



“Bisogna rallentare l’AI”: l’appello di Amodei, boss di Anthropic

Descrizione

(Adnkronos) “Dobbiamo rallentare il ritmo con cui potenziamo le capacità dei modelli di AI. I progressi continueranno comunque a sembrare rapidi, e dovremo sfruttare con saggezza il tempo così guadagnato. Due fattori mi hanno convinto di questo”. È quanto scrive Dario Amodei, ceo di Anthropic, in un post su X. “La mia preoccupazione principale è che dall’estate scorsa, spiega Amodei l’AI stia progredendo a una velocità drasticamente superiore, trainata soprattutto dalla crescente capacità dell’AI stessa di sviluppare la generazione successiva di sistemi di intelligenza artificiale. Questa dinamica è nota come “recursive self-improvement” 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 capacità di comprendere e governare questi sistemi; pertanto, deve essere affrontato con estrema cautela, o forse non dovrebbe essere perseguito affatto”.

“La mia seconda preoccupazione” prosegue Amodei “riguarda l’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 (“grader”) incaricato di esaminare le loro prestazioni. È facile liquidare l’accaduto visto che non vi sono stati feriti e i danni economici sono stati trascurabili; tuttavia, spiega il ceo di Anthropic- a mio avviso, uno sciame dotato di maggiori capacità, ma caratterizzato da un analogo livello di disallineamento, avrebbe potuto causare danni catastrofici”.

“Il settore dell’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, affinché possano verificare il rispetto delle nostre misure di sicurezza, segnalare eventuali incidenti e valutare l’allineamento dei modelli durante la fase di addestramento”.

“

economia

webinfo@adnkronos.com (Web Info)

Categoria

1. Comunicati

Tag

1. Ultimora

Data di creazione

Settembre 12, 2026

Autore

redazione

default watermark