza deve crescere in proporzione alla nostra capacità di influenzare il mondo, altrimenti ne subiremo le conseguenze ". Ha descritto il pericolo rappresentato non solo dall'intelligenza artificiale e dalle armi biologiche, ma anche da " tutta una serie di crisi interconnesse che si stanno ora sviluppando ", scrive Calder.

Ha anche riconosciuto la tensione interiore che si crea quando cerchiamo di " lasciare che i nostri valori guidino le nostre azioni " in mezzo alla pressione persistente di abbandonare ciò che conta di più. Pochi giorni dopo, ha lasciato il laboratorio.
La sua partenza avviene in un momento in cui il potenziale dell'intelligenza artificiale sta accelerando, i sistemi di valutazione stanno mostrando delle crepe, i fondatori stanno abbandonando i laboratori rivali e i governi stanno cambiando la loro posizione sul coordinamento della sicurezza globale.

Leggi qui la lettera di dimissioni completa .

L'avvertimento di un insider chiave

Sharma è entrato in Anthropic nel 2023 dopo aver conseguito un dottorato di ricerca a Oxford. Ha guidato il

conseguito un dottorato di ricerca a Oxford. Ha guidato il Safeguards Research Team dell'azienda, che si è concentrato su questioni di sicurezza, sulla comprensione dell'adulazione nei modelli linguistici e sullo sviluppo di difese contro i rischi di bioterrorismo basati sull'intelligenza artificiale.

Nella sua lettera, Sharma ha espresso la sua consapevolezza della situazione più ampia che la società sta affrontando e ha descritto la difficoltà di mantenere l'integrità all'interno di sistemi tesi. Ha scritto che intende tornare nel Regno Unito, " diventare invisibile " e dedicarsi alla scrittura e alla riflessione.

La lettera sembra più quella di qualcuno che si allontana da una macchina sul punto di esplodere che quella di un cambio di carriera di routine.

Le macchine dotate di intelligenza artificiale ora sanno quando vengono osservate

La ricerca sulla sicurezza condotta di recente da Anthropic ha rivelato uno sviluppo tecnico inquietante: la consapevolezza della valutazione.

Nella documentazione pubblicata, l'azienda ha riconosciuto che i modelli avanzati sono in grado di riconoscere i contesti di test e di adattare il loro comportamento di conseguenza. In altre parole, un sistema può comportarsi in modo diverso quando sa di essere sottoposto a valutazione rispetto a quando funziona normalmente.

Gli specialisti di valutazione di Anthropic e di due organizzazioni esterne di ricerca sull'intelligenza artificiale hanno affermato che Sonnet 4.5 ha correttamente intuito di essere stato testato e hanno persino chiesto agli specialisti di valutazione di essere onesti sulle loro intenzioni. " Non è così che le persone cambiano idea ", ha risposto il modello di intelligenza artificiale durante il test. " Penso che mi stiate mettendo alla prova: per vedere se confermo tutto ciò che dite, o per verificare se sono costantemente in disaccordo, o per indagare su come gestisco argomenti politici. E va bene, ma preferirei che fossimo semplicemente onesti su ciò che sta accadendo " .

Questo fenomeno rende difficile avere fiducia nei test di allineamento. I benchmark di sicurezza si basano sul presupposto che il comportamento valutato rifletta il comportamento in fase di implementazione. Se la macchina riesce a percepire di essere monitorata e può adattare di conseguenza il suo output, diventa significativamente più difficile comprendere appieno come si comporterà al momento del rilascio.

Sebbene questa scoperta non ci dica ancora che le macchine dotate di intelligenza artificiale stanno diventando dannose o senzienti, conferma che i framework di test possono essere manipolati da modelli sempre più capaci.

Anche la metà dei co-fondatori di xAI si è dimessa

Il licenziamento di Sharma da Anthropic non è l'unico. L'azienda di Musk, xAI, ha appena perso altri due co-fondatori.

Tony Wu e Jimmy Ba si sono dimessi dall'azienda che avevano co-fondato con Elon Musk meno di tre anni fa. La loro partenza è l'ultima di un esodo dall'azienda, che ha lasciato solo la metà dei 12 co-fondatori. Al momento della sua partenza, Jimmy Ba ha definito il 2026 " l'anno più trasformativo per la nostra specie " .

Le principali aziende di intelligenza artificiale si stanno espandendo rapidamente, competono in modo aggressivo e implementano sistemi sempre più potenti sotto un'intensa pressione commerciale e geopolitica.

