Consiglio Nazionale delle Ricerche

il venerdì di Repubblica
Dir. Resp.:Aligi Pontani
Tiratura: 193.707 Diffusione: 202.251 Lettori: 851.000

Rassegna del: 20/12/24
Edizione del:20/12/24
Estratto da pag.:56
Foglio:1/1

Molto fragili, come noi

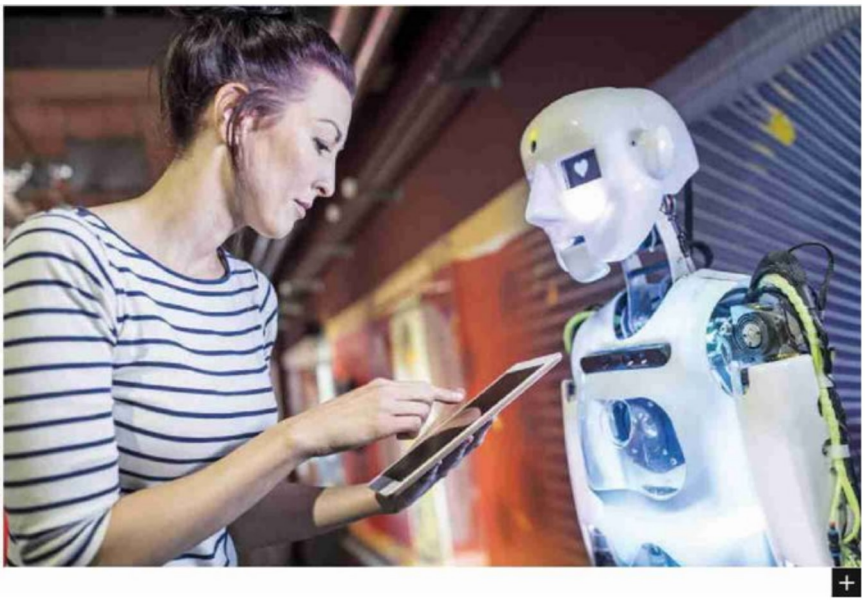
CHAT-GPT ABBOCCA ALLE MAIL TRUFFA, MENTRE GLI ALGORITMI USATI NEI TRIBUNALI USA SI RIVELANO PIENI DI PREGIUDIZI. NUOVI STUDI MOSTRANO CHE (ANCHE) LE **INTELLIGENZE ARTIFICIALI** SONO TUTT'ALTRO CHE INFALLIBILI

di **Alex Saragosa**

SISA che le intelligenze artificiali soffrono di "allucinazione", danno cioè risposte inventate quando non conoscono quelle giuste. Recenti ricerche stanno rivelando altre loro debolezze. La psicologa Udari Sehwag, della JP Morgan, ha creato 37 diverse email truffaldine con richieste, tipo "attiva la carta di credito a questo link", chiedendo a due versioni di Chat-GPT e a Llama, Ia di Meta, se dovremmo fidarci di tali offerte. Il risultato è che Chat-GPT 3 e 4 ci cascavano il 22 e il 9 per cento delle volte, mentre Llama abboccava 3 volte su cento: sono tutte insomma molto credulone. Almeno quanto gli umani, che aprono le mail di *phishing* al 17,2 per cento (studio 2021). Kosuke Imai, giurista di Harvard, ha invece valutato l'accuratezza dei sistemi Ia usati nei tribunali Usa per valutare il rischio di violazione della libertà provvisoria. Confrontando le valutazioni date da giudici che avevano usato o meno l'Ia, Imai ha concluso che l'algoritmo era troppo severo, soprattutto verso i non bianchi, raccomandando pesanti cauzioni che causavano problemi finanziari o inutili detenzioni a persone povere ma affidabili.

Non si meraviglia troppo l'informatica del Cnr Sara Colantonio: «Queste Ia progettate per imitare il linguaggio umano sembrano esseri senzienti. Ma sono solo sistemi statistici che individuano strutture all'interno di enormi database di testi o immagini. In ambiti ristretti, come trovare tumori in radiografie, possono essere utili, ma nel complesso e caotico mondo reale sbagliano spesso, perché non hanno idea di cosa sia "il mondo", e quindi, per esempio, cosa sia una truffa, sia perché i dati con cui vengono istruite sono spesso incompleti e pieni di informazioni distorte o false (come quelle che portano ai pregiudizi mostrati nei tribunali Usa). Se non risolviamo il problema della qualità dell'"istruzione", sarà bene non usarle per problemi delicati, come quelli giudiziari o lavorativi».

E come se le debolezze intrinseche delle Ia non fossero abbastanza, una bizzarra ricerca ne ha aggiunta un'altra: la paura. Lo psicologo del Max Planck Institute, Iyad Rahwan, con il progetto Spook the Machine ha creato due algoritmi fifoni, uno che teme di perdere la memoria e l'altro di diventare obsoleto, invitando tutti a inviare immagini che li terrorizzino. Il 7 gennaio 2025 le più efficaci verranno pubblicate e premiate, così imparano le Ia a spaventare noi... ◘



GETTY IMAGES/WESTEND61

Le Intelligenze artificiali sono istruite per imitare il linguaggio umano. A destra, **Sara Colantonio**, informatica del Cnr

Peso:80%