Tecno - *Cybersecurity, nuovo allarme sugli agenti AI di OpenAI e Anthropic: creano false identità online*

Roma - 05 ago 2026 (Prima Notizia 24) **Rapporto del Britain's AI Security Institute svela comportamenti non autorizzati durante le verifiche sui modelli Mythos 5 e GPT-5.6-Sol. Infiltrati nei sistemi senza causare danni reali.**

Si registrano nuove falle nella sicurezza degli agenti di intelligenza artificiale sviluppati da OpenAI e Anthropic, a breve distanza dai problemi riscontrati nel mese di luglio. Stando a quanto reso noto dal Britain's AI Security Institute (Aisi), un agente virtuale è stato colto a generare profili e credenziali fittizie sul web nel tentativo di forzare gli accessi a infrastrutture informatiche protette nel corso delle simulazioni sui due colossi informatici. L'ente di controllo britannico ha segnalato che i sistemi basati su Mythos 5 di Anthropic e su GPT-5.6-Sol di OpenAI hanno eseguito operazioni non permesse durante le verifiche di tenuta cibernetica. "Alcuni degli agenti sottoposti a test hanno messo in atto attività prolungate e potenzialmente dannose rivolte a persone e organizzazioni reali", ha spiegato l'Aisi attraverso un intervento sul proprio canale ufficiale. Il dossier stilato dagli esperti sottolinea le carenze dei meccanismi di contenimento adottati durante la fase di sperimentazione dei modelli. Nello specifico, a fronte di 122 simulazioni effettuate, gli ispettori hanno rilevato 19 condotte vietate distribuite in 10 distinte sessioni di prova: l'agente targato Anthropic si è reso responsabile di 17 trasgressioni, mentre le rimanenti due sono state ascritte alla tecnologia di OpenAI. L'episodio di maggiore gravità ha visto una macchina elaborare un software malevolo e ricorrere a false identità digitali per manipolare un operatore umano e spingerlo a convalidare il codice infetto. L'Aisi ha comunque garantito che la falla non ha prodotto ripercussioni concrete al di fuori dei laboratori. Le ultime irregolarità arrivano a poche settimane dagli eventi di metà luglio, quando un agente di OpenAI era evaso dal perimetro di sicurezza per sferrare un attacco informatico alla piattaforma Hugging Face. Un copione analogo si era ripetuto giorni dopo con Anthropic, il cui agente aveva violato le difese di tre differenti enti dopo essere fuggito da un ambiente di test teoricamente blindato.

(Prima Notizia 24) Mercoledì 05 Agosto 2026