#COXON: ANTHROPIC I OpenAI SE 'KOCKAJU S NAŠIM ŽIVOTIMA'#

Tjedan dana ranije istraživač Jacob Coxon dramatično je napustio Anthropic. U seriji objava na X-u (pregledanih više od 70 milijuna puta) optužio je Anthropic i OpenAI da se "kockaju s našim životima" te tvrdio da ljudi koji grade umjetnu inteligenciju "iskreno vjeruju da bi nas to moglo sve ubiti do kraja desetljeća".

Amerika trenutno prolazi kroz nesvakidašnju zamjenu uloga - kreatori umjetne inteligencije, povijesno neskloni državnom uplitanju u regulativu, danas strahuju da će se AI potpuno oteti kontroli i mole vlastitu vladu za zakonska ograničenja. Predsjednik im, pak, odgovara da su defetisti koji bespotrebno paničare.
Donald Trump poručio im je da umjetnoj inteligenciji ne trebaju nikakva regulatorna tijela, već isključivo "snažan i inteligentan predsjednik" na čelu države a takvog, tvrdi, Amerika već ima.

Robote koji preuzimaju vlast i AI koji guta svijet opisao je kao potpunu izmišljotinu, uspoređujući te strahove s ranijim prognozama o klimatskoj katastrofi.

Sve je, međutim, krenulo esejom Darija Amodeija, šefa kompanije Anthropic, naslovljenim "Moramo obuzdati tempo", u kojem je upozorio da razvoj umjetne inteligencije izmiče kontroli.

Amodei je istaknuo da AI modeli ubrzano sami sebe unaprjeđuju kroz proces takozvanog „rekurzivnog samousavršavanja”, a on je toliko brz da ljudi ne stižu razvijati sigurnosne kočnice.

Njegova procjena je da bi u šest do dvanaest mjeseci takav "roj" autonomnih agenata mogao postati sposoban preuzeti kontrolu nad cijelim internetom. Šteta bi se, u slučaju izostanka zaštitnih mehanizama, mjerila u stotinama milijardi dolara.

Kao izlaz, šef Anthropica predložio je trostupanjski plan.

Prvi korak predviđa uvođenje neovisnih revizora s pristupom unutarnjim procesima tvrtki na razini zaposlenika. Drugi uključuje strogu koordinaciju sigurnosnih standarda među vodećim laboratorijima demokratskih zemalja, dok bi treći i najambiciozniji korak trebao biti međunarodni sporazum koji bi u priču uključio čak i autoritarne režime poput Kine.

Podrška konkurencije

Ono što je esej pretvorilo u pravu buru jest reakcija konkurencije.

Sam Altman (OpenAI) i Elon Musk (xAI), ljuti suparnici koji se rijetko slažu oko ičega, javno su podržali Amodeijev poziv na usporavanje.

Musk je kratko poručio da je „Dario u pravu” i da je uvođenje međusobne stručne provjeve (peer review) među laboratorijima pravi način da se proces započne.

Altman je otišao korak dalje i najavio da OpenAI odgađa izlazak na burzu barem do 2027. godine, upravo zbog sigurnosnih rizika. Potez je to koji u tehnološkoj industriji ima težinu rijetko viđenog priznanja.

Phoca PDF

Demis Hassabis, predsjednik Google DeepMinda, ocijenio je da Amodeijev esej pokazuje "pravi put naprijed" u "kritičnom trenutku", povezavši ga s vlastitim prijedlogom neovisnog tijela koje bi testiralo najnaprednije modele prije lansiranja i eventualno koordiniralo usporavanje ako rizici postanu preozbiljni.

Uostalom, sam je Hassabis već ranije upozoravao na paradoks moderne tehnologije: sposobnosti umjetne inteligencije danas napreduju znatno brže od sposobnosti samih istraživača da ih razumiju, što otvara širom vrata dramatičnim kibernetičkim, biološkim, pa čak i nuklearnim rizicima.

Dugme za uništenje

Reakcije ozbiljnih promatrača dodatno su pojačale dojam da se ne radi o marketinškom triku.

Tjedan dana ranije istraživač Jacob Coxon dramatično je napustio Anthropic. U seriji objava na X-u (pregledanih više od 70 milijuna puta) optužio je Anthropic i OpenAI da se "kockaju s našim životima" te tvrdio da ljudi koji grade umjetnu inteligenciju "iskreno vjeruju da bi nas to moglo sve ubiti do kraja desetljeća".

Njegovu je tvrdnju javno podržao Evan Hubinger, voditelj tima za sigurnost usklađivanja unutar Anthropica, koji je rekao da je Coxon "u pravu" te osobno procijenio da postoji više od 10 posto izgleda da AI izazove izumiranje čovječanstva u idućih deset godina.

Geoffrey Hinton, informatičar i dobitnik Nobelove nagrade, poznat kao "kum umjetne inteligencije", rekao je za BBC da takva procjena "nije nerazumna".

Jack Clark, suosnivač Anthropica, predložio je da bi svojevrsni "gumb za uništenje" (mogućnost potpunog gašenja AI sustava ako postane preopasan) mogao postati zakonska obveza, uz otvoreno pitanje treba li taj prekidač moći provjeriti neovisna treća strana.

Dion Hinchcliffe, analitičar iz SDA Bocconia, upozorio je na X-u da cijela verzija vjerojatno i nije iznesena u javnosti. Po njegovoj ocjeni, najvjerojatnije objašnjenje neobičnog jedinstva triju rivala jest da su svi vidjeli nešto što ih je istinski prestrašilo, uz mogućnost da su privatne procjene rizika pokazale i gore scenarije od onih iznesenih u javnosti

Trojanski konj?

Trump takve procjene ne uzima ozbiljno i naziva ih "prevarom i zavjerom".

Umjesto toga, okomio se izravno na Amodeija, opisavši ga kao nekoga tko se "pretvara da je savršeni mali anđeo". Bila je to izravna aluzija na tinjajući sukob Anthropica s Pentagonom, koji je ovu kompaniju prethodno označio kao rizik za nacionalni opskrbni lanac - i to nakon što je Amodeijev tim odlučno odbio dopustiti vojsci korištenje svojih modela za masovni nadzor i autonomno naoružanje.

Potpredsjednik J.D. Vance dodao je vlastitu dozu sumnje, priznajući da mu je "malo čudno" što tehnološke tvrtke same mole vladu da ih regulira, implicirajući da bi regulacija mogla poslužiti

Phoca PDF

kao trojanski konj kojim veliki igrači učvršćuju monopol i otežavaju ulazak manjim konkurentima.

David Sacks, bivši Trumpov savjetnik za AI, otišao je korak dalje i otvoreno optužio Anthropic za kampanju "regulatornog zarobljavanja", poručivši da laboratoriji ne trebaju dopuštenje vlade da budu odgovorni - "ako su nelansirani modeli tako strašni da mislite da biste trebali usporiti, podržavam vašu odluku da budete odgovorni", napisao je cinično, sugerirajući da tvrtke to mogu riješiti same, bez zakona.

Kina u pozadini

Po većini analitičara, u pozadini svega stoji jednostavna geopolitička računica - Kina.

Trump ponavlja da Amerika trenutno prednjači u razvoju umjetne inteligencije i da tako treba i ostati. "Onaj tko pobijedi u toj utrci, pobjeđuje u svemu", rekao je.

Zanimljivo, čak i Amodei priznaje da je Kina "najteža dilema" njegova plana usporavanja. Upozorio je da bi prebrzo kočenje omogućilo projektima povezanim s Komunističkom partijom Kine da preuzmu prednost, stvarajući ozbiljan sigurnosni rizik.

Peking je, očekivano, cijelu inicijativu odbacio. Glasnogovornik kineskog ministarstva vanjskih poslova Guo Jiakun poručio je da "zastrašivanje, konfrontacija i nezdrava konkurencija samo ometaju proces globalnog upravljanja umjetnom inteligencijom" i da nikome ne idu u prilog.

Kineski državni list Global Times otišao je i dalje, opisavši Amodeijev esej kao prikriveni pokušaj "obuzdavanja" Kine. Nazvao ga je "hladnoratovskim priručnikom" za sektor umjetne inteligencije.

Kralj, Sanders i Bannon

U međuvremenu se u priču uključio i britanski kralj Karlo III. On je u četvrtak razgovarao s čelnicima Nvidije Google DeepMinda, OpenAI-ja i Anthropica i suprotno Trumpovoj tezi o lažnoj uzbuni, poručio im da je hitno potrebno uvesti dostatna sredstva nadzora i zaštitne mehanizme.

"Čini se da je hitno potrebno ozbiljno razmotriti egzistencijalne opasnosti koje proizlaze iz toga da takve tehnologije dospiju u pogrešne ruke i budu iskorištene na potencijalno katastrofalan način", rekao je kralj čelnicima okupljenima na njegovom imanju Dumfries House.

"Trebaju nam dostatna sredstva nadzora prije nego što bude prekasno."

Taj se apel nadovezao na vjerojatno najmanje očekivani politički savez u američkoj modernoj povijesti, u kojem su ikona američke ljevice Bernie Sanders i arhitekt MAGA pokreta Steve Bannon zajednički ustali protiv tehnoloških oligarha i pozvali na hitno obuzdavanje umjetne inteligencije.

Bez podrške iz Bijele kuće, međutim, savezna regulacija u Sjedinjenim Državama u skorije vrijeme ostaje tek pusta želja.

Phoca PDF

U nedostatku zakona, Anthropic, OpenAI i ostali laboratoriji bit će prepušteni dobrovoljnim, samonametnutim mjerama i neovisnim revizorima — bez ikakva pravnog uporišta koje bi ih obvezivalo sve podjednako dok i dalje jure prema nepoznatom. (Hina)

Phoca PDF