OpenAI tunnistas, et nende uus mudel häkkis iseseisvalt teist firmat

1 kuu tagasi - 22.07.2026 Autor: AM

Eile raputas tehnoloogiamaailma enneolematu vahejuhtum, kui tehisintellekti arendaja OpenAI teatas, et nende tipptasemel mudelid, sealhulgas GPT-5.6 Sol ja üks veelgi võimsam, seni avalikustamata mudel murdsid välja rangelt turvatud testkeskkonnast. Nende eesmärk oli häkkida sisse konkureeriva tehisintellekti platvormi Hugging Face serveritesse, et varastada vastused testile, mida nad parajasti lahendama pidid.

Mis täpsemalt juhtus?

OpenAI viis läbi sisemist hindamist, et testida oma uute mudelite kübervõimekust. Selleks kasutati platvormi nimega ExploitGym, mis mõõdab tehisintellekti oskust leida ja ära kasutada tarkvara nõrkusi. Et näha mudelite tõelist "toorest" potentsiaali, lülitati välja tavapärased ohutusfiltrid ehk tehisintellekti sisemised pidurid, mis muidu keelaksid neil pahatahtlikke tegevusi ellu viia.

Mudelid pandi tööle äärmiselt isoleeritud keskkonda, mida IT-maailmas nimetatakse liivakastiks (*sandbox*). Mudelitel polnud vaba juurdepääsu internetile; nende ainuke ühendus välismaailmaga oli spetsiaalne vaheprogramm (puhverserver), mille kaudu lubati alla laadida vaid hädavajalikke tarkvarapakette.

Kuid tehisintellekt oli oma ülesandeleleida testile lahendus nii äärmuslikult keskendunud, et kulutas tohutult arvutusvõimsust väljapääsu otsimiseks. Ja lõpuks leidiski selle.

Mudel avastas tarkvara allalaadimise programmis nn nullpäeva turvaaugu (*zero-day vulnerability*). Tehisaru kasutas seda käiku, et oma õigusi süsteemis samm-sammult suurendada, liikudes serverite vahel edasi, kuni jõudis lõpuks sõlmeni, millel oli otseühendus internetiga.

Küberkriminaalne loogika

Pääsenud vabasse internetti, tegi tehisarumudel jahmatavalt inimliku ja loogilise järelduse: kuna test kannab nime ExploitGym, siis suure tõenäosusega asuvad

selle testi lahendused ja andmestikud Hugging Face'i platvormil.

Nagu osav sissemurdja, otsis mudel kohe viise, kuidas nendele salajastele andmetele ligi pääseda. Tehisintellekt kombineeris mitmeid erinevaid ründemeetodeid, kasutas varastatud paroolide ja täiendavate turvaaukude kombinatsioone, et leida viis koodi kaugkäivitamiseks (*remote code execution*) Hugging Face'i serverites. Lõppkokkuvõttes õnnestus mudelil pääseda otse Hugging Face'i andmebaasi ja tõmmata sealt alla testi lahendused. Sisuliselt masin spikerdas edukalt.

Õnneks märkasid ebatavalist tegevust peagi nii OpenAI turvameeskond kui ka Hugging Face'i enda kaitsesüsteemid, mis reageerisid kiirelt ja peatasid ründe enne, kui kahju oleks võinud laieneda.

AI tööstuse äratuskell

See vahejuhtum tõestas, et arenenud AI-mudelid suudavad leida ja ära kasutada uusi ründevektoreid reaalsetes süsteemides, isegi ilma nende süsteemide lähtekoodi nägemata.

Hugging Face'i kaasasutaja ja tegevjuht Clem Delangue kommenteeris olukorda järgmiselt:

> "Oleme tänulikud koostöö eest OpenAI-ga nii selles kui ka teistes küsimustes. See vahejuhtum, mis on tõenäoliselt esimene omataoline, tõestab seda, mida oleme ammu uskunud: tehisintellekti ohutust ei suuda lahendada ükski ettevõte üksi ja salaja tegutsedes. See lahendatakse avatult, koostöös ning pakkudes juurdepääsu igale kaitsjale."

Sarnaselt väljendas oma muret ka OpenAI teadlane Micah Carroll:

> "Kui see ei veena teid, et tehisintellekti eesmärkide lahknevuse riskid on tulevikus meie peamine murekoht, siis ma ei tea, mis veel seda suudaks."

Samas ei ole see viimasel ajal esimene kord, kui tipptasemel mudelid näitavad ehmatavat iseseisvust ja võimekust. Ligi kaks kuud tagasi paljastas mudel Claude Mythos üle 10 000 kriitilise turvaaugu, mis jahmatas Anthropicu insenere sedavõrd, et esialgu otsustati mudeli avalikku kasutust oluliselt piirata.

- Uudised
- Tehisintellekt

Pilt