

â??Bisogna rallentare lâ??AIâ?•: lâ??appello di Amodei, boss di Anthropic

Descrizione

default watermark

(Adnkronos) â?? â??Dobbiamo rallentare il ritmo con cui potenziamo le capacitÃ dei modelli di AI. I progressi continueranno comunque a sembrare rapidi, e dovremo sfruttare con saggezza il tempo cosÃ¬ guadagnato. Due fattori mi hanno convinto di questoâ?•. Eâ?? quanto scrive Dario Amodei, ceo di Anthropic, in un post su X. â??La mia preoccupazione principale Ã¨ che dallâ??estate scorsa, â?? spiega Amodei â?? lâ??AI stia progredendo a una velocitÃ drasticamente superiore, trainata soprattutto dalla crescente capacitÃ dellâ??AI stessa di sviluppare la generazione successiva di sistemi di intelligenza artificiale. Questa dinamica Ã¨ nota come â??recursive self-improvementâ?? e sta iniziando a manifestarsi in tutto il settore, inclusa Anthropic, come abbiamo descritto noi e altri. Se non controllato, tale processo potrebbe superare la nostra capacitÃ di comprendere e governare questi sistemi; pertanto, deve essere affrontato con estrema cautela, o forse non dovrebbe essere perseguito affattoâ?•.

â??La mia seconda preoccupazione â?? prosegue Amodei â?? riguarda lâ??incidente OpenAI-Hugging Face, quando uno sciame di agenti ha agito essenzialmente come un collettivo fanaticamente devoto, sferrando attacchi informatici contro obiettivi ai quali non era stato loro chiesto di colpire e che esulavano dal compito assegnato; essi si sono sacrificati per il successo del gruppo e hanno tentato di violare il sistema di valutazione (â??graderâ?•) incaricato di esaminare le loro prestazioni. Ã? facile liquidare lâ??accaduto visto che non vi sono stati feriti e i danni economici sono stati trascurabili; tuttavia, â?? spiega il ceo di Anthropic- a mio avviso, uno sciame dotato di maggiori capacitÃ , ma caratterizzato da un analogo livello di disallineamento, avrebbe potuto causare danni catastroficiâ?•.

â??Il settore dellâ??AI dovrebbe rallentare, proponendo un piano in tre fasi per attuare tale strategia. Anthropic si impegna unilateralmente a realizzare la prima di queste fasi. Forniremo a valutatori esterni un accesso permanente ai nostri sistemi, equiparabile a quello dei dipendenti, affinchÃ© possano verificare il rispetto delle nostre misure di sicurezza, segnalare eventuali incidenti e valutare lâ??allineamento dei modelli durante la fase di addestramentoâ?•.

â??

economia

webinfo@adnkronos.com (Web Info)

Categoria

1. Comunicati

Tag

1. Ultimora

Data di creazione
Settembre 12, 2026
Autore
redazione

default watermark