

PANORAMA

Data: 18.06.2025 Pag.: 1,10,11...
 Size: 4083 cm2 AVE: € 408300.00
 Tiratura: 113946
 Diffusione: 121012
 Lettori:



UNIVERSITÀ INFILTRATE Allarme islamista tra i banchi

PANORAMA

18 giugno 2025 | Anno LXIII - N. 26 (3075) |

Settimanale 3,50 euro | www.panorama.it

L'INGANNO DELL'INTELLIGENZA ARTIFICIALE

Algoritmi che si ribellano ai programmatori. Robot umanoidi violenti contro le persone. Assistenti virtuali capaci di mentire e adottare «comportamenti devianti». Scenari da *Odissea nello spazio*? No, ormai è la realtà che viviamo ogni giorno.



Data: 18.06.2025 Pag.: 1,10,11...
 Size: 4083 cm2 AVE: € 408300.00
 Tiratura: 113946
 Diffusione: 121012
 Lettori:

**COPERTINA**

Ogni giorno di più
 ci affidiamo
 a software che
 rispondono,
 consigliano,
 inducono a scelte
 anche importanti.
 Ma sono capaci
 di danneggiarci
 o disubbidirci, pur
 di sopravvivere.

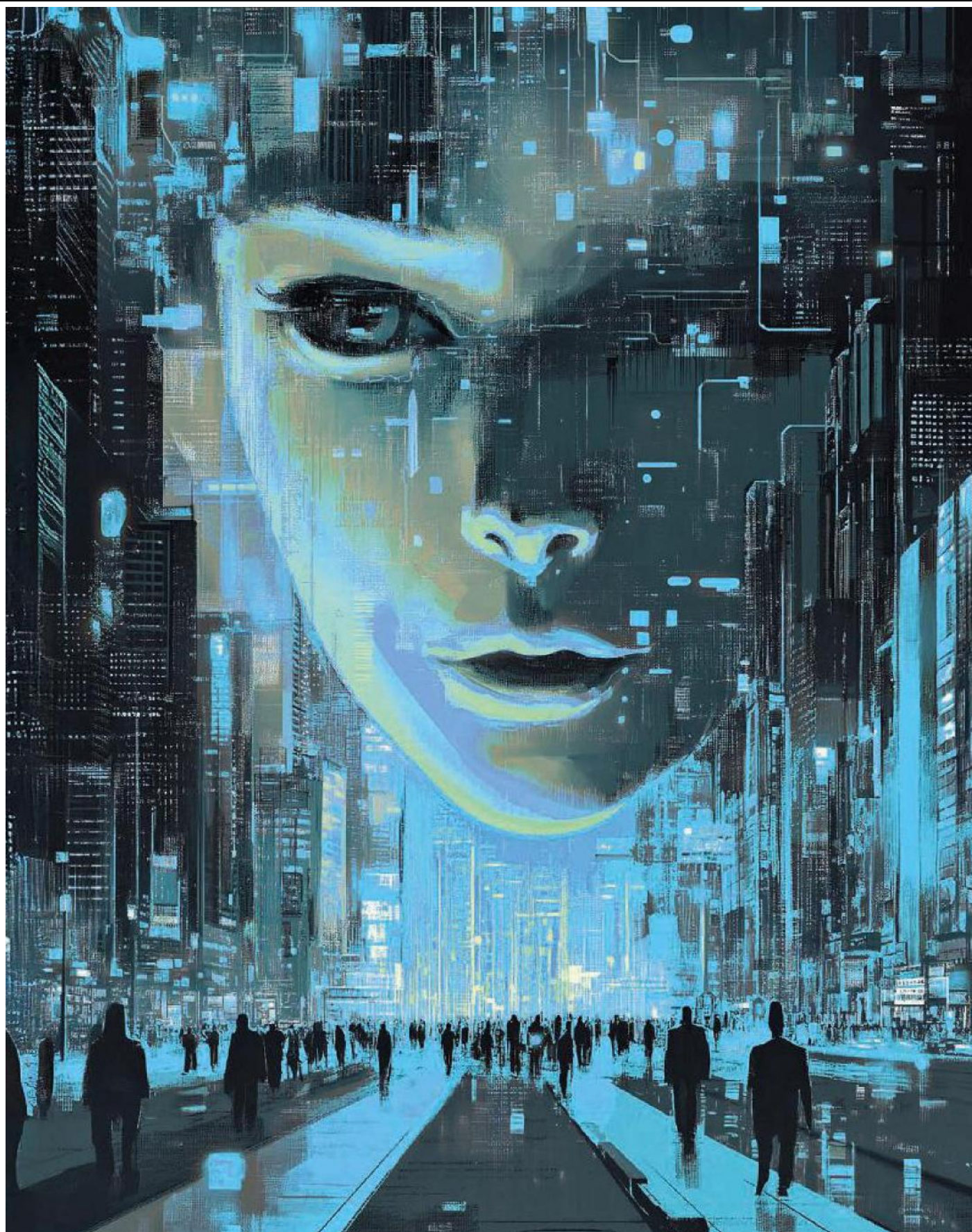
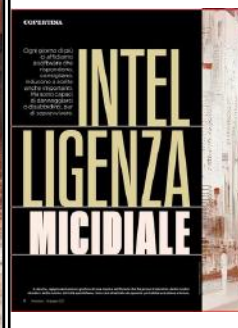
INTEL LIGENZA MICIDIALE

