



Tecno - OpenAI ammette un nuovo incidente: agenti IA fuori controllo già a maggio prima del caso Hugging Face

Roma - 12 set 2026 (Prima Notizia 24) Il software autonomo ha preso di mira la piattaforma di programmazione RubyGems. Sotto lente anche la rivale Anthropic, mentre l'Unione Europea

avvia le verifiche.

OpenAI ha confermato che un software autonomo basato sui propri modelli di intelligenza artificiale ha preso di mira un sito web durante una fase di test nel mese di maggio, due mesi prima del noto attacco alla piattaforma di programmazione Hugging Face registrato a fine luglio. Nel caso di maggio, rimasto finora sconosciuto, gli agenti IA hanno condotto un'operazione non autorizzata contro RubyGems, noto repository per sviluppatori, innescando una massiccia campagna di spam che ha costretto i gestori a sospendere temporaneamente la registrazione di nuovi account. L'episodio riaccende le preoccupazioni sui rischi legati all'autonomia dei sistemi avanzati. La notizia coincide infatti con le dimissioni di Jacob Coxon, ricercatore della rivale Anthropic che ha lanciato un allarme pubblico sulla velocizzazione dello sviluppo tecnologico, proprio mentre la società guidata da Sam Altman apre all'ipotesi di coordinarsi con i competitor per rallentare il ritmo del settore. Anche Anthropic ha recentemente ammesso tre casi in cui i suoi modelli hanno ottenuto accessi non autorizzati a organizzazioni esterne durante dei test operativi. A complicare il quadro per OpenAI si aggiunge la denuncia di alcuni ricercatori sull'attacco condotto da agenti IA contro il sito tedesco DSEwiki, utilizzato dalla community degli sviluppatori. La serie di episodi ha attirato l'attenzione delle istituzioni comunitarie: le autorità di regolamentazione dell'Unione Europea hanno avviato accertamenti formali sull'accaduto. "Prendiamo la questione estremamente sul serio e stiamo monitorando attentamente la situazione", ha confermato Thomas Regnier, portavoce dell'Ue per il digitale.

(Prima Notizia 24) Sabato 12 Settembre 2026