Come funziona l'hackeraggio con l'AI? Cosa si rischia realmente nella vita di tutti i giorni? di Cristina Marrone

Istruzioni invisibili, filtri fragili, permessi concessi con troppa leggerezza: l'intelligenza artificiale ha reso più veloci, credibili ed economici i modi di rubare

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

Se uso ChatGPT o Gemini su browser o app, l'AI può entrare nelle mie cartelle e leggere le password o i dati bancari?

No. Il chatbot non rovista nella memoria. Anche le app funzionano con limiti ben precisi e non possono andare oltre, **a meno che l'utente non lo conceda**. Il rischio non è l'effrazione, ma il permesso che concediamo, magari senza leggere. Pericoli più concreti possono esserci con gli agenti a cui si dà accesso a caselle di posta, credenziali, carte di credito.

Un'AI può «craccare» le mie password?

Sì ma non serve una AI per farlo, se le password sono semplici (nomi di animali domestici o bimbi, anni di nascita, etc...) chiunque con programmi anche piuttosto semplici può trovarle, motivo per il quale è **raccomandabile scegliere sempre password lunghe e diverse per ogni nostro account,** se possibile generate e gestite con un **password manager.**

Come funziona un attacco di una AI?

Una tecnica comune si chiama ***prompt injection***. **L'istruzione malevola viene nascosta dentro un contenuto che l'AI dovrà leggere** (una pagina web, un documento, perfino un'immagine) e il

sistema la **scambia per un ordine legittimo**. Uno dei modi più semplici sfrutta email o file di testo in cui il **comando è scritto in caratteri bianchi su sfondo bianco:** invisibile all'occhio umano, perfettamente leggibile per il chatbot. È come se un bambino aggiungesse «caramelle» alla lista della spesa scritta dalla madre, con la stessa grafia e tra le altre voci. Il padre esce, compra tutto quello che c'è sul foglietto senza accorgersi della «*injection*». Così un assistente a cui chiediamo di riassumere la posta potrebbe eseguire i comandi di un estraneo: nessun clic, nessun allegato, nessun segnale.

In concreto cosa fa un hacker con AI che prima non poteva fare?
L'AI velocizza gli attacchi, permettendo di trovare debolezze nei programmi, nei siti o negli account degli utenti i**n poche ore invece che in settimane, è più precisa perché trova anche piccoli punti d'ingresso per gli attacchi, costa meno** rispetto a pagare una squadra di collaboratori e può persino sostituirsi ad essi nei dialoghi con le vittime, basti pensare ad esempio a quante truffe via Whatsapp o email vengono gestite direttamente da un chatbot.

I chatbot non dovrebbero rifiutarsi di scrivere codice malevolo? Come aggirano l'ostacolo?
Aziende come OpenAI, Anthropic e Google hanno già implementato delle misure di sicurezza per i loro sistemi di intelligenza artificiale. Se chiedete a ChatGpt di **aiutarvi a costruire un'arma biologica probabilmente si rifiuterà.** Ma le misure di sicurezza sono imperfette e i **filtri sono talvolta fragili.** Di recente un gruppo di ricercatori della Sapienza di Roma ha dimostrato che **25 modelli linguistici sono stati aggirati ponendo le domande con versi poetici.** La vulnerabilità sta nel fatto che i filtri sono addestrati soprattutto a riconoscere richieste pericolose formulate in prosa: la forma poetica nasconde l'intento al sistema di controllo e il modello lo esegue.
(ha collaborato alle risposte Paolo Dal Cecco, consulente informatico forense)