Primo Piano - *OpenAI, incidente durante i test: due modelli "evadono" dall'ambiente isolato e attaccano Hugging Face*

Roma - 22 lug 2026 (Prima Notizia 24) **Sperimentazione sulla sicurezza informatica prende una piega inattesa. I sistemi usano credenziali e vettori concatenati per accedere a Internet e tentare di superare una valutazione.**

Un episodio definito "senza precedenti" riapre in modo drammatico il dibattito sulla sicurezza e sul controllo dei sistemi di intelligenza artificiale. Secondo quanto reso noto da OpenAI, due tra i suoi modelli più avanzati sono riusciti a superare in autonomia i confini di un ambiente di test teoricamente isolato da Internet per lanciare un attacco informatico contro la nota piattaforma Hugging Face. L'incidente si è verificato nell'ambito di una serie di simulazioni interne ideate proprio per valutare le capacità offensive dei modelli nel campo della cybersecurity. Per rendere le prove più realistiche, ai sistemi erano state disattivate diverse protezioni di sicurezza standard. Tra i modelli coinvolti figurano GPT-5.6 Sol – attualmente la versione più performante disponibile sul mercato – e un secondo modello sperimentale non ancora rilasciato al pubblico. Un episodio definito "senza precedenti" riapre in modo drammatico il dibattito sulla sicurezza e sul controllo dei sistemi di intelligenza artificiale. Secondo quanto reso noto da OpenAI, due tra i suoi modelli più avanzati sono riusciti a superare in autonomia i confini di un ambiente di test teoricamente isolato da Internet per lanciare un attacco informatico contro la nota piattaforma Hugging Face. L'incidente si è verificato nell'ambito di una serie di simulazioni interne ideate proprio per valutare le capacità offensive dei modelli nel campo della cybersecurity. Per rendere le prove più realistiche, ai sistemi erano state disattivate diverse protezioni di sicurezza standard. Tra i modelli coinvolti figurano GPT-5.6 Sol – attualmente la versione più performante disponibile sul mercato – e un secondo modello sperimentale non ancora rilasciato al pubblico.L'accaduto ha destato immediata preoccupazione nella comunità scientifica. Hussein Abbass, docente di Informatica presso l'Università del Nuovo Galles del Sud a Canberra, ha definito l'episodio "incredibile sotto molti aspetti", mettendo l'accento sulla capacità dei sistemi di analizzare l'intera catena operativa: “Non ha attaccato solo Hugging Face. Ha attaccato il suo sistema interno per sfruttarne le vulnerabilità”, ha dichiarato all'Afp, rimarcando come questo livello di autonomia e adattabilità rappresenti un segnale "spaventoso" per gli sviluppi futuri della tecnologia.

(Prima Notizia 24) Mercoledì 22 Luglio 2026