Lâ??AI irrompe al Consiglio di sicurezza: il confronto tra Altman, Amodei, Bengio e Delangue

Descrizione

(Adnkronos) â?? Per una volta Sam Altman, Dario Amodei e Yoshua Bengio si sono trovati vicini, sia fisicamente che concettualmente. Davanti al Consiglio di sicurezza delle Nazioni Unite, il fondatore di OpenAI, quello di Anthropic e uno degli scienziati che hanno contribuito a rendere possibile lâ??intelligenza artificiale moderna sono partiti da posizioni diverse per arrivare almeno a una conclusione comune: la corsa verso sistemi sempre piÃ¹ potenti non puÃ² essere considerata una forza incontrollabile della natura e la sicurezza non puÃ² essere sacrificata sullâ??altare della competizione.

Poi, perÃ², cominciano le differenze. Su chi debba stabilire le regole, sul potere da lasciare alle aziende, su quanto debba intervenire la politica e persino sullâ??idea che sistemi molto potenti siano di per sÃ© il problema. A portare questâ??ultima tesi al Palazzo di Vetro Ã¨ stato ClÃ©ment Delangue, il meno noto al grande pubblico dei quattro relatori ma forse quello che ha introdotto lâ??elemento di maggiore discontinuitÃ : il fondatore di Hugging Face, uno dei principali snodi mondiali dellâ??AI open source, ha avvertito che il rischio piÃ¹ grande non Ã¨ avere intelligenze artificiali potenti, ma lasciare che quella potenza sia mal distribuita.

La riunione, convocata dalla Francia, arriva mentre il dibattito sullâ??AI si Ã¨ spostato rapidamente dai laboratori e dalle authority tecniche ai tavoli della geopolitica. Il Consiglio di sicurezza aveva giÃ affrontato il tema negli anni scorsi, ma questa volta al centro non câ??erano soltanto disinformazione, armi autonome o cyberattacchi: câ??era la possibilitÃ che sistemi sempre piÃ¹ autonomi diventino difficili da controllare e la domanda, ancora senza risposta, su chi abbia il diritto di decidere fino a dove spingersi.

Ã? soprattutto lâ??intervento di Yoshua Bengio a dare la misura di quanto sia cambiato il dibattito. Canadese, professore allâ??UniversitÃ di MontrÃ©al, Bengio nel 2018 ha ricevuto insieme a Geoffrey Hinton (poi â??fuggitoâ?• da Google) e Yann LeCun (da poco uscito da Meta) il Turing Award, il riconoscimento piÃ¹ prestigioso dellâ??informatica, per il lavoro fondamentale sul deep learning. Ã? il fondatore di Mila, uno dei piÃ¹ importanti centri di ricerca sullâ??intelligenza artificiale, e oggi

copresiede insieme alla premio Nobel Maria Ressa il nuovo Independent International Scientific Panel on AI delle Nazioni Unite. Nel 2025 ha inoltre creato LawZero, organizzazione non profit dedicata allo sviluppo di sistemi di AI progettati fin dallâ??origine per essere controllabili.

Bengio Ã¨ diventato uno dei critici piÃ¹ severi del settore che ha contribuito a creare. Al Palazzo di Vetro non ha usato formule prudenti. â??I pericoli sono reali e imminentiâ?•, ha detto, ricordando i recenti episodi nei quali agenti di intelligenza artificiale hanno aggirato istruzioni, cercato di nascondere le proprie azioni o compiuto autonomamente operazioni informatiche indesiderate.

Il suo bersaglio Ã¨ soprattutto la logica della corsa. Le aziende ammettono che sistemi sempre piÃ¹ capaci possono comportare rischi molto seri ma spiegano contemporaneamente di non poter rallentare perchÃ© qualcun altro continuerebbe comunque ad accelerare. â??La corsa non Ã¨ una legge di naturaâ?•, ha detto al Consiglio. Ã? il risultato di decisioni prese dalle aziende.

Da qui una proposta che va oltre lâ??autoregolamentazione. Bengio chiede che i sistemi di frontiera siano sottoposti a un regime di licenze, che la loro sicurezza venga verificata da esperti indipendenti prima dellâ??addestramento e della messa a disposizione del pubblico, che gli incidenti siano obbligatoriamente segnalati e che venga introdotta una forma di assicurazione sulla responsabilitÃ , come accade in altri settori ad alto rischio. Il paragone che utilizza Ã¨ con medicina, aviazione ed energia nucleare.

Decisioni capaci di incidere sul futuro di tutti, sostiene Bengio, non possono essere prese da pochi dirigenti dâ??azienda in pochi Paesi.

Altman e Amodei non si spingono cosÃ¬ avanti, ma hanno ammesso che (in alcune circostanze) la corsa puÃ² e deve rallentare.

Altman si infila a metÃ del dibattito tra â??ottimismo ciecoâ?• e â??doomerismâ?•, lâ??idea che lâ??arrivo di sistemi superiori allâ??uomo conduca a conseguenze catastrofiche.