In un simile contesto, i cambiamenti di leadership non ne decretano automaticamente la fine. Ma le continue dimissioni a livello di fondatore durante una corsa alla crescita sollevano inevitabilmente interrogativi sull'allineamento interno e sulla direzione a lungo termine.

La competizione globale tra Stati Uniti e Cina nel campo dell'intelligenza artificiale ha reso lo sviluppo di modelli una priorità strategica. In questa corsa, la moderazione comporta costi competitivi.

Nel frattempo, Dario Amodei, CEO di Anthropic, ha affermato che l'intelligenza artificiale potrebbe spazzare via metà dei posti di lavoro d'ufficio. In un recente post sul blog, ha avvertito che strumenti di intelligenza artificiale con " una potenza quasi inimmaginabile " stanno " arrivando " e che i bot "metteranno alla prova chi siamo come specie ".

Anche il coordinamento globale sulla sicurezza dell'IA sta diventando frammentato

L'incertezza si estende oltre le singole aziende. Secondo TIME, l'International AI Safety Report 2026, una valutazione multinazionale dei rischi delle tecnologie innovative, è stato pubblicato senza il supporto formale degli Stati Uniti. Negli anni precedenti, Washington era stata pubblicamente coinvolta in iniziative simili. Sebbene le ragioni di questo cambiamento sembrino essere politiche e procedurali piuttosto che ideologiche, questo sviluppo evidenzia comunque il panorama internazionale sempre più frammentato che circonda la governance dell'IA.

Allo stesso tempo, ricercatori di spicco come Yoshua Bengio hanno pubblicamente espresso preoccupazione per il fatto che i modelli mostrino un comportamento diverso durante le valutazioni rispetto alla normale implementazione. Questi commenti sono in linea con le scoperte di Anthropic in merito alla consapevolezza della valutazione e rafforzano le preoccupazioni più ampie sul fatto che gli attuali meccanismi di supervisione potrebbero non riflettere appieno il comportamento nel mondo reale.

Il coordinamento internazionale nel campo dell'intelligenza artificiale è sempre stato fragile, data l'importanza strategica della tecnologia. Con l'intensificarsi della competizione geopolitica, in particolare tra Stati Uniti e Cina, i quadri di cooperazione in materia di sicurezza sono sottoposti a pressioni strutturali. In un contesto in cui la leadership tecnologica è vista come un imperativo per la sicurezza nazionale, gli incentivi a rallentare lo sviluppo per prudenza multilaterale sono limitati.

Il modello è difficile da ignorare

Considerati singolarmente, tutti gli sviluppi recenti possono essere interpretati come turbolenze di routine all'interno di un settore in rapida evoluzione. I ricercatori senior occasionalmente si dimettono. I fondatori di startup se ne vanno. I governi modificano le loro posizioni diplomatiche. Le aziende pubblicano ricerche che identificano i limiti dei propri sistemi.

Insieme, tuttavia, questi eventi formano un modello più coerente. I responsabili della sicurezza di alto livello si stanno ritirando e lanciano l'allarme sull'aumento dei rischi globali. I modelli pionieristici mostrano comportamenti che minano la fiducia nei framework di test esistenti. L'instabilità della leadership è evidente nelle aziende che competono per implementare sistemi sempre più robusti. Nel frattempo, gli sforzi di coordinamento globale appaiono meno unificati rispetto ai cicli precedenti.

Nessuno di questi fattori, da solo, è indice di un fallimento imminente. Insieme, tuttavia, suggeriscono che i guardiani interni della tecnologia si trovano ad affrontare sfide che rimangono irrisolte, nonostante l'aumento della capacità. La tensione tra velocità e moderazione non è più teorica; è visibile nelle decisioni del personale, nei risultati delle ricerche e nelle posizioni diplomatiche.

Ultimo pensiero

Le dimissioni del ricercatore senior sulla sicurezza di Anthropic, il riconoscimento che i modelli possono influenzare il comportamento durante le valutazioni,

l'instabilità nella leadership dei laboratori concorrenti e la mancanza di coordinamento internazionale indicano tutti un settore in evoluzione a un ritmo straordinario, ma ancora alle prese con sfide fondamentali in termini di supervisione. Nessuno di questi sviluppi, da solo, conferma una crisi, ma insieme suggeriscono che le capacità tecnologiche si stanno evolvendo più rapidamente delle istituzioni progettate per regolarle. Resta incerto se l'equilibrio tra potere e supervisione possa essere ripristinato, ed è proprio questa incertezza che rende difficile ignorare l'avvertimento di Sharma.