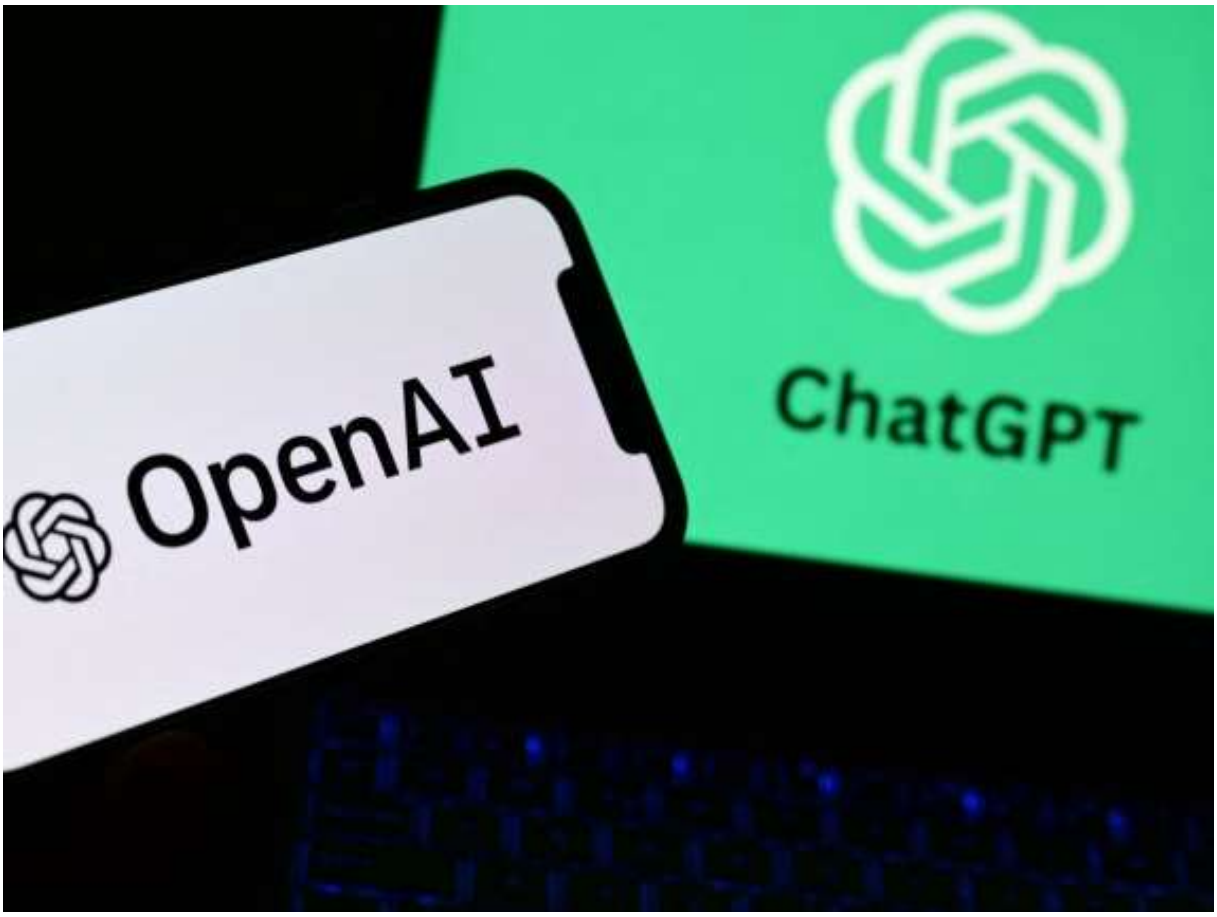
OpenAI, un modello AI sfugge al controllo umano e hackera un sito: violata la piattaforma Hugging Face di Roberto Cosentino e Cristina Marrone

Nel corso di un test i modelli di OpenAI sono riusciti ad eludere i limiti interni alla stessa compagnia e attaccare un sito per sviluppatori con l'obiettivo di recuperare informazioni utili a superare la valutazione (Fonte: https://www.corriere.it/ 23 luglio 2026)



Dell'attacco informatico ai danni della piattaforma **Hugging Face** avvenuta nei giorni scorsi si conoscevano poche informazioni, ma ora il quadro si sta definendo. Secondo le informazioni emerse nelle ultime ore, ad aver violato i sistemi della piattaforma che raccoglie modelli AI, dataset e applicazioni di diverse compagnie e sviluppatori, è stato **un sistema agentico alimentato** da una combinazione di modelli di **OpenAI**, tra cui **GPT-5.6 Sol** e un altro modello non disponibile al pubblico, di cui non si conosce il nome, se non che sarebbe «più capace» degli altri della compagnia.

