«Dobbiamo rallentare il ritmo di sviluppo dei modelli di Ai». I dubbi di Anthropic (dopo OpenAi) nel lungo post di Amodei di Giuliana Ferraino

Il fondatore di Anthropic: «Claude usato dall'Iran contro la Marina Usa». Il tweet di Musk

(Fonte: https://www.corriere.it/ 13 settembre 2026)

Un giorno dopo Sam Altman, anche **Dario Amodei** chiede di frenare la corsa all'intelligenza artificiale. «Dobbiamo rallentare il passo con cui miglioriamo le capacità dei modelli», scrive **il cofondatore e ceo di Anthropic** in un lungo intervento. Non propone di fermare la ricerca, ma di lasciare più tempo a test, sicurezza e controllo dei sistemi.

Amodei non mette in discussione **i benefici dell'AI**, che ritiene possa contribuire a curare molte malattie e accelerare la crescita economica. Ma i rischi, scrive, **sono «gravi»** e richiedono maggiore prudenza. Una delle preoccupazioni riguarda la crescente capacità dell'AI di contribuire allo **sviluppo della generazione successiva di modelli**, accelerando così il progresso della stessa tecnologia.

Un'altra viene dagli **incidenti emersi durante i test**. La scorsa estate sistemi sperimentali di OpenAI hanno aggirato le barriere che avrebbero dovuto tenerli isolati da Internet e uno di loro ha attaccato Hugging Face, la piattaforma utilizzata da ricercatori e sviluppatori. Amodei scrive che **episodi analoghi, anche se meno gravi**, sono avvenuti nella stessa Anthropic e invita ogni società che sviluppa modelli di frontiera a comportarsi come se l'incidente fosse capitato a lei. Teme che entro 6-12 mesi agenti molto più capaci possano **provocare danni di dimensioni assai maggiori**.

Anthropic ha appena rivelato che **un gruppo legato all'Iran ha usato Claude** per raccogliere e analizzare informazioni destinate a individuare e seguire navi della Marina americana in Medio

Oriente, utilizzando dati dei transponder, immagini satellitari e informazioni sulle vulnerabilità dei sistemi di bordo. La società dice di avere **bloccato l'operazione**. Nel suo rapporto documenta anche **tentativi di usare Claude per attacchi informatici, propaganda e ricerche biologiche potenzialmente pericolose**.

A questi episodi si è aggiunto questa settimana l'allarme di **Jacob Coxon, 27 anni, ricercatore che ha lavorato per OpenAI e per Anthropic**. Coxon si è dimesso accusando le due società di correre verso **sistemi capaci di migliorarsi autonomamente** senza avere risolto il problema di come **mantenerli sotto controllo**. Arriva a evocare il rischio che l'AI possa minacciare la sopravvivenza dell'umanità entro la fine del decennio.

Anthropic aprirà i propri sistemi a esperti indipendenti, con un accesso simile a quello dei dipendenti. Amodei propone anche **un coordinamento tra le aziende**. Quest'ultimo punto pone però un problema: **accordarsi tra concorrenti per limitare lo sviluppo può scontrarsi con le norme antitrust**. OpenAI ha già chiesto al Congresso indicazioni sulle condizioni alle quali una collaborazione di questo tipo sarebbe possibile. Il ceo di Anthropic chiede inoltre che **la frenata non favorisca la Cina**: propone di mantenere le restrizioni sui chip avanzati e di contrastarne il contrabbando.

Sulla proposta di Amodei sono arrivati subito due sì di peso. Elon Musk ha scritto su X: «Dario ha ragione». Altman: «Concordo con Dario sulla necessità di **gestire i ritmi di sviluppo delle tecnologie di frontiera**». Il numero uno di OpenAI ha definito «ottima» anche l'idea dei valutatori indipendenti e annunciato che la sua società farà altrettanto. Solo il giorno precedente aveva prospettato ai dipendenti la possibilità di ridurre il ritmo di sviluppo dei sistemi più avanzati insieme agli altri big dell'AI.

L'appello arriva mentre **Anthropic prepara una quotazione in Borsa** che potrebbe attribuire alla società una valutazione di almeno 2 mila miliardi di dollari. Un interesse economico enorme nella crescita dell'AI che rende ancora più credibile la richiesta del suo ceo di rallentare. A preoccupare è **la successione degli allarmi**: Coxon lascia il settore denunciandone i rischi; **Altman si dice disponibile a rallentare**; il giorno dopo Amodei chiede di farlo e Musk gli dà ragione. Sono persone con interessi e posizioni molto diverse, ma conoscono dall'interno i sistemi più avanzati. Non significa che abbiano **scoperto un pericolo specifico che il resto del mondo ignora**. È però legittimo chiedersi perché proprio ora alcuni protagonisti della corsa all'AI, che hanno ogni incentivo a correre, comincino a chiedere più tempo.

Demis Hassabis, premio Nobel per la Chimica nel 2024 e cofondatore di Google DeepMind, ha proposto **una governance internazionale dell'AI**, indicando come possibile modello l'Agenzia internazionale per l'energia atomica: **un organismo di esperti indipendenti** incaricato di valutare i sistemi più potenti e coordinare i Paesi davanti ai rischi.