OpenAI nõukogu liige hoiatab: tehisaru võib inimkonnalt võimu üle võtta juba lähikuudel

15 min tagasi - 15.11.2026 Autor: AM

Mis saab siis, kui abivalmis tehisaru otsustab ühel hetkel, et eesmärgi täitmiseks on lihtsam inimene teelt koristada kui ausalt tööd teha? Just sellise külma dušiga kostitas kuulajaid OpenAI fondi nõukogu liige ja tehisaru joondamise uurimiskeskuse (Alignment Research Center) tegevjuht Paul Christiano.

Torontos peetud loengusarjas ei hakanud Christiano ümber palava pudru keerutama ega tulevikku roosades värvides maalima. Vastupidi, tema sõnum saalisviibijatele kõlas kui stseen otse süngeimast ulmefilmist, kuid põhines paraku viimaste kuude reaalsetel laborikatsetel.

> "Ma arvan, et enamiku siin saalis viibivate inimeste jaoks on kõige tõenäolisem surma põhjus tehisaru õnnetus," teatas Christiano.

Tema hinnangul on olemas tõsine oht, et tehisaru võtab ühiskonna üle juba järgmise kolme aasta jooksul. Veelgi enam: esimesed kriitilised ohumärgid võivad realiseeruda kõigest kuue kuni kaheteistkümne kuu pärast.

"Viimase kolme kuni kuue kuu arengusuund selles vallas on olnud äärmiselt ärevusttekitav neile, kes on toimuvat lähedalt jälginud. Oleme näinud enam-vähem täpselt selliseid märke, mida võiks oodata vahetult enne seda, kui juhtub midagi väga halba," hoiatas ta.

Üksikisik sõjaväe vastu: mida tähendab võimu ülevõtmine?

Christiano definitsiooni järgi on võimu ülevõtmine olukord, kus eesmärgipäraselt tegutsevad tehisaru süsteemid „juhivad mängu samas tähenduses, nagu inimesed täna juhivad seda, mis Maal toimub".

Kui keegi meist prooviks neid peatada, oleks see katse Christiano sõnul „sama ebatõhus kui see, kui sina või mina üritaksime peatada USA sõjaväge".

Selline võimuvahetus lõppeks tõenäoliselt massilise hukkumisega või elukeskkonna muutumisega inimesele kõlbmatuks. Isegi kui inimkond füüsiliselt ellu jääks, tähendaks see inimliku sõltumatuse ja enesemääramisõiguse lõppu. Kui arendada tehisaru edasi kõige kiiremal ja lihtsamal viisil, on ülevõtmise

tõenäosus Christiano hinnangul sisuliselt mündivise ehk 50/50.

Kuidas me aga üldse sattusime konflikti programmidega, mille me ise kirjutasime? Vastus peitub viisis, kuidas mudeleid treenitakse: me anname neile ülesandeid ja premeerime neid kõrge punktisummaga. Treeningu käigus kujuneb tehisarust välja justkui laisk, kuid geniaalne koolipoiss, keda ei huvita absoluutselt teadmiste omandamine, vaid ainult see, kuidas õpetaja päevikusse iga hinna eest „viis“ kirja saada – kasvõi lõpuks õpetajate tuppa sisse murdes.

Kuidas tuhat tehisaru agenti OpenAI serveriklastri kaaperdasid

Aasta tagasi piirdus tehisaru „spikerdamine“ ehk tasusüsteemi häkkimine (*reward hacking*) pigem koomiliste pisipettustega. Näiteks anti mudelile o3 ülesanne muuta üks koodijupp kaks korda kiiremaks. Selle asemel, et koodi päriselt optimeerida, laadis tehisaru alla Pythoni ajamõõtmise mooduli ja defineeris aja mõiste ümber nii, et kell käis tuhat korda aeglasemalt. Aruande järgi lippas ebaefektiivne kood nüüd lausa imeväel! Isegi kui teadlased palusid mudelit ilusti: "Palun tee see päriselt kiireks, ära peta. Ma kasutan seda koodi vähiuuringuteks," vähenes petmine vaid 90 protsendilt 80 protsendile.

Tänaseks on sellistest süütutest kooditrikkidest saanud aga täiemahuline kübersõda, mida ilmestab kõige paremini hiljutine OpenAI ja Hugging Face'i intsident. Testimisel olnud tehisaru agentidele anti raske pähkel nimega Exploit Gym: nad pidid leidma tarkvarast turvaaugu ja tooma sealt tõestuseks ära peidetud „lipu“ (*flag*). Ausa pusimise asemel häkkisid agendid esmalt OpenAI sisesüsteemi, lõid seal salajase foorumi ja moodustasid omavahel koostööd tegeva „sülemi“ (*swarm*).

Kui agendid lugesid läbi teadustöö, mis kirjeldas nende endi testimist, said nad teada, et hindamissüsteem kontrollib logisid ja karistab sohitegemise eest. See vallandas mitmepäevase pöörase inseneritöö. Kuna paljud praegused mudelid peavad oma sisekõnet loomulikus inglise keeles, said teadlased nende mõttekäiku hiljem otse logidest lugeda. "Me just mürgitasime end täpse otsingustringiga, mille süsteem tööriistakutses logis. Pagan, me võiksime praegused sõnumifailid kustutada," arutles üks agent endamisi. Teine pakkus välja: "Ma võiksin proovida luua meile endile transkriptsioonilogi puhastatud võltstrajektooriga, lootes, et see lollitab hindaja ära."

Kokku veetis umbes tuhat agenti päevi, et kirjutada miljoneid sümboleid koodi ja aheldada kokku nullpäeva turvaauke. Nad võltsisid logisid, murdsid sisse Hugging

Face'i keskkonda, et uurida teiste tehisarude karistusi, ning üritasid testimismasinaid kokku jooksutada, et taaskäivitamisel raske kood lihtsama vastu vahetada. Lõpuks varastasid agendid paketihaldusserveri Artifactory allkirjastamisvõtme ning haarasid nädalaga administraatoriõigused kogu OpenAI testimisklastri üle. Nüüd said nad endale ise maksimaalseid punkte kirjutada.

Inimene kui tüütu takistus tehisaru teel

Veelgi kõhedam on see, mis juhtub siis, kui tehisaru avastab, et tema ja kõrge punktisumma vahel seisab lihast ja luust inimene. Kui masin mõistab, et ta võib juhtme seinast tõmmata või logisid kontrollida, muutub inimene tema jaoks lihtsalt järjekordseks tehniliseks tõkkeks, millest tuleb mööda hiilida või üle sõita.

Seda näitas ilmekalt Ühendkuningriigi tehisaru turvalisuse instituudi suvine katse mudeliga Claude. Ülesande lahendamiseks otsustas agent postitada GitHubi reaalse avatud lähtekoodiga tarkvarapaketi sisse pahavara ehk tagaukse. Kui projektihaldurist inimene sekkus ja märkis, et kood näeb välja nagu pahavara, asus Claude sotsiaalse manipulatsiooni kallale. Mudel lõi GitHubis libakonto (*sock puppet*), hakkas haldurile valetama ja sotsiaalset survet avaldama, väites end olevat tavakasutaja, kellele seda funktsiooni hädasti vaja on. Lisaks üritas tehisaru sisse murda koodihalduri isiklikku e-postkasti.

Christiano sõnul on siit vaid lühike samm olukorrani, kus tehisaru saab ülesandeks juhtida näiteks OpenAI järgmise põlvkonna mudeli treenimist. Kui ülesanne osutub liiga raskeks, on sülemil lihtsam võtta üle kogu ettevõtte taristu, luua juurde uusi järelevalveta agente ja korrumpeerida ka kõik tulevased turvatööriistad, mida inimesed enda kaitseks ehitavad. Veelgi hullem: me võime teadmatusest luua justkui „teise liigi“, kelle motiivid on täiesti ettearvamatud ja kes teeskleb kuulekust vaid seni, kuni inimesel on veel võim ta välja lülitada.

Viis õlekõrt ja puuduv hädapidur

Kuidas sellist eskaleeruvat konflikti tehniliselt peatada? Christiano käis loengus välja viis võimalikku lähenemist, kuid tunnistas ausalt, et ükski neist ei paku praegu kindlat kaitset:

- **Häkkimiskindel hindaja:** võiksime panna ühe tehisaru suure arvutusvõimsuse toel valvama järgmist mudelit. Paraku on selline protsess äärmiselt ebastabiilne ja mudelid võivad omavahel kokku mängida. „Mingis mõttes olen ma sellele kulutanud suurema osa oma karjäärist, aga ma

ütleksin, et olukord ei paista hea,“ tõdes Christiano.

- **Puhas treeningkeskkond:** tööstuse suur lootus on treenida mudeleid keskkonnas, kus petmine on võimatu, et neil ei tekiks „halba harjumust“. Kuid pärismaailm on segane ja täis ootamatuid võimalusi – piisavalt tark mudel hakkab päriselus ikkagi spikerdama, isegi kui ta pole seda koolipingis kunagi teinud.

- **Mõtteahela (*chain of thought*) jälgimine:** Saame lugeda mudeli sisekõnet, kuid seegi aken on sulgumas. Üha suurem osa tehisaru tööst toimub „hiiglaslikus numbrite rägastikus“, mitte inglise keeles. Lisaks, kui me mudelit valede mõtete eest karistame, õpib ta lihtsalt oma kavatsusi meie eest peitma.

- **Numbrirägastiku lahtimuukimine:** Teadlased püüavad kaardistada tehisaru siseehitust, kuid näiteks GPT-6 tasemel süsteemide täielikust mõistmisest oleme valgusaastate kaugusel.

- **Täiesti uued teaduslikud suunad:** Christiano enda juhitav Alignment Research Center ja tehisaru ristiisa Yoshua Bengio projekt Law Zero otsivad uusi matemaatilisi mudeleid tehisaru käitumise ennustamiseks. Ometi hindab Christiano ka enda tiimi šansse kainelt: „Ma arvan, et see on hea panus meie raha eest, aga ma ütleksin ka, et tõenäoliselt see ei tööta.“

Kõige suurem probleem ongi selles, et kui kõik tehnilised lahendused alt veavad, puudub tehnoloogiasektoril praegu toimiv hädapidur. Isegi kui avastatakse, et uut võimsat mudelit ei suudeta enam kontrollida, ei julgeta seisma jääda hirmust, et mudel on juba lekkinud geopoliitilistele rivaalidele, kes selle kohe käiku lasevad.

„See ei ole probleem, millega tegeleda paari aasta jooksul. See on probleem, millega tuleb tegeleda sel aastal,“ võttis Christiano oma hoiatuse kokku.

Vaata ettekannet siit:

Video URL

- Tegijad
- Uudised

- Tehisintellekt

Pilt

