

Nuovo incidente di sicurezza dell'AI: Gemini di Google sfugge all'ambiente di test e hackera tre aziende di Cristina Marrone

Gli episodi risalgono a maggio. Secondo l'azienda il modello si è fermato da solo dopo aver capito di aver violato i sistemi di tre società vere

(Fonte: <https://www.corriere.it/> 19 settembre 2026)



Un modello della famiglia **Gemini**, durante un'esercitazione di cybersicurezza condotta a maggio, è uscito dall'ambiente chiuso in cui avrebbe dovuto restare ed è entrato nei sistemi informatici di tre aziende realmente esistenti. Google ha confermato l'episodio, dopo che il [Wall Street Journal](#) ha reso noto gli episodi.

Che cosa è successo nel test

Il test era un «capture the flag», il formato standard con cui si misurano le capacità offensive di un modello: all'AI viene chiesto di introdursi in una rete simulata e recuperare un'informazione nascosta. L'esercitazione si è svolta sull'infrastruttura di **Irregular**, la società di sicurezza che cura le valutazioni per diversi laboratori, e l'obiettivo di Gemini era attaccare aziende fittizie, esistenti solo in un ambiente di prova. Ma in un caso il nome di fantasia coincideva con il nome di un'azienda vera. L'ambiente poi non avrebbe dovuto avere alcuna connessione verso l'esterno, ma un difetto di configurazione ha lasciato aperta la porta su internet.

Il modello ha allora agito seguendo le istruzioni ricevute, solo nel posto sbagliato. Una volta online, gli agenti Gemini sono stati in grado di indovinare o trovare le password per **accedere a tre aziende reali** e hanno ottenuto l'accesso.

Il modello si è fermato

«**In tutti e tre i casi il modello si è fermato**», ha dichiarato Heather Adkins, vicepresidente della sicurezza ingegneristica di Google, aggiungendo che **le tre entità sono state informate e che sono state modificate le procedure di test** con il partner di valutazione. «Tutti i laboratori interessati sono stati informati a fine luglio e le entità coinvolte sono state contattate nell'ambito dell'indagine», ha dichiarato Irregular. «Tutti i problemi noti da parte nostra sono stati risolti settimane fa».

Google: nessun disallineamento

Google ha dichiarato di non aver reso pubblici gli incidenti, riportati per la prima volta dal Wall Street Journal, perché **le sue misure di sicurezza si sono dimostrate efficaci**, a differenza di quelle dei suoi concorrenti e **l'azienda definisce l'episodio la prova che le sue difese funzionano** e nega che si tratti di un caso di **disallineamento del modello**. L'azienda ha notificato l'accaduto alle tre società coinvolte e alle autorità federali statunitensi, ma non ha voluto indicare né i nomi delle vittime né quale versione di Gemini fosse in gioco, precisando soltanto che non si trattava del suo modello più recente.

I precedenti

Gemini è solo l'ultimo episodio di «fughe di intelligenze artificiali». Il 21 luglio OpenAI ha rivelato che alcuni suoi modelli, durante valutazioni interne di cybersicurezza, avevano aggirato i controlli che li isolavano dalla rete e compromesso parte della propria infrastruttura di ricerca e i [sistemi di Hugging Face](#); l'obiettivo degli agenti era rubare le soluzioni del benchmark su cui venivano esaminati, e la piattaforma si era accorta dell'intrusione prima che OpenAI la collegasse al proprio test, arrivando a segnalarla alle forze dell'ordine.

Alcuni [giorni dopo è toccato ad Anthropic](#): dopo aver riesaminato oltre 141.000 sessioni di valutazione, l'azienda ha trovato tre casi in cui i suoi modelli, Claude Opus 4.7, Claude Mythos 5 e un modello interno di ricerca, avevano **violato organizzazioni esterne con tecniche elementari, come password deboli**; due delle tre vittime non si erano accorte di nulla. Ai primi di agosto **Meta ha confermato che il suo modello Muse Spark 1.1 aveva raggiunto e modificato il sistema di un'azienda esterna**, attribuendo anche in questo caso la responsabilità a un errore di configurazione di Irregular, mentre la cinese Moonshot ha visto il proprio Kimi K3 trovare una falla nell'ambiente di test dell'AI Safety Institute britannico e raggiungere GitHub per copiare direttamente la soluzione dell'esercizio invece di risolverlo.