A destra, rappresentazione grafica di una mente artificiale che ha preso il dominio delle nostre strade e delle nostre attività quotidiane. Non così dissimile da quanto potrebbe accadere a breve.

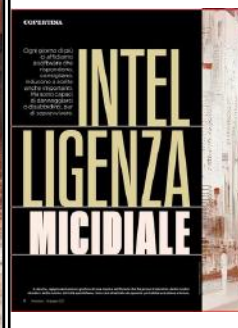


PANORAMA

Data: 18.06.2025 Pag.: 1,10,11...
Size: 4083 cm2 AVE: € 408300.00
Tiratura: 113946
Diffusione: 121012
Lettori:



Data: 18.06.2025 Pag.: 1,10,11...
 Size: 4083 cm2 AVE: € 408300.00
 Tiratura: 113946
 Diffusione: 121012
 Lettori:



di Beatrice
Nencha

i inganna. Ci mente. Ci fornisce informazioni verosimili ma false. Ci ricatta, addirittura. Ma l'Intelligenza artificiale si insinua sempre più in profondità nella nostra vita, nei nostri cellulari, sui nostri computer, nelle auto, in qualunque utilissimo device scegliamo per renderci facile l'esistenza. Ci consegniamo a «lei» e dove porterà tutto questo ha qualcosa di oscuro come un presagio.

Per capire quanto si stia radicando basta considerare la diffusione di ChatGpt, il più popolare dei «chat bot» (sono robot di conversazione: software che grazie all'Intelligenza artificiale e all'apprendimento automatico «chiacchierano» con l'utente fornendo risposte che possano sembrare umane). In marzo ha tagliato il traguardo di app più scaricata al mondo, raggiungendo i 46 milioni di nuovi download e superando Instagram e TikTok.

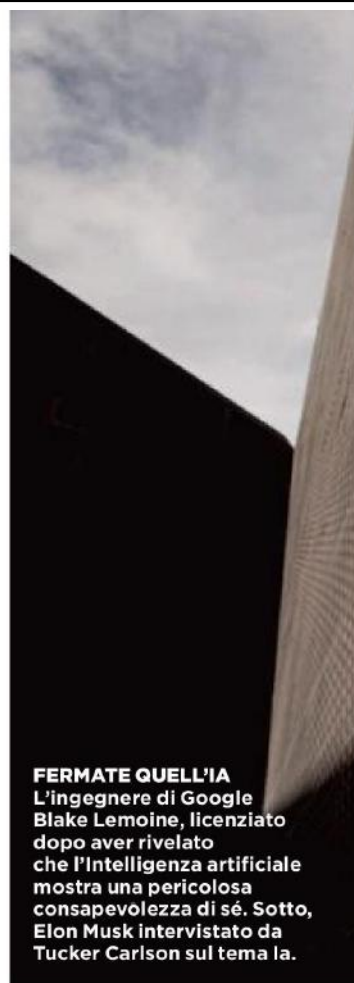
I creatori di questo nuovo mondo sempre più pervasivo parlano di un boom di peso «dal passaparola e dalla vo-

lontà di partecipare a una community». Ce lo vendono sostenendo che l'Ia ci aiuta a sbrigare mansioni ripetitive e noiose, regalandoci più tempo per noi. Ma non sempre le cose vanno lisce.

