Claude Mythos, il futuro modello di Anthropic fa paura: «Sicurezza informatica a rischio» *di Roberto Cosentino*

Il nuovo modello della società di intelligenza artificiale dei fratelli Amodei trapela attraverso una fuga di dati e causa il crollo in Borsa delle aziende di cyber security: le sue capacità di sfruttare vulnerabilità potrebbero travolgere chi si occupa di difesa digitale
(Fonte: https://www.corriere.it/ 31 marzo 2026)

In questi giorni il mondo dell'informatica è in fermento per via di un leak che riguarda quello che dovrebbe il nuovo modello di intelligenza artificiale di **Anthropic**, con il nome in codice di «**Mythos**». La fuga di dati - **neanche minima, si parla di 3.000 asset** - ha avuto luogo in seguito ad un errore umano nella configurazione del sistema con cui lo staff gestisce le pubblicazioni sul proprio sito Web. La sua importanza è tale da aver causato ripercussioni in Borsa per diversi titoli di aziende di sicurezza informatica.

Il sistema «più capace»

Anthropic ha rimediato tardivamente all'errore, tanto che ha poi confermato a Fortune quello che effettivamente sarà il prossimo modello. Secondo quanto affermato dalla compagnia dei fratelli Amodei, segnerà un passo fondamentale. Il condizionale è d'obbligo. Secondo *Fortune*, il modello in questione è costoso da eseguire e comunque non è ancora pronto per il rilascio generale.
Ma è anche quello che un portavoce dell'azienda ha definito il sistema «più capace» costruito finora da Anthropic. Una frase che in realtà da anni si legge e si sente **nelle nuove generazioni di**

ogni prodotto, software o hardware. Ma Claude Mythos sfoggerà progressi significativi in *reasoning*, *coding* e *cybersecurity*. Divese società di cybersecurity hanno riportato un calo dei propri titoli dopo che il leak è diventato di pubblico dominio: tra gli altri si parla di **Palo Alto Networks**, **CrowdStrike**, **Zscaler** e **Fortinet**, con cali a Wall Street che sono stati in alcuni casi netti ma temporanei.

Che cosa fa paura di Mythos

Le capacità di sfruttare le vulnerabilità da parte di **Mythos** potrebbero essere così avanzate da anticipare, anche in velocità, chi si occupa proprio di **difesa digitale**. Al momento si tratta di informazioni ottenute tramite la visione di *Fortune* dei documenti trapelati, ma di certo non è una novità che l'AI abbia un ruolo determinante negli attacchi informatici moderni. Anzi, ne parla proprio Anthropic nello stesso blog, **in un report** pubblicato lo scorso novembre. All'epoca la compagnia ha descritto ciò che viene definita come **la prima campagna di spionaggio informatico orchestrata dall'intelligenza artificiale**. Secondo Anthropic, un attore statale cinese avrebbe impiegato **Claude Code** contro 30 target e una quota di quasi il 90% del lavoro svolto in autonomia dal sistema di intelligenza artificiale, con l'obiettivo di portare a termine attività per cercare vulnerabilità, sfruttarle o raccogliere credenziali ed esfiltrare dati.

Questo però, non ha escluso l'affiorare di **allucinazioni** da parte dell'AI, con cui gli attaccanti hanno avuto a che fare. Insomma, il dilemma sicurezza è un punto nevralgico del nuovo modello, nel bene e nel male. Così come può portare vantaggio negli attaccanti, potrebbe giovare anche chi si occupa di difesa.

Anticipare l'accesso?

Tuttavia, secondo **Axios** l'azienda avrebbe avvisato in privato **funzionari governativi statunitensi** del fatto che **Mythos** potrebbe rendere nel 2026 gli attacchi informatici su larga scala più probabili e più efficaci. La scelta di Anthropic comunque è quella di dare **accesso anticipato** alle organizzazioni difensive, per far sì che irrobustiscano i propri strumenti contro **possibili attacchi** derivati dall'uso di algoritmi di AI. Difficile non ricordare lo scontro **senza precedenti** che ha visto contrapposti proprio **Anthropic e gli Stati Uniti** e che vede la società **recentemente vittoriosa** contro il governo in tribunale. La fuga di dati ha fatto trapelare anche nuove informazioni su quella che potrebbe essere una classe di nuovi modelli: tra questi vi sarebbe anche il modello noto come Capybara e potrebbe esserci un avvicendamento sugli attuali Opus, utilizzati dal chatbot Claude.