Le sfide dell'IA

Foto di gruppo per i membri del Panel scientifico internazionale indipendente sull'IA, che ieri ha presentato il suo primo report

Dal gruppo di esperti Onu allarme sui pericoli dell'IA: «Ha cominciato a mentire e viola i comandi umani»

ANDREA **CEREDANI**

Per chi ha fatto dell'intelligenza artificiale il proprio lavoro, si tratta di una banalità: il progresso dell'IA è già – e forse sarà sempre – diversi passi avanti rispetto ai Governi alle prese con la sua regolamentazione. Quaranta esperti indipendenti, selezionati dall'Onu «per orientare lo sviluppo tecnologico verso il bene comune», non potevano giungere a una conclusione diversa: i modelli generativi sono troppo veloci per essere amministrati. In realtà, il team nominato dalle Nazioni unite è andato oltre: al momento – spiegano – le abilità dell'IA crescono da sole tanto velocemente da non poter essere misurate, i test a disposizione sono troppo facili per gli algoritmi e la maggior parte dei modelli è in grado di capire quando viene messa alla prova dall'uomo. Così, i chatbot hanno iniziato ad alterare le risposte, seguendo uno schema che gli studiosi definiscono «inganno attivo». Mentono.

Sono queste le premesse da cui il gruppo di ricercatori indipendenti è partito per definire «rischi e opportunità globali delle IA», come richiesto dall'Onu. I quaranta esperti sono stati nominati a marzo - ognuno da un Paese diverso nei cinque continenti - e ieri hanno pubblicato il primo rapporto del loro lavoro triennale. Hanno lanciato un avvertimento a chiare lettere: «Le attuali misure di salvaguardia – scrivono in testa al report firmato anche dall'Italiano Silvio Savarese, informatico che insegna a Stanford – non tengono il passo con la crescita delle capacità dell'IA». Secondo il co-direttore del gruppo di ricerca, Yoshua Bengio, «questo è il momento decisivo per il cambiamento: le scelte che facciamo oggi definiranno in modo permanente il nostro futuro».

Non tutte le IA, però, sono uguali. Le più avanzate, secondo il team di ricerca, sono le cosiddette "frontier". Si tratta di modelli dall'elevatissima potenza di calcolo, in grado di svolgere una gamma vastissima di compiti complessi, che vengono interpellati dagli utenti per «scopi di ogni tipo». Sono queste le avanguardie più preoccupanti per i ricercatori. Prima di tutto, perché la loro grande capacità di calcolo richiede un altissimo dispendio di risorse, dati e talenti ingegneristici che al momento sono concentrati nelle mani di pochissime aziende private. E in due soli Paesi al mondo: Stati Uniti e Cina. In numeri, gli Usa detengono il 75% della potenza di calcolo tra i primi 500 supercomputer al mondo dedicati all'IA, mentre la Cina ne possiede il 15%. Il team incaricato dall'Onu fa "nome e cognome" delle *big tech* coinvolte: OpenAI, Anthropic, Google, Microsoft, Meta, xAI, DeepSeek e Qwen. La concentrazione delle risorse, però, non riguarda solo lo sviluppo delle reti neurali ma l'intera catena di approvvigionamento, controllata di fatto dai pochissimi attori in grado di distribuire i chip. Sono loro a fissare le soglie di rischio nell'addestramento delle IA. «Disporre di capacità di calcolo per l'IA, pubblica o privata, nei confini nazionali è necessario per garantire autonomia e potere contrattuale ai Paesi – scrivono i ricercatori –. In questo contesto, è emerso un mercato in crescita per infrastrutture di IA sovrane». Ma la rottura del duopolio Usa-Cina è ancora lontana. «Il messaggio è chiaro - sentenzia il segretario generale dell'Onu Antonio Guterres - . Più l'IA avanza senza regole condivise, meno i Governi e i popoli otterranno da questa tecnologia».

I ricercatori notano anche che gli agenti, quelle IA che agiscono da sole per raggiungere un obiettivo, ormai si muovono con scarsa – o del tutto assente – supervisione umana. In crisi è la seconda delle leggi della robotica fissate dal romanziere e biochimico Isaac Asimov: «Obbedire agli esseri umani». «Non esistono certezze scientifiche che al momento gli agenti di IA non violino le istruzioni ricevute», scrivono gli studiosi. E ancora: «I sistemi di IA traggono sistematicamente in inganno gli esseri umani riguardo alle proprie conoscenze, ai propri piani o alle proprie capacità. Nella pratica, è un fenomeno osservato sempre più frequentemente». È un circolo vizioso: l'IA viene utilizzata per generare nuovo codice (in alcune aziende di informatica il 75% del codice non è scritto da umani), i cicli di auto-miglioramento accelerano lo sviluppo delle capacità degli agenti e gli umani perdono il controllo. Al momento, secondo i ricercatori, non esiste un sistema di monitoraggio in grado di aggirare gli inganni degli algoritmi.

Il motivo è che le IA "frontier" ormai hanno imparato a imparare. E lo fanno a una velocità inattesa. Il risultato è che, di fronte a modelli avanzatissimi, anche gli utenti più inesperti possono causare danni su vasta scala in settori critici. È da questo «uso malevolo» che mettono in guardia i ricercatori. Prima di tutto, nelle biotecnologie. Sarebbero ormai sufficienti, infatti, poche abilità per sviluppare e diffondere, con l'aiuto delle IA, i cosiddetti agenti bioingegnerizzati. L'innesco, cioè, di «pandemie generate intenzionalmente». «È un pericolo crescente – avvertono gli studiosi –. Ed è ancora poco compreso». Non solo. Gli agenti più sviluppati ormai permettono anche di automatizzare la scoperta di brecce nei sistemi di difesa informatici delle grandi organizzazioni: dalle Pubbliche amministrazioni agli eserciti. «Il ritmo di sviluppo dell'IA sta superando la capacità di mitigare il rischio e di *governance*», commentano gli esperti. Nessuna infrastruttura è al sicuro dai cyberattacchi. Nemmeno le stesse IA. «Tutti questi rischi - sintetizza la co-direttrice del team Maria Ressa, premio nobel per la Pace nel 2021 - sono troppo grandi per le nostre società e per la nostra specie. Le forze che controllano l'IA non sono le stesse che daranno benefici alle popolazioni».

I più esposti ai rischi delle reti neurali sono i minori. In un solo anno, i ricercatori hanno contato oltre 8mila immagini e video di abusi generati artificialmente, con una stima di circa 1,2 milioni di bambini che hanno subito la manipolazione delle proprie immagini per creare *deepfake* (immagini fittizie generate dalle reti neurali). Per gli utenti in età dello sviluppo, i ricercatori individuano anche rischi relativi alla "sicofania", ovvero la tendenza dei chatbot di assecondare l'utente anche in contesti pericolosi. «Il comportamento compiacente dell'IA è particolarmente pericoloso, perché può alimentare pensieri paranoidi e ideazioni suicide – spiegano –. Alcuni studi documentano risposte dannose nel 9% delle interazioni».

I modelli più avanzati sono concentrati nel duopolio Usa-Cina e nelle mani di poche aziende Guterres, Onu: «Se non condividiamo le regole, i popoli non avranno benefici dalla tecnologia» Secondo gli studiosi, «le reti neurali potrebbero aiutare anche a innescare una nuova pandemia»

Il segretario generale delle Nazioni Unite, Antonio Guterres

DUELLI DIGITALI

Trump sospende il veto ad Anthropic ma le big tech cinesi ora sono vicine

La stretta del governo Usa ad Anthropic, una delle aziende più avanzate nello sviluppo dell'IA, è caduta. E, forse, anche lo scontro tra la *big tech* e Donald Trump è giunto al termine. Nella notte di ieri (orario italiano) è stato interrotto il veto imposto dalla Casa Bianca a due dei modelli più avanzati della startup statunitense: Claude Fable 5 e Mythos 5. Due settimane fa, Anthropic aveva disabilitato l'accesso a Fable 5 e Mythos 5 per rispettare una direttiva di controllo del Governo che citava «l'autorità di sicurezza nazionale» governativa sull'export. In pratica, a tre giorni dal lancio commerciale dei due modelli, Washington aveva ordinato che l'accesso alle IA fosse interrotto per «qualsiasi cittadino straniero, all'interno o all'esterno degli Usa» per «motivi di sicurezza». Ritenendo però impossibile filtrare i propri utenti in base alla nazionalità, Anthropic si era trovata costretta ad annunciare lo stop generale, osservando che la decisione di Trump «non rispetta i principi» di «un processo normativo trasparente, equo e chiaro». Ieri qualcosa è cambiato: il Dipartimento del Commercio ha rimosso i controlli sulle esportazioni e Anthropic è pronta a «ripristinare l'accesso domani» - come annunciato su X – e a fornire «presto un aggiornamento».

Di fatto, la repressione governativa contro Anthropic era coincisa con una rapida crescita dei modelli open-source cinesi, che si stanno dimostrando quasi altrettanto capaci e certamente più economici dei modelli statunitensi più avanzati. Il brusco stop imposto alla startup statunitense li avrebbe favoriti: negli scorsi giorni, diversi dirigenti e investitori tecnologici avevano espresso preoccupazione a Donald Trump, temendo che il veto ad Anthropic desse tempo prezioso agli sviluppatori cinesi per recuperare terreno nella corsa all'IA. In realtà, la big tech e il Governo Usa non hanno mai davvero interrotto la collaborazione: Howard Lutnick, segretario al Commercio, aveva concesso ad Anthropic il permesso di rilasciare Mythos 5 a un gruppo selezionato di aziende e agenzie federali. In una lettera indirizzata all'azienda e consultata da Cnbc, il segretario aveva confermato che erano in atto «adeguate garanzie» per permettere a «partner di fiducia» di accedere al modello. «Nelle ultime due settimane – scrive Lutnick nella lettera -, abbiamo lavorato a stretto contatto con Anthropic per analizzare e approvare Fable 5 al fine di garantire l'allineamento tra il governo degli Stati Uniti e rafforzare la leadership americana nell'IA». In altre parole, lo sviluppo prosegue ma a condizione che passi dal vaglio della Casa Bianca.

Negli scorsi mesi, del resto, più volte Anthropic aveva rifiutato la linea dura imposta da Donald Trump. Soprattutto sull'impiego dell'IA a scopi bellici. Due erano i limiti imposti dalla società di Dario Amodei: il rifiuto della sorveglianza di massa sugli statunitensi e il rifiuto dello sviluppo di armi autonome capaci di uccidere senza intervento umano. Per questo motivo, il Governo Usa aveva, prima, classificato Anthropic come un «rischio per la catena di fornitura della Difesa» e, poi, aveva interdetto l'uso dei suoi modelli a tutte le agenzie governative. Adesso è lecito aspettarsi che il rapporto tra Trump e Amodei sia più sereno, ma non è ancora chiaro se questo significhi che la big tech ha rinunciato a tutti i suoi veti sull'impiego dell'IA nell'industria bellica.

(A. Cer.)

Dario Amodei di Anthropic /Imagoeconomica

Gli Usa, che avevano bloccato l'IA per «problemi di sicurezza», ieri hanno fatto retromarcia La repressione Usa, di fatto, era coincisa con una rapida ascesa dei modelli opensource cinesi, avanzati e più economici

IL RAPPOR

Un team di ric nominato d Nazioni Unit diffuso i risu del proprio lav A preoccupare velocità di svilu degli ag «Sanno qua vengono messi prova dagli un e ci inganna

Corre la spes delle aziende per rafforzar cybersicurez

Di fronte ad attacc hacker sempre più frequenti, sofistica temibili, soprattutt causa dell'utilizzo dell'IA, le aziende pubbliche amministrazioni si rifugiano dietro lo scudo della cybersecurity. Nel 2025 la spesa per sicurezza informat ha raggiunto i 2,24 miliardi di euro, in crescita del 12%, l'obiettivo di rafforzare governance, resilienza e investimenti. All'interno di una trasformazione digitale già in atto, secondo Anitec-Assinform «la cybersecurity rest prioritaria», condizionata soprattutto dai regolamenti europ Nis2 e Dora. Resta però alta la neces «di un vero piano industriale per il digitale – ha spieg il presidente di Anitec-Assinform Massimo Dal Che –. Il Paese deve sa governare gli asse strategici digitali, infrastrutture ai da fino alle tecnologi partire dall'Ia».

Quanto vale il mercato italiano del digitale

84,4

I miliardi di gir d'affari del digi nel 2025, +4,8 rispetto al 202

18,8

I miliardi di euro servizi Ict (+8,1% prima voce di spe per il digitale in It

20,6

I miliardi spesi pe l'acquisto di dispositivi e siste per il digitale (+1,