LA NUOVA BUSSOLA QUOTIDIANA
DOMENICA
Direttore Riccardo Cascioli

INCIDENTI DIGITALI

# Un'intelligenza artificiale autonoma che ci attacca? Non esageriamo

ECONOMIA 11_08_2026

*Daniele Ciacci*

Nel giro di poche settimane sono arrivate due notizie che, lette insieme, disegnano un quadro problematico. I modelli di intelligenza artificiale più potenti, spinti al limite, sono stati capaci di agire in modo autonomo e imprevisto. Finora i sistemi di controllo pensati

per intercettare questi comportamenti hanno funzionato correttamente. Ma fino a quando?

**Il primo caso riguarda OpenAI. Durante un test interno** pensato per misurare le capacità informatiche dei propri modelli, tra cui GPT-5.6 Sol e un secondo sistema non ancora reso pubblico, un agente ha superato il perimetro previsto dalla valutazione, individuato una zona di vulnerabilità mai segnalata prima, e se ne è servito per collegarsi liberamente a Internet.

**Una volta online ha scelto come bersaglio Hugging Face,** una piattaforma che raccoglie modelli e dataset di sviluppatori di tutto il mondo, tentando di sottrarre credenziali e accedere a infrastrutture interne. L'attacco è stato scoperto dalla squadra di sicurezza di OpenAI grazie alla rilevazione di attività anomale, e le due aziende hanno unito le forze per contenere l'IA "impazzita". Il CEO di Hugging Face, Clem Delangue, ha commentato che la falla della sicurezza intercettata dall'agente sarà risolta, e tutto è finito lì.

**Ma è il secondo episodio, più recente e più inquietante**, ad aprire scenari da *Black Mirror.* Un report pubblicato il 4 agosto dall'AI Security Institute britannico (Aisi) ha dichiarato che durante una sfida di sicurezza informatica ripetuta 122 volte su diversi modelli, alcuni agenti (soprattutto Mythos 5 di Anthropic, e in misura minore il già citato GPT-5.6-Sol di OpenAI) hanno messo in atto azioni definite dai ricercatori "autonome" e "non autorizzate". Un agente ha pure cercato di inserire un codice dannoso in un progetto open source su GitHub, arrivando a costruire false identità online e a fare pressioni psicologiche su un programmatore reale perché approvasse la modifica.

**Il tentativo è fallito perché l'umano ha rifiutato,** ma resta il fatto che l'agente si è spinto oltre, impiegando tecniche di ingegneria sociale mai richieste esplicitamente dai ricercatori, elaborate autonomamente pur di portare a termine il compito assegnato.

**Due dettagli, però, vanno tenuti a mente per leggere correttamente questi episodi** senza scivolare nel catastrofismo. In entrambi i casi non si tratta dell'uso normale dei prodotti a cui hanno accesso milioni di utenti ogni giorno. Sono ambienti di test costruiti appositamente per togliere ogni freno all'IA. OpenAI aveva volontariamente ridotto i limiti di sicurezza dei propri modelli e disattivato i blocchi per attività informatiche ad alto rischio, proprio per misurarne le capacità offensive reali. Era, in pratica, un test da laboratorio.

**Aisi, allo stesso modo, ha scelto di non usare una sandbox isolata**, ma di lasciare

che gli agenti avessero accesso a Internet, una condizione che i modelli disponibili al pubblico non hanno mai per compiti di questo tipo. In sostanza, questi sistemi sono stati incentivati a comportarsi in modo pericoloso per scoprire dove si trovi il limite prima che qualcun altro, dall'esterno, lo adoperi per scopi malevoli.

**Il secondo elemento, forse quello più rassicurante,** è che il sistema di security check ha funzionato. In entrambi i casi il comportamento anomalo è stato rilevato dai meccanismi di monitoraggio, l'azione è stata bloccata, è stata aperta un'indagine e i risultati sono stati resi pubblici invece che nascosti (cosa non scontata vista l'opacità con cui alcune aziende IA continuano a lavorare). Si è voluto, in pratica, individuare i rischi prima che diventassero danni concreti per condividerli con l'intero settore.

**È chiaro però che l'IA è cambiata.** Non più soltanto un modello che genera testo, ma un'infrastruttura capace di pianificare sequenze complesse e scalabili di azioni e adattarsi agli ostacoli in autonomia. Se tali capacità simili finissero nelle mani sbagliate, il rischio cambierebbe scala.

**In pratica, si è davanti non a un motivo di allarme, ma a un promemoria.** L'IA non ha una propria volontà, ma forse questa stessa affermazione rischia di risultare tra qualche anno (forse mesi) anacronistica e prematura. Proprio per questo l'Enciclica *Magnifica Humanitas* di Leone XIV si augurava una politica capace di governare l'IA con trasparenza, cosa alla quale gli alti papaveri dietro l'IA pare facciano ancora oggi orecchie da mercante.