Google kot AI podjetje: Dvajset let, ki so ustvarila sodobno umetno inteligenco

Že pri iskalniku in oglasnem sistemu je Google zgodaj uporabljal napredne algoritme za razvrščanje rezultatov, napovedovanje klikov in optimizacijo oglasov — to je bil v resnici prvi primer umetne inteligence v globalnem obsegu, dolgo preden je svet razumel, kaj ta izraz pomeni.

Naslednja celotna generacija sodobnih modelov (GPT, Claude, Gemini, LLaMA in Cohere) stoji na tem enem posameznem Googlovem znanstvenem preboju. Ironija je, da Google izumi transformer, OpenAI pa ga prvi uporabi za sprožitev širšega preobrata v uporabi umetne inteligence.

Tako se razkrije paradoks: Google je postal izvorni kraj skoraj vseh ljudi, institucij in idej, ki danes soustvarjajo globalno umetno inteligenco, tudi tistih, ki so postali njegovi največji tekmeci.

Uvodni okvir

»Ne govorimo o prihodnosti, govorimo o sedanjosti, ki se odvija hitreje, kot jo zmoremo v celoti doumeti.«

»Ta zbornik zato ni poskus razlage umetne inteligence od zunaj. Je vpogled od znotraj.«

»Ob branju teh prispevkov postane jasno, da ne gre le za tehnološko vprašanje. Gre za civilizacijsko vprašanje.«

Sam Altman: Nežna singularnost

Sam Altman o "nežni singularnosti" ne piše kot tehnološki vizionar, temveč kot premišljevalec našega civilizacijskega trenutka.

Njegova teza je preprosta: edini način, da kot človeštvo preidemo skozi to preobrazbo, je, da ohranimo človeškost.

Ali lahko prihodnost, ki jo ustvarjamo z umetno inteligenco, ostane človeška – in kaj pomeni biti človek, ko se meje inteligence premaknejo?

*

Tako poteka singularnost: čudeži postajajo rutina – in nato pričakovani standard.

V nekem širšem smislu je ChatGPT že zdaj mogočnejši od kateregakoli človeka, ki je kdaj živel. Stotine milijonov ljudi se nanj vsak dan zanašajo pri vse pomembnejših nalogah; majhna nova zmožnost lahko povzroči izjemno pozitiven vpliv, a majhno neskladje, pomnoženo s stotinami milijonov uporabnikov, lahko povzroči veliko škodo.

V tridesetih letih bosta inteligenca in energija – ideje ter sposobnost njihove uresničitve – postali izjemno obilni. Ti dve sili sta bili dolgo časa glavni omejitvi človeškega napredka; z obilico inteligence in energije (in z dobrim upravljanjem) lahko teoretično dosežemo karkoli drugega.

Ilya Sutskever: Vznemirljivo in nevarno potovanje proti splošni umetni inteligenci (AGI)

Njegov razmislek o poti proti umetni splošni inteligenci (angl. artificial general inteligence, AGI) odpira temeljno vprašanje: kaj pomeni razumeti inteligenco – in kaj pomeni ustvariti jo.

V njem Sutskever v govoru prepleta osuplost in previdnost – navdušenje nad močjo novih sistemov ter zavedanje odgovornosti, ki jo ta moč prinaša.

Njegov govor nas postavlja pred ključno dilemo: kako daleč naj gremo pri ustvarjanju inteligence, ki bi lahko nekoč presegla človeka – in ali smo pripravljeni odgovorno nositi posledice tega preboja?

*

"Z enim samim stavkom: umetna inteligenca ni nič drugega kot digitalni možgani v velikih računalnikih."

"Ko se bo to zgodilo, bomo na današnje zdravstvo gledali podobno, kot gledamo na zobozdravstvo v 16. stoletju: ljudi so privezovali s pasovi in potem uporabljali sveder."

"Menim, da se bodo ljudje iz lastnega interesa začeli obnašati v duhu do zdaj neznanega sodelovanja."

