Experţii în cybersecurity au analizat 800.000 de aptitudini ale AI, iar 25.000 au fost considerate suspecte (analiza)

Peste 25.000 de aptitudini ale Inteligenței Artificiale (AI) dintr-un total de 800.000 analizate au fost considerate suspecte de catre experții în securitate cibernetica Eset, în urma efectuarii unei ample analize.

Potrivit datelor prezentate într-un articol publicat, recent, pe blogul din România al producatorului de soluții antivirus, din martie 2026 au fost cercetate peste 800.000 de AI skills, fiind identificate mai mult de 25.000 suspecte, dintre care aproximativ 2.500 au fost clasificate drept malițioase.

'Unele dintre cele mai evidente exemple au fost concepute pentru a descarca pe dispozitivele utilizatorilor programe malware de tip infostealer, cu scopul de a colecta informații sensibile, precum chei API, parole sau date de plata pentru aplicații de criptomonede şi servicii bancare. Altele conţineau troieni (programe malware ascunse) şi backdoor-uri concepute pentru a oferi atacatorilor acces discret la sistemele victimelor', explica specialistul Eset, Phil Muncaster.

Un studiu separat, realizat în luna februarie a acestui an, a identificat peste 800 de aptitudini maliţioase listate pe platforma ClawHub, ce conţineau programe infostealer, malware pentru furtul criptomonedelor şi 'reverse shell'-uri care le permit hackerilor sa preia controlul asupra dispozitivelor victimelor.

'Hackerii nu vizeaza doar AI skills, ci folosesc inteligența artificiala şi ca metoda de deghizare în contexte deja cunoscute, precum extensiile de browser. Microsoft a descoperit o campanie ampla în care extensii maliţioase imitau asistenți AI legitimi. Au fost concepute pentru a colecta istoricul conversațiilor utilizatorilor, care poate conţine informaţii sensibile, precum date de autentificare. Microsoft a identificat aproape un milion de descarcari doar în cadrul acestei campanii. Una dintre cele mai serioase amenințari pentru AI generativa şi agenții AI este prompt injection. Funcţioneaza în doua moduri. Prompt injection direct presupune introducerea de catre atacatori a unor instrucţiuni într-un model lingvistic mare (LLM), pentru a-l determina sa ignore mecanismele de siguranţa integrate. Însa, pentru agenții AI, mai relevant este prompt injection indirect. Instrucţiuni maliţioase sunt ascunse în pagini web, postari, cod sau alte tipuri de conţinut. Atunci când un agent interacţioneaza cu acel conţinut, acesta poate fi manipulat sa efectueze activitați maliţioase', este de parere Muncaster.

În acelaşi timp, AI poate fi utilizata şi pentru generarea în timp real a unor scripturi maliţioase sau a unor componente de malware şi ransomware.

În context, experţii în securitate cibernetica vin cu o serie de recomandari pentru utilizatori, pentru a reduce riscurile de securitate asociate utilizarii AI: precauţie la orice instrument care pare sa necesite mai multe permisiuni decât ar fi necesar; actualizari recente şi remedierea problemelor de securitate; un semnal de alarma apare atunci când pluginul trebuie sa ruleze scripturi necunoscute sau sa acceseze domenii/resurse suspecte în timpul instalarii; monitorizarea agentului/pluginului şi dupa instalare, pentru a identifica eventuale modificari de comportament realizate discret şi maliţios; menţinerea întotdeauna a software-ului agentului AI actualizat, pentru a beneficia de cea mai sigura versiune disponibila.

'În doar câţiva ani, inteligența artificiala a devenit un instrument indispensabil pentru mulţi utilizatori casnici. Acest fapt nu a trecut neobservat de infractorii cibernetici, care au dezvoltat diverse metode prin care pot transforma AI-ul utilizat zilnic într-un canal pentru distribuirea de malware, furtul de date şi alte activitaţi maliţioase. În prezent, cele mai mari riscuri apar în domeniul emergent al agenților AI. Deşi chatboţii sunt mai siguri decât în trecut, este important sa nu renunţam la masurile de precauţie. Din fericire, se contureaza deja bune practici care pot contribui la utilizarea sigura a inteligenţei artificiale. Combinarea lor cu funcţii de securitate

oferite de furnizori de încredere poate contribui la menținerea unui nivel ridicat de protecție', se arata în articolul de specialitate.

În opinia specialiştilor, un chatbot AI legitim nu 'fura' datele utilizatorului, deşi poate stoca şi partaja cu alte parţi informațiile introduse în conversații.

Eset a fost fondata în anul 1992 în Bratislava (Slovacia) şi se situeaza în topul companiilor care ofera servicii de detecție și analiza a conținutul malware, fiind prezenta în peste 180 de țari.