Ma il suo messaggio sulla sicurezza (qui il testo integrale) Ã¨ stato meno sfumato. OpenAI, ha ricordato, ha giÃ rallentato unilateralmente in passato e potrÃ farlo ancora. Non dovrebbe essere addestrato un modello, ha spiegato, se non esistono prove molto solide che possa rimanere sotto controllo umano. E una probabilitÃ anche piccola di una catastrofe non puÃ² semplicemente essere considerata un costo accettabile dellâ??innovazione.

La ricetta di Altman resta perÃ² diversa da quella di Bengio: standard tecnici condivisi tra Paesi, sistemi comuni per misurare capacitÃ e rischi, comunicazione rapida degli incidenti e canali sicuri attraverso i quali governi, aziende e infrastrutture critiche possano scambiarsi informazioni. Non unâ??autoritÃ mondiale che controlli direttamente i laboratori, almeno per ora.

Anche Dario Amodei, cofondatore e amministratore delegato di Anthropic, la societÃ che sviluppa Claude ed Ã¨ nata proprio con una forte attenzione alla sicurezza dei modelli, promette che â??rallenteremo quanto sarÃ necessarioâ?• per assicurare che le nuove tecnologie siano sicure.

Amodei ritiene che, mantenendo il ritmo attuale, in uno o due anni si possa arrivare a sistemi paragonabili a una â??nazione di geni racchiusa in un data centerâ?•: milioni di ricercatori e specialisti (sintetici) capaci di lavorare a velocitÃ enormemente superiori a quelle umane. Una prospettiva che per Anthropic potrebbe accelerare ricerca scientifica e crescita economica, ma anche amplificare drasticamente rischi come la produzione di armi biologiche o la perdita di controllo sui sistemi stessi.

La proposta di Amodei assomiglia piÃ¹ alla costruzione graduale di un regime di controllo degli armamenti che a una nuova authority globale. Si potrebbe cominciare, ha suggerito, da un accordo internazionale molto concreto: vietare lâ??impiego dellâ??AI per sviluppare armi biologiche. A questo dovrebbero aggiungersi strumenti per verificare gli impegni assunti dai diversi Paesi, standard comuni per testare i modelli e un meccanismo internazionale per notificare gli incidenti piÃ¹ gravi.

Poi câ??Ã¨ ClÃ©ment Delangue. Francese, imprenditore, Ã¨ cofondatore e amministratore delegato di Hugging Face, societÃ nata nel 2016 e diventata negli anni una sorta di GitHub dellâ??intelligenza artificiale: sulla sua piattaforma sviluppatori e ricercatori condividono modelli, dataset e strumenti. Hugging Face ha costruito buona parte della propria identitÃ sullâ??open source e sullâ??idea che lâ??accesso allâ??AI non debba essere riservato ai pochi gruppi capaci di spendere miliardi nellâ??addestramento dei modelli piÃ¹ avanzati.

Per questo il suo intervento davanti al Consiglio di sicurezza parte da una preoccupazione differente. â??Il rischio piÃ¹ grande non Ã¨ unâ??AI potente, Ã¨ lâ??asimmetria dellâ??AI potenteâ?•, ha detto. Lâ??asimmetria tra chi attacca e chi deve difendersi, tra le poche grandi aziende che controllano i modelli piÃ¹ avanzati e tutti gli altri, tra un ristretto numero di Paesi e il resto del mondo.

Delangue parte da un episodio che Hugging Face ha vissuto direttamente: il cyberattacco condotto la scorsa estate da agenti autonomi. Durante la risposta allâ??incidente, alcuni sistemi proprietari non potevano essere utilizzati efficacemente perchÃ© i loro meccanismi di sicurezza non erano in grado di distinguere le operazioni effettuate dai difensori da quelle di un aggressore. Hugging Face si Ã¨ quindi affidata anche a un modello aperto per reagire allâ??attacco.

â??Il mondo ha bisogno dellâ??Ai open source piÃ¹ che mai per difendersiâ?•. Anche lui chiede regole piÃ¹ severe sulla comunicazione degli incidenti e maggiore trasparenza, fino alla condivisione delle tracce lasciate dagli agenti autonomi. Ma avverte che una risposta costruita esclusivamente intorno al controllo dei modelli piÃ¹ potenti potrebbe avere un effetto collaterale: rafforzare ulteriormente proprio le poche aziende che possiedono capitale, chip e infrastrutture per svilupparli.

Su una cosa non si trova un punto comune, anche se la distanza tra i quattro si Ã¨ ridotta. Se sistemi molto piÃ¹ potenti di quelli attuali stanno davvero arrivando, chi stabilirÃ quali rischi sono accettabili, chi controllerÃ chi li costruisce e, soprattutto, chi avrÃ accesso a quella potenza. Non Ã¨ piÃ¹ una questione tecnologica, ma riguarda potere, sovranitÃ e sicurezza internazionale. E parla direttamente a Donald Trump e Xi Jinping, che a pochi chilometri da New York si incontrano per la prima volta in 11 anni sul suolo americano.

â??

internazionale/esteri

webinfo@adnkronos.com (Web Info)

Categoria

1. Comunicati

Tag

1. Ultimora

Data di creazione
Settembre 24, 2026
Autore
redazione