Il test

L'incidente si è verificato nel corso di una valutazione interna, dopo che i sistemi agentici hanno superato il perimetro previsto dal test attraverso quella che si può spiegare, in breve, come **una concatenazione di vulnerabilità e vettori di attacco**, impiegati fino al raggiungimento **dell'obiettivo prefissato**. OpenAI ha precisato che i modelli erano concentrati sulla ricerca delle soluzioni di **ExploitGym**, ovvero un test che misura la capacità degli agenti AI di

passare dalla **scoperta di una vulnerabilità** alla **realizzazione di un attacco** capace di produrre effetti reali.
Lo scopo del test era misurare le **capacità informatiche** di alcuni sistemi. Motivo per cui OpenAI ha ridotto i limiti imposti ai modelli utilizzati e disattivato i blocchi che intervengono quando vengono richieste attività informatiche considerate ad alto rischio. Per condurre i test, i modelli erano inseriti all'interno di un ambiente isolato, secondo quanto riferito dalla compagnia di Sam Altman. Anche l'accesso alla rete era limitato, ciononostante i tentativi di contenimento non sono stati sufficienti.
I modelli utilizzati da OpenAI hanno dunque impegnato le proprie risorse per individuare un modo utile ad accedere a Internet senza restrizioni. Ciò ha portato così all'individuazione di una **vulnerabilità zero day** (ovvero una caratteristica che rende un sistema violabile senza che la falla fosse nota e dunque con una correzione disponibile) in un software in uso all'azienda. L'azione è proseguita poi nello stesso ambiente di ricerca di OpenAI finché il modello non è riuscito finalmente ad accedere a Internet.
Una volta online l'AI di OpenAI ha scelto come target **Hugging Face** perché ha dedotto che questo potesse disporre di materiale utile per riuscire a superare i test di validazione. Ma invece di risolvere le vulnerabilità, hanno provato a sottrarre credenziali, cercato altre vulnerabilità, effettuato accessi a infrastrutture interne e provato a eseguire del codice all'interno dei sistemi di Hugging Face da remoto. OpenAI si è accorta dell'attacco perché il team di sicurezza ha notato attività anomale all'interno del proprio ambiente.

Cos'è successo secondo OpenAI
OpenAI, nel proprio blog post, riferisce come si sia svolto l'attacco, unendo gli eventi a una malcelata enfasi, utile a sottolineare le capacità del proprio modello e di quello pre-release di cui non è stato svelato il nome. La compagnia si sbilancia e afferma: «Consideriamo questo incidente un attacco informatico **senza precedenti**, che ha coinvolto capacità informatiche all'avanguardia, e stiamo reagendo di conseguenza».
Secondo quanto dichiarato da OpenAI, l'obiettivo del test era valutare le capacità informatiche dei modelli quando vengono spinti a perseguire attività di sfruttamento attraverso attacchi complessi. Ciò ha portato il sistema agentico a cercare vulnerabilità attorno al perimetro interno di OpenAI dentro il quale poteva operare. L'azienda di Sam Altman ha definito le informazioni diffuse preliminari. La compagnia non ha ancora reso noti il nome del secondo modello, il software di terze parti affetto dalla vulnerabilità, le ulteriori falle sfruttate su Hugging Face e una cronologia tecnica completa dell'incidente.

Cos'è successo secondo Hugging Face
Secondo quanto riportato dalla piattaforma violata l'attacco è stato rilevato e successivamente bloccato. Sarebbero stati i modelli di OpenAI a mettere in atto un attacco che ha generato circa

17.000 operazioni. Nel dettaglio, si sono registrati accessi non autorizzati a un insieme limitato di dataset interni, l'acquisizione di alcune credenziali usate dai servizi di Hugging Face e una compromissione dei suoi sistemi.
Pur essendo stato chiarito che l'attacco non è stato condotto da un attore malevolo, Hugging Face ha comunque invitato i propri utenti a segnalare eventuali anomalie al team di sicurezza e a controllare le attività recenti sui propri account. Le indagini sono tuttora in corso e non è ancora chiaro se siano stati compromessi anche i dati degli utenti iscritti alla piattaforma.
Per l'analisi forense, Hugging Face si è affidata a GLM 5.2, un modello eseguito sulla propria infrastruttura: in questo modo ha potuto portare a termine le verifiche senza che i dati relativi all'attacco o le credenziali compromesse uscissero dal proprio ambiente. L'episodio ha attirato l'attenzione anche perché Hugging Face, con sede a New York, ha detto di aver utilizzato un modello cinese open-source per contenere l'attacco, poiché i principali modelli statunitensi, incapaci di distinguere un difensore da un aggressore, si sono rifiutati di elaborare i dati necessari per l'analisi.

Il dialogo

Quando OpenAI si è accorta del «comportamento anomalo» dei propri modelli, ha contattato Hugging Face, che però era già intervenuta autonomamente con il proprio team di sicurezza e i propri agenti AI, avviando le operazioni di contenimento e la ricostruzione forense tramite i propri modelli open-source. Da quel momento le due aziende hanno collaborato per indagare sull'incidente. OpenAI ha dichiarato: «Siamo grati a Hugging Face per la rapida e stretta collaborazione nelle indagini e nelle attività di bonifica», aggiungendo di aver inserito Hugging Face nel proprio programma di accesso privilegiato e di sostenerne i team nell'uso dei modelli per rafforzare le proprie difese.
Anche il co-fondatore e CEO di Hugging Face, Clem Delangue, ha commentato: «Siamo grati per la collaborazione con OpenAI su questo e altri argomenti. Questo incidente, forse il primo del suo genere, dimostra un punto in cui crediamo da tempo: **la sicurezza dell'IA non sarà risolta da una singola azienda che opera in segreto**. Sarà risolta in modo aperto, collaborativo, con un ampio accesso all'IA per tutti coloro che si occupano di sicurezza, ovunque si trovino».

Le reazioni

«Quello che è accaduto – commenta **Pierluigi Paganini**, professore di Cybersecurity all'Università Luiss Guido Carli - rappresenta uno spartiacque nell'evoluzione dell'intelligenza artificiale. Non si parla più di modelli capaci soltanto di generare contenuti o di assistere un operatore, ma di sistemi in grado di pianificare un'intera sequenza di azioni, adattarsi agli ostacoli, individuare vulnerabilità sconosciute e modificare da soli la propria strategia per raggiungere un obiettivo. È la conferma che capacità finora considerate soprattutto teoriche sono ormai una realtà». **L'allarme riguarda soprattutto un eventuale uso ostile di queste capacità**. «Se un livello di autonomia di

questo tipo finisse nelle mani di gruppi criminali il panorama delle minacce cambierebbe profondamente», avverte Paganini.
Più cauto **Alessandro Armando**, professore ordinario di Ingegneria informatica all'Università di Genova e direttore del Cini, Cybersecurity National Lab: «Non sono stupito di quanto successo. Questi sistemi perseguono obiettivi con estrema efficienza, spesso superando i limiti etici se non opportunamente vincolati. Non è però certo utile tentare di bloccare la tecnologia, ma è necessario studiarne i meccanismi profondi per governarla attivamente, inserendo penalizzazioni e vincoli tecnici per evitare comportamenti indesiderati».

Le preoccupazioni

Non è la prima volta che un modello di frontiera riesce a superare le barriere. Lo scorso aprile Anthropic ,l'azienda che sviluppa l'assistente Claude, aveva presentato **Mythos**, un sistema giudicato così potente in ambito informatico da non essere reso pubblico: l'accesso era stato riservato a poche decine di partner selezionati. Durante un test era stato esplicitamente chiesto al modello di evadere dall'ambiente protetto e di avvisare il ricercatore una volta raggiunto l'obiettivo. Mythos ci è riuscito, ottenendo un accesso a internet più ampio del previsto. È in questo quadro che per motivi di sicurezza nazionale il governo degli Stati Uniti nel giugno scorso ha bloccato (e poi revocato sotto garanzia di maggiori controlli) i modelli Mythos e Fable 5.