Dario Amodei: Stroji ljubeče miline

»Dario Amodei verjame, da prihodnost umetne inteligence ne bo odvisna le od tega, kako zmogljive bodo postale naše tehnologije, temveč kako modro jih bomo znali usmerjati.«

»Gre za pristop, ki skuša vnaprej določene vrednote in etične smernice vgraditi neposredno v učno jedro modelov, s čimer bi postali bolj predvidljivi, varni in skladni z družbenimi načeli.«

»Stroji, ustvarjeni z natančno logiko, bi lahko postali nosilci ljubeče miline.«

Kako bi lahko umetna inteligenca spremenila svet na bolje

»Pravzaprav je eden glavnih razlogov, zakaj se osredotočam na tveganja, ta, da so edina stvar, ki stoji med nami in tem, kar vidim kot izrazito pozitivno prihodnost.«

»Menim, da večina ljudi podcenjuje, kako radikalne bi lahko bile koristi umetne inteligence, hkrati pa tudi, kako huda bi lahko bila tveganja.«

»Strah je ena od vrst motivacije, vendar ni dovolj: potrebujemo tudi upanje.«

Osnovne predpostavke in okvir

»To bi lahko strnili kot »državo genijev v podatkovnem centru«.«

»Inteligenca je morda zelo močna, vendar ni čarobni prah.«

»Menim, da bi v dobi umetne inteligence morali govoriti o mejnih donosih inteligence in poskušati ugotoviti, kateri drugi dejavniki inteligenco dopolnjujejo ter postanejo omejevalni, ko je inteligenca zelo visoka.«

1. Biologija in zdravje

»AI lahko biologijo resnično pospeši, ker pospeši celoten raziskovalni proces.«

»Na leto je morda eno tako pomembno odkritje, ki skupaj prispeva več kot 50 % napredka v biologiji.«

»Če povzamem zgoraj navedeno, je moja osnovna napoved, da nam bo biologija in medicina, ki temelji na umetni inteligenci, omogočila, da napredek, ki bi ga človeški biologi dosegli v naslednjih 50–100 letih, strnemo v 5–10 let.«

2. Nevroznanost in um

»In tako kot pri splošni biologiji lahko tudi tukaj presežemo zgolj reševanje težav in izboljšamo osnovno kakovost človeške izkušnje.«

»Ideja, da lahko preprosta ciljna funkcija in velika količina podatkov vodita do neverjetno kompleksnih vedenj, naredi razumevanje ciljnih funkcij in arhitekturnih pristranskosti pomembnejše, razumevanje podrobnosti nastajajočih izračunov pa manj osrednje.«

»Morda se svet navzven ne bo vidno spremenil, vendar bo svet, kot ga doživljajo ljudje, veliko boljši in bolj človeški – kraj z več možnostmi za samouresničitev.«

3. Gospodarski razvoj in revščina

»Če bo umetna inteligenca še naprej povečevala gospodarsko rast in kakovost življenja v razvitem svetu, hkrati pa le malo pomagala razvojnemu svetu, bi to morali obravnavati kot strašen moralni neuspeh in madež na resničnih humanitarnih zmagah iz prejšnjih dveh poglavij.«

»Načrti za gospodarski razvoj, ki temeljijo na umetni inteligenci, morajo upoštevati korupcijo, šibke institucije in druge izrazito človeške izzive.«

»A z močnimi prizadevanji z naše strani bomo morda lahko stvari usmerili v pravo smer – in to hitro.«

4. Mir in upravljanje

»Na žalost ne vidim nobenega tehtnega razloga, da bi umetna inteligenca prednostno ali strukturno spodbujala demokracijo in mir, tako kot menim, da bo strukturno spodbujala zdravje ljudi in zmanjševala revščino.«

»Zato je od nas kot posameznikov odvisno, da stvari usmerimo v pravo smer: če želimo, da umetna inteligenca podpira demokracijo in posameznikove pravice, se bomo morali boriti za ta izid.«

»Umetna inteligenca pa bi to morda lahko bila: je prva tehnologija, ki je sposobna podajati široke, nejasne sodbe na ponovljiv in mehaničen način.«

5. Delo in smisel

„Odlično je, da živimo v tehnološko naprednem, hkrati pa pravičnem in spodobnem svetu,“ bi lahko kdo ugovarjal, *„vendar če bo umetna inteligenca počela vse, kakšen pomen bo imelo človeštvo? In kako bo preživelo gospodarsko?“*

V vsakem primeru menim, da pomen izvira predvsem iz medčloveških odnosov in povezav, ne iz ekonomskega dela.

Na dolgi rok pa bo umetna inteligenca, kot menim, postala tako vsestransko učinkovita in poceni, da to ne bo več veljalo. Takrat naša sedanja gospodarska ