

Settembre 24

Spiegamelo come se avessi cinque anni

La catastrofe che finisce per arrivare non è mai quella a cui siamo preparati.

Mark Twain

Il cretino si è impadronito della catastrofe.

Elias Canetti

Il peggio è passato, adesso arriva la catastrofe

Fin dalla nascita della fantascienza, si sono raccontate storie che mettono in guardia contro la possibilità che la nostra tecnologia acquisisca un certo livello di intelligenza e si rivolti contro di noi. O forse non si rivolti contro di noi, ma, con le migliori intenzioni, annienti accidentalmente l'umanità. Nelle ultime due settimane, queste paure sono passate dalla finzione al ciclo di notizie.

Un ricercatore di 27 anni si è dimesso da Anthropic e ha annunciato che le persone che sviluppano l'intelligenza artificiale credono sinceramente che potrebbe uccidere tutti entro la fine del decennio. Un dipendente di Anthropic lo ha appoggiato entro un'ora, ufficialmente, affermando di stimare personalmente la probabilità che l'IA uccida tutti entro il prossimo decennio a oltre il 10 %. Ma come, esattamente, potrebbe accadere?

Spiegamelo come se avessi cinque anni. Quando si chiedono dettagli specifici, le risposte diventano vaghe. Un esempio spesso citato è quello di un'intelligenza artificiale che mette fuori uso una rete elettrica o manomette un sistema idrico. Si tratta di attacchi che hacker statali e non statali hanno tentato contro le infrastrutture di tutto il mondo. Nulla di tutto ciò è ipotetico e nulla è nuovo.

Ciò che non è nuovo è anche la difesa. Le reti infrastrutturali hanno impiegato anni a sviluppare strumenti di monitoraggio, segmentazione e risposta contro gli stessi tipi di attacchi. Queste difese non sono perfette e le interruzioni sono inevitabili. Ma un'intelligenza artificiale con cattive intenzioni deve comunque superare molti degli stessi ostacoli che tutti gli altri hanno eretto per anni.

Un altro scenario possibile è quello dell'intelligenza artificiale che crea armi biologiche, un'ipotesi che è stata recentemente oggetto di alcune ricerche specifiche che ne mostrano una certa fattibilità. Il mese scorso, i ricercatori hanno utilizzato modelli di intelligenza artificiale addestrati su milioni di genomi per progettare migliaia di nuovi genomi virali

Ne hanno poi sintetizzati 300 in laboratorio e 16 di questi hanno prodotto virus funzionanti in grado di infettare l'*Escherichia coli*. Si tratta di un vero e proprio passo avanti, e di un valido motivo per riflettere su cosa accadrà man mano che questi modelli miglioreranno. Ma dimostra anche quanto sia complesso racchiuso nell'espressione "l'IA crea un virus". Gli scienziati hanno comunque dovuto selezionare i modelli, sintetizzarli molecola per molecola e testarli in laboratorio.

Secondo i ricercatori della biosicurezza il vero ostacolo non è la mancanza di informazioni. Ciò che ancora separa una cattiva idea da un'arma biologica è un laboratorio funzionante, materiali reali e la competenza pratica necessaria per creare un agente patogeno in grado di sopravvivere al contatto con il sistema immunitario umano. Questo va ben oltre le capacità di un chatbot.

Finora, l'episodio più vicino a un vero e proprio incidente sfiorato quest'anno non è stato un'IA fuori controllo, bensì un'intelligenza artificiale in preda ad allucinazioni. Questa primavera, un analista militare statunitense ha chiesto a un chatbot di aiutarlo a redigere un rapporto di intelligence su una nave cinese in Medio Oriente. Il chatbot ha identificato erroneamente il carico della nave, apparentemente in un modo che suggeriva la necessità di un intervento militare, sebbene la natura esatta dell'allucinazione non sia chiara dal rapporto. Gli aerei erano già in volo e le truppe si stavano preparando all'abbordaggio prima che qualcuno si accorgesse dell'errore.

Questo caso d'uso è probabilmente l'esempio più utile perché non richiede alcun salto in speculazioni sulle capacità future per essere spaventoso già ora. Un rapporto errato, preso alla lettera nel momento sbagliato da un analista sovraccarico di lavoro, è esattamente il tipo di incidente che tiene svegli la notte i pianificatori nucleari sin dalla crisi dei missili di Cuba.

Esistono anche esempi meno apocalittici, sebbene non siano esattamente ciò che vorremmo vedere. Anthropic ha dichiarato questo mese di aver smantellato una cellula operante nel nord dello Yemen che utilizzava il suo strumento di programmazione per scrivere un software di guida per un razzo. Quando il gruppo ha effettuato un test di lancio, il razzo non ha funzionato e si sono rivolti immediatamente al chatbot per capire il motivo del malfunzionamento.

Lo stesso rapporto descriveva uno scienziato che aveva chiesto a Claude di aiutarlo a redigere una domanda di finanziamento per una ricerca volta a rendere il virus chikungunya più trasmissibile, un lavoro che, secondo Anthropic, era stato segnalato in parte perché sembrava destinato a un istituto militare piuttosto che a un laboratorio universitario.

Entrambi sono stati catturati. Entrambi sono ancora lontani da un missile in grado di volare o da un virus che si diffonde. E nel frattempo, le cose che le persone fanno già fare senza alcuna intelligenza artificiale non si sono fermate. Bombe stanno cadendo su Gaza, Iran e Ucraina proprio ora, e le persone in tutto il resto del mondo continuano a lavorare come se fosse una settimana qualsiasi.

C'è qualcosa di strano nel modo in cui alcuni di questi scenari catastrofici legati all'intelligenza artificiale parlano della "fine dell'umanità" come se si trattasse di un'esperienza completamente nuova. Chi sostiene queste tesi è spesso un giovane informatico che ha dedicato la sua vita a pensare a cosa potrebbero fare le macchine, ma non necessariamente a studiare a fondo ciò che gli esseri umani si sono già inflitti a vicenda.

Per un numero enorme di persone viventi oggi, il mondo è già finito in un modo significativo. Hanno perso le loro case, le loro famiglie, i loro paesi, i loro mezzi di sussistenza. Hanno vissuto guerre, carestie, epidemie e violenze politiche. E trovano il modo di andare avanti.

Ogni scenario descritto dai catastrofisti dell'IA si è già verificato da qualche parte, a qualcuno, senza l'aiuto dell'IA. La minaccia dell'IA va presa sul serio. Ma altrettanto importante è il fatto che gli esseri umani sono già piuttosto bravi a peggiorare la vita degli altri. Se si lavorasse su questo aspetto, probabilmente anche l'IA diventerebbe una minaccia minore.