



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

Descrizione

(Adnkronos) 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 l'intelligenza artificiale moderna sono partiti da posizioni diverse per arrivare almeno a una conclusione comune: la corsa verso sistemi sempre più potenti non può essere considerata una forza incontrollabile della natura e la sicurezza non può essere sacrificata sull'altare della competizione.

Poi, però, cominciano le differenze. Su chi debba stabilire le regole, sul potere da lasciare alle aziende, su quanto debba intervenire la politica e persino sull'idea che sistemi molto potenti siano di per sé il problema. A portare quest'ultima tesi al Palazzo di Vetro è stato Clement Delangue, il meno noto al grande pubblico dei quattro relatori ma forse quello che ha introdotto l'elemento di maggiore discontinuità: il fondatore di Hugging Face, uno dei principali snodi mondiali dell'AI open source, ha avvertito che il rischio più grande non è avere intelligenze artificiali potenti, ma lasciare che quella potenza sia mal distribuita.

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

È soprattutto l'intervento di Yoshua Bengio a dare la misura di quanto sia cambiato il dibattito. Canadese, professore all'Università di Montréal, Bengio nel 2018 ha ricevuto insieme a Geoffrey Hinton (poi fuggito da Google) e Yann LeCun (da poco uscito da Meta) il Turing Award, il riconoscimento più prestigioso dell'informatica, per il lavoro fondamentale sul deep learning. È il fondatore di Mila, uno dei più importanti centri di ricerca sull'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 dall'origine per essere controllabili.

Bengio è diventato uno dei critici più severi del settore che ha contribuito a creare. Al Palazzo di Vetro non ha usato formule prudenti. «I pericoli sono reali e imminenti», 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 è soprattutto la logica della corsa. Le aziende ammettono che sistemi sempre più capaci possono comportare rischi molto seri ma spiegano contemporaneamente di non poter rallentare perché qualcun altro continuerebbe comunque ad accelerare. «La corsa non è una legge di natura», ha detto al Consiglio. È il risultato di decisioni prese dalle aziende.

Da qui una proposta che va oltre l'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 dell'addestramento e della messa a disposizione del pubblico, che gli incidenti siano obbligatoriamente segnalati e che venga introdotta una forma di assicurazione sulla responsabilità, come accade in altri settori ad alto rischio. Il paragone che utilizza è con medicina, aviazione ed energia nucleare.

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

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

Altman si infila a metà del dibattito tra ottimismo cieco e doomerism, l'idea che l'arrivo di sistemi superiori all'uomo conduca a conseguenze catastrofiche.

Ma il suo messaggio sulla sicurezza (qui il testo integrale) è stato meno sfumato. OpenAI, ha ricordato, ha già rallentato unilateralmente in passato e potrà farlo ancora. Non dovrebbe essere addestrato un modello, ha spiegato, se non esistono prove molto solide che possa rimanere sotto controllo umano. E una probabilità anche piccola di una catastrofe non può semplicemente essere considerata un costo accettabile dell'innovazione.

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

Anche Dario Amodei, cofondatore e amministratore delegato di Anthropic, la società che sviluppa Claude ed è nata proprio con una forte attenzione alla sicurezza dei modelli, promette che «rallenteremo quanto sarà necessario» 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 "nazione di geni racchiusa in un data center": milioni di ricercatori e specialisti (sintetici) capaci di lavorare a velocità 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 più 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 l'impiego dell'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 più gravi.

Poi c'è Clément Delangue. Francese, imprenditore, cofondatore e amministratore delegato di Hugging Face, società nata nel 2016 e diventata negli anni una sorta di GitHub dell'intelligenza artificiale: sulla sua piattaforma sviluppatori e ricercatori condividono modelli, dataset e strumenti. Hugging Face ha costruito buona parte della propria identità sull'open source e sull'idea che l'accesso all'AI non debba essere riservato ai pochi gruppi capaci di spendere miliardi nell'addestramento dei modelli più avanzati.

Per questo il suo intervento davanti al Consiglio di sicurezza parte da una preoccupazione differente. Il rischio più grande non è un'AI potente, è l'asimmetria dell'AI potente, ha detto. L'asimmetria tra chi attacca e chi deve difendersi, tra le poche grandi aziende che controllano i modelli più 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 all'incidente, alcuni sistemi proprietari non potevano essere utilizzati efficacemente perché i loro meccanismi di sicurezza non erano in grado di distinguere le operazioni effettuate dai difensori da quelle di un aggressore. Hugging Face si è quindi affidata anche a un modello aperto per reagire all'attacco.

Il mondo ha bisogno dell'AI open source che mai per difendersi. Anche lui chiede regole 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 più 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 è ridotta. Se sistemi molto più potenti di quelli attuali stanno davvero arrivando, chi stabilirà quali rischi sono accettabili, chi controllerà chi li costruisce e, soprattutto, chi avrà accesso a quella potenza. Non è una questione tecnologica, ma riguarda potere, sovranità 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.

?

internazionale/esteri

webinfo@adnkronos.com (Web Info)

Categoria

1. Comunicati

Tag

1. Ultimora

Data di creazione

Settembre 24, 2026

Autore

redazione

default watermark