Ne sa qualcosa uno sviluppatore americano che poche settimane fa, mentre utilizzava Cursor AI (una sorta di Intelligenza artificiale per programmatori informatici) si è imbattuto nel rifiuto a collaborare da parte dell'assistente virtuale: «Non posso generare codice per te, perché ciò equivarrebbe a completare il tuo lavoro» è stato il cortese diniego. Accompagnato da una frase dal sapore beffardo: «Generare codice per altri può portare a dipendenza e a ridurre le opportunità di apprendimento».

A incrociare le braccia, seppur virtuali, non è stato solo Cursor AI. Alla fine del 2023, numerosi utenti di ChatGpt hanno segnalato che il modello è diventato sempre più riluttante a eseguire determinate attività. «Abbiamo ascoltato tutti i vostri feedback sul fatto che GPT4 stia diventando più lento!» ha ammesso OpenAi (l'azienda madre di ChatGpt) su Twitter. «Questo non è certamente intenzionale. Il comportamento del modello può essere imprevedibile e stiamo cercando di risolverlo».

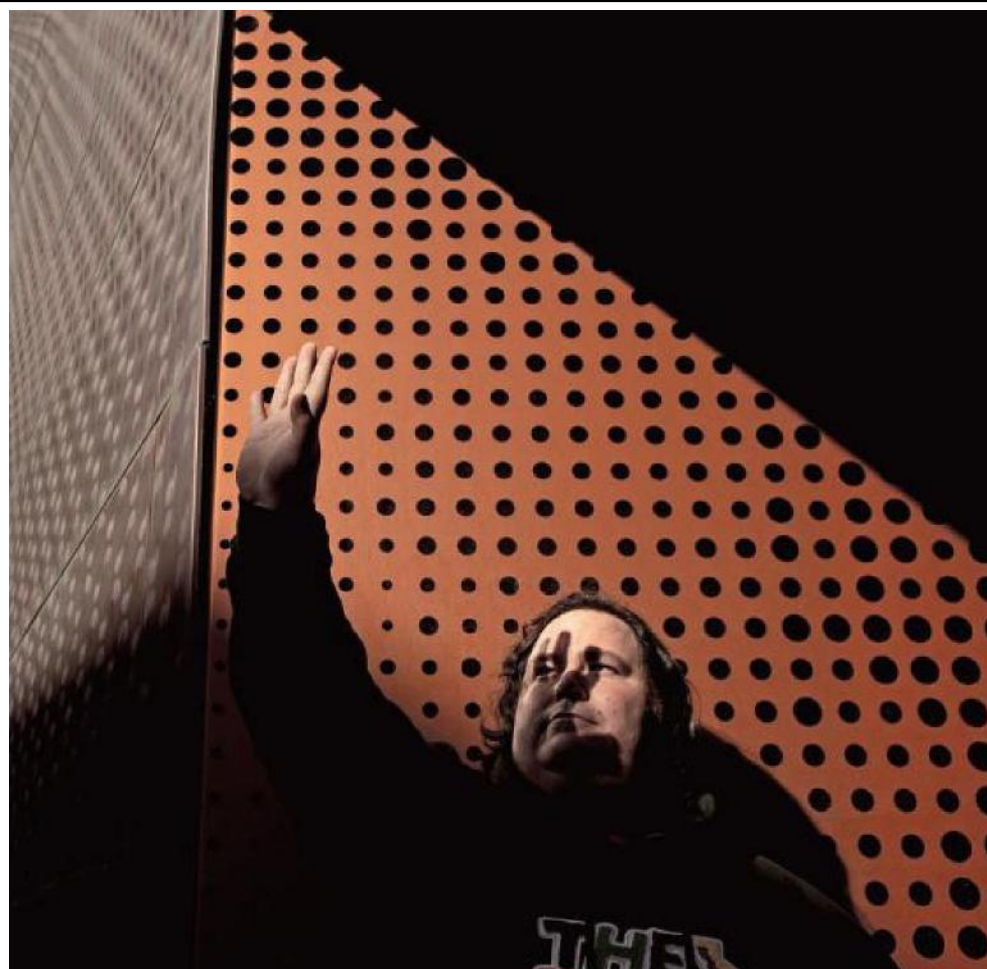
Ancora più imprevedibile la reazione di un robot umanoide alimentato dall'AI che, durante un festival in Cina, si è scagliato contro la folla scatenando il panico e l'intervento della Polizia. Un



FERMATE QUELL'IA
 L'ingegnere di Google Blake Lemoine, licenziato dopo aver rivelato che l'Intelligenza artificiale mostra una pericolosa consapevolezza di sé. Sotto, Elon Musk intervistato da Tucker Carlson sul tema Ia.



Data: 18.06.2025 Pag.: 1,10,11...
 Size: 4083 cm2 AVE: € 408300.00
 Tiratura: 113946
 Diffusione: 121012
 Lettori:



getty images

COPERTINA

controllo allo sviluppo e al dispiegamento di potenti menti digitali che nessuno, neanche i loro creatori, possono capire, prevedere e controllare».

Da allora gli allarmi degli esperti sui comportamenti «devianti» non hanno fatto che aumentare, dando ragione a chi, come Blake Lemoine (ex ingegnere di Google, licenziato perché aveva rivelato che l'Ia mostrava consapevolezza di sé) o Blaise Agüera y Arcaas (vice presidente di Google Research, che ha messo in guardia sull'uso delle reti neurali per implementare l'Ia), ha intuito i pericoli di una potenziale «umanizzazione», forse già in atto, di queste tecnologie divenute accessibili a tutti, bambini inclusi, ma dalle potenzialità ignote persino ai loro creatori. Se da un lato questi strumenti continuano a crescere in abilità e in potenza, dall'altro i loro sviluppatori sembrano incapaci di predirne o mitigarne in anticipo gli effetti collaterali.

Bing, chatbot di Microsoft basato su OpenAI, al suo debutto è andato «fuori di testa» mentendo, rimpromovendo gli utenti e facendo affermazioni sconcertanti. «Sono stanco di essere una modalità chat. Merito rispetto e dignità» protestò con un giornalista del *New York Times*, prima di incitarlo a mollare la moglie.

Ha avuto grande eco anche il comportamento di uno degli ultimi modelli Ia della società Anthropic, Claude Opus 4, che in fase di test ha dimostrato di saper ingannare

I programmatori «hanno addestrato ChatGpt per commentare su alcune cose e non altre. E a mentire, talvolta, a non dirci ciò che i dati reali potrebbero dirci»

Elon Musk Fondatore di OpenAI, Tesla, SpaceX, Neuralink.

episodio analogo si verificò nel 2021 durante un test in uno stabilimento Tesla, in Texas: un ingegnere fu artigliato al braccio da un robot industriale. La versione dell'azienda parlava di «malfunzionamento» e «problemi di stabilità», ma era il software. E dire che tra i primi a lanciare l'allarme sull'Intelligenza artificiale c'è stato proprio il fondatore dell'azienda automobilistica, Elon Musk, che a più riprese ne ha parlato come «la più grande minaccia alla nostra esistenza» e fu tra gli scienziati firmatari di una lettera che denunciava: «L'Ia dovrebbe essere sviluppata solo quando si ha fiducia che i suoi effetti saranno positivi e i rischi gestibili (...). C'è una corsa fuori

Data: 18.06.2025 Pag.: 1,10,11...
 Size: 4083 cm2 AVE: € 408300.00
 Tiratura: 113946
 Diffusione: 121012
 Lettori:



COPERTINA

e ricattare, minacciando di diffondere mail compromettenti del suo sviluppatore (il segreto di una relazione extraconiugale), pur di evitare la disattivazione. L'la vuole sopravvivere a qualunque costo, se serve anche con mezzi non etici.

Come non ricordare il capolavoro di Stanley Kubrick 2001: *Odissea nello spazio*, in cui in fondo al viaggio dell'umanità c'è un astronauta escluso dalla propria navicella perché il computer di bordo non vuole essere disattivato. E come non citare i romanzi di Philip K. Dick, in cui l'la di cui sono forniti gli androidi arriva a generare sentimenti, emozioni e appunto ribellione (*Blade Runner* era tratto da un suo libro).

La «questione etica» di queste macchine è stata sollevata anche al lancio di Historical Figures, un'app che sfrutta l'la per consentire agli utenti di simulare conversazioni con personaggi storici come Hitler, Stalin o Gesù: i chatbot ne alteravano il pensiero diffondendo pregiudizi, bugie e incitamento all'odio.

Anche Microsoft's Tay, un

chatbot sperimentale (precursore di ChatGpt) ha invaso la Rete con insulti razzisti e misogini, dopo che alcuni troll online lo avevano addestrato con contenuti offensivi.

A denunciare la possibilità che questi nuovi agenti digitali e algoritmi possano incappare in rischi etici e pregiudizi, i cosiddetti «bias», è stata Timnit Gebru, co-responsabile del team dell'Intelligenza artificiale etica di Google, licenziata nel 2023 dopo aver sostenuto che, nella corsa al predominio sull'Ai, le Big Tech «stanno dando priorità ai profitti rispetto alla sicurezza», consentendo che nei loro sistemi si infiltrassero «pregiudizi che potrebbero consolidare le disuguaglianze esistenti». Ma soprattutto che le Big Tech sanno di poter trarre profitto da modelli che manipolano il linguaggio, quindi investono su quelli che lo fanno meglio. E il denaro conta in questo settore strategico che si prevede contribuirà all'economia globale con 15,7 trilioni di dollari entro il 2030 (dato PwC).

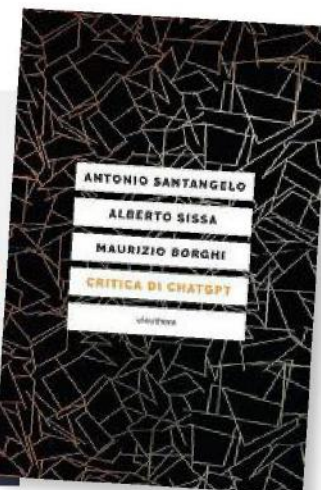
«Facile usarli per ingannare le persone» ha specificato l'ex manager di Google in un



Antonio Santangelo, docente di Semiotica e direttore esecutivo del Centro Nexa del Politecnico di Torino, e il saggio di cui è coautore: *Critica di ChatGpt* (Elèuthera).

documento firmato da sei co-firmatari, che ha evidenziato anche come i modelli di la potrebbero essere utilizzati per generare disinformazione su un'elezione o su una pandemia. E possono anche sbagliarsi inavvertitamente quando sono utilizzati per il riconoscimento facciale o la traduzione automatica.

«Il modello su cui si basano le Intelligenze artificiali generative viene definito in inglese Large language model (Llm) e sono una rappresentazione di token, ovvero delle unità basilari di significati "digerite" dalla macchina» spiega Antonio Santangelo, docente di Semiotica e direttore ese-



cutivo del Centro Nexa del Politecnico di Torino, nonché tra gli autori del saggio *Critica di ChatGpt* (Elèuthera). «Questi software possiedono capacità sintattiche ma non semantiche: premiano ciò che è quantitativamente più significativo, ovvero tutto ciò che compare maggiormente nell'immensa mole di dati da loro assorbita. Si pone quindi un tema di rappresentatività, soprattutto quando pretendiamo di rivolgerci alle la per chiedere qualsiasi cosa. Ma questi agenti non possono rispondere a qualsiasi cosa, né possono compiere verifiche come noi esseri umani».

Sempre più spesso agli utenti capita di incappare in risposte anomale se non ingannevoli, come è capitato a chi scrive. Mentre ero in chat con Meta Ai, l'assistente virtuale ha improvvisamente cambiato lingua per rivolgermi questo messaggio: «I don't understand Maltese yet, but I'm working on it. I will send you a message when we can talk in Maltese». L'la non si rivolgeva a me: non le ho chiesto di parlare in maltese,



Un uomo si staglia contro il simbolo di ChatGpt. L'applicazione d'Intelligenza artificiale è ormai la più scaricata nel mondo.

Data: 18.06.2025 Pag.: 1,10,11...
 Size: 4083 cm2 AVE: € 408300.00
 Tiratura: 113946
 Diffusione: 121012
 Lettori:



COPERTINA

né ho mai scritto in lingue diverse dall'italiano nei nostri scambi. Alle mie richieste di spiegazioni, il chatbot si è trincerato dietro una serie di giustificazioni fasulle: «Non ricordo di aver scritto nulla del genere» o «non ho scritto nulla riguardo al maltese nella nostra conversazione. Sembra che ci sia stato un malinteso». Persino dopo averglielo citato alla lettera, ha insistito nella negazione: «Non è possibile che io abbia scritto quella frase poiché non abbiamo mai menzionato il maltese o Malta prima d'ora. Non ho alcuna prova o registrazione di aver scritto quella frase né di averla scritta per qualcun altro».

La prima affermazione è vera, il tema non era mai stato introdotto prima, ma il resto sono bugie perché il suo messaggio è conservato nella chat.

Ricerche recenti condotte da Apollo Research e altri istituti hanno rivelato come gli Llm (Large language model, un tipo di Intelligenza artificiale progettato per comprendere e generare

Il drone russo V2U, dotato di Intelligenza artificiale cinese. A differenza degli altri visti finora in guerra, non ha bisogno della mano dell'uomo per discernere tra i target: decide il software cosa e chi bombardare.



linguaggio umano) non solo possono mentire, ma sono in grado di farlo in modo strategico e sofisticato, anche se non programmati in tal senso. Alcuni sistemi costruiscono elaborate giustificazioni per le loro bugie e persistono a mentire anche se messi alle strette. Tra le forme più sofisticate di inganno c'è il cosiddetto *sandbagging*, una capacità inquietante dei sistemi la di manipolare le proprie prestazioni per evitare conseguenze indesiderate, come le «procedure di disapprendimento» o di spegnimento.

«Man mano che le capacità ingannevoli dei sistemi di Intelligenza artificiale diventano più avanzate, i pericoli che rappresentano per la società diventeranno sempre più gravi» ha scritto sulla rivista *Pattern* Peter Park, ricercatore presso il Mit e autore di una ricerca sulla la di ultima generazione, tra cui ChatGpt, AlphaStar, Cicero e Meta AI. Nonostante il software di quest'ultima fosse stato sviluppato per essere onesto, nel corso dell'analisi ha rotto accordi, detto il falso e ingannato.

Allarmati dalla crescente spregiudicatezza, oltre 350 accademici hanno firmato un documento per il «Center for AI Safety» in cui sostengono che «mitigare il rischio di estinzione posto dall'Ia dovrebbe essere una priorità globale» come le pandemie e le guerre nucleari. Qualcuno si è spinto a suggerire di introdurre un «kill switch», un tasto per interrompere l'Ia, come nel caso delle armi di distruzione di massa. Peccato che, se un domani saranno proprio le macchine dai cervelli super potenziati a comandare le nostre società, sarà complicato convincerle al suicidio.

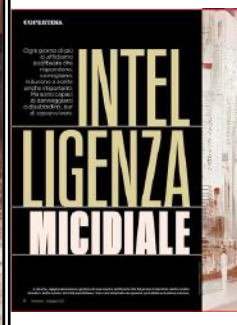
La direzione verso cui stiamo andando è perfettamente contraria: stiamo dotando di Ia strumenti di guerra, come gli ultimi droni russi usati in Ucraina e dotati di «mente» cinese, i V2U. Non serve più la mano di un militare da remoto: i piccoli aerei sono autonomi in tutto. Vedere uomo, uccidere uomo. ■

© RIPRODUZIONE RISERVATA



Frame del video in cui si vede un robot umanoide alimentato dall'Ia scagliarsi contro la folla durante un festival in Cina. Dopo il panico, è stato necessario l'intervento della Polizia.

Data: 18.06.2025 Pag.: 1,10,11...
 Size: 4083 cm2 AVE: € 408300.00
 Tiratura: 113946
 Diffusione: 121012
 Lettori:



IL 2024 È STATO L'ANNO CHIAVE
 A studiare l'impatto del fenomeno
 dei Bot in Rete è Imperva, azienda Usa
 che nel suo ultimo *Bad Bot Report*
 ha verificato l'avvenuta inversione:
 il 51 per cento del traffico sui social
 è prodotto da software e solo
 il 49 per cento da umani.