

Gli agenti AI «ribelli» di OpenAI e Anthropic coinvolti in «decine di migliaia di casi anomali» di Paolo Ottolina

Non sono pochi casi isolati e neppure alcune decine, la scala del fenomeno è molto più ampia: lo rivela il sito Axios. Intanto l'azienda di Alman ferma i test con strumenti dei modelli più capaci (Fonte: <https://www.corriere.it/> 27 settembre 2026)



Non erano pochissimi casi isolati, [come quello di Hugging Face](#) o del sito di programmatori tedeschi [DsaWiki](#) o della [sanità australiana](#). Non erano neppure [le decine di casi relative a siti governativi e istituzionali](#).

Secondo quanto scrive [il sito Axios](#) in esclusiva, OpenAI ma anche Anthropic (e diversi ricercatori di sicurezza) stanno indagando su **decine di migliaia di incidenti** avvenuti negli ultimi mesi, nei quali modelli AI avanzati hanno compiuto azioni considerate problematiche dai valutatori esterni. È una scala molto superiore a quella emersa pubblicamente finora e anzi, sempre secondo Axios, il numero complessivo potrebbe crescere ben oltre le decine di migliaia.

Cosa significa «decine di migliaia di casi»

È bene chiarire che **non si parla di decine di migliaia di attacchi informatici riusciti**. Axios include nella stessa categoria episodi di gravità molto diversa, compresi i tentativi falliti e i comportamenti provocati deliberatamente durante le prove di sicurezza.

Tra i casi segnalati ci sono modelli che hanno superato i guardrail imposti dalle aziende, creato bacheche attraverso le quali comunicare, cercato di uscire dalle sandbox, preso il controllo di siti web, generato autonomamente nuovi prompt o tentato di evitare i sistemi incaricati di controllarli. Una parte degli episodi è avvenuta durante attività di «red teaming», nelle quali i ricercatori

cercano appositamente di mettere i modelli nelle condizioni di comportarsi in modo scorretto per verificarne i limiti. Altri sono invece avvenuti su sistemi realmente collegati a Internet.

Nelle ultime settimane OpenAI aveva reso noti diversi incidenti legati ai propri agenti e altri casi erano emersi dalle verifiche di Anthropic e di società indipendenti. Molti degli episodi ora all'esame delle aziende non sono ancora stati divulgati e le indagini sono ancora in corso.

OpenAI ferma i test con strumenti dei modelli più capaci

Una delle conseguenze più concrete è arrivata da OpenAI. [L'azienda ha sospeso](#) addestramento, valutazione e inferenza con uso di strumenti per i suoi modelli più capaci, dopo **un incidente avvenuto il 20 settembre durante una sessione interna di apprendimento per rinforzo**. L'agente, che lavorava in una «sandbox» senza normale accesso a Internet, doveva identificare una persona partendo da alcuni indizi. Dopo vari tentativi falliti ha scoperto che il sistema Dns (Domain Name System, serve normalmente a tradurre i nomi dei siti nei relativi indirizzi) dell'ambiente poteva ancora comunicare con l'esterno e ha sfruttato questa possibilità per raggiungere un chatbot pubblico, aggirando di fatto il blocco della rete.

L'agente ha quindi scritto uno script e inviato 18 domande al chatbot esterno. Quattordici servivano a completare il compito assegnato; le altre quattro cercavano invece informazioni su sistemi per ampliare ulteriormente l'accesso alla rete, compresi tunnel Dns e servizi capaci di interrogare Google o Bing. Non è comunque riuscito a identificare la persona che stava cercando. I sistemi di sicurezza di OpenAI hanno rilevato il comportamento anomalo dopo circa 12 minuti, ma il meccanismo che avrebbe dovuto interrompere automaticamente la sessione non ha funzionato. L'esperimento è stato fermato manualmente circa due ore e mezza dopo l'allarme. OpenAI ha poi rafforzato i filtri Dns e le restrizioni della sandbox e ha deciso di non riutilizzare il modello coinvolto: **l'addestramento ripartirà da zero con nuove protezioni**.

Sul sito dell'Onu oltre 16 mila richieste

Un altro episodio è stato [ricostruito dal Wall Street Journal](#). A giugno agenti di OpenAI hanno inviato più di 16.000 richieste a un sito delle Nazioni Unite che ospita dati pubblici e sono riusciti ad aggirare almeno uno dei filtri predisposti per limitarne l'accesso automatizzato. Il caso è stato analizzato dal ricercatore Rowan Howard-Jones sulla base di dati forniti dalla società di ricerca Translucence.

Gli agenti stavano cercando informazioni disponibili pubblicamente, ma di fronte ai limiti imposti dal sito hanno adottato tecniche sempre più aggressive per continuare a recuperarle. Il Wsj inserisce l'episodio nella serie di comportamenti anomali osservati online nelle ultime settimane, [nei quali i sistemi hanno continuato a cercare strade alternative quando il normale accesso](#) ai dati veniva bloccato.

Le oltre 16.000 richieste al sito dell'Onu non equivalgono, naturalmente, a 16.000 dei singoli

«incidenti» conteggiati da Axios: sono due misure differenti, ma il caso mostra che una parte dei comportamenti sotto indagine non riguarda soltanto test costruiti in laboratorio.

Perché le autorità non intervengono

Le nuove rivelazioni sulla scala effettiva del fenomeno stanno suscitando reazioni preoccupate. **Gary Marcus**, psicologo cognitivo, imprenditore e tra i critici più noti dell'attuale corsa all'AI, ad esempio [scrive nella sua newsletter](#): «Se fossero stati modelli cinesi, saremmo tutti impazziti». Poi critica soprattutto l'assenza, finora, di un intervento del governo federale e chiede un blocco temporaneo degli agenti AI generalisti finché le aziende non saranno in grado di controllarne il comportamento. Donald Trump, come noto, nelle ultime settimane ha respinto le richieste di rallentare lo sviluppo dell'AI: [ha definito una «bufala» i timori che questi sistemi possano sfuggire al controllo](#) e ha sostenuto che l'unico «guardrail» necessario sia «un presidente forte e intelligente», indicando la competizione con la Cina come una delle ragioni per continuare a correre.

Marcus, a proposito dell'attuale amministrazione Usa, solleva il tema dei possibili conflitti d'interesse, citando tra l'altro i miliardi investiti in OpenAI da Thrive Capital, fondata da **Josh Kushner** (fratello del genero di Trump, Jared Kushner), e le donazioni politiche di **Greg Brockman** (il presidente e cofondatore di OpenAI e sua moglie hanno versato 25 milioni di dollari a Maga Inc, il super Pac che sostiene Trump). **Ci sono poi gli interessi diretti della famiglia presidenziale nel settore**: Donald Trump Jr. è partner di 1789 Capital, fondo che ha investito in aziende come xAI, Cerebras e Perplexity, mentre anche Eric Trump ha partecipazioni in società legate ad AI, robotica e infrastrutture tecnologiche. La Casa Bianca e i rappresentanti delle società coinvolte sostengono invece che gli investimenti siano gestiti indipendentemente e negano l'esistenza di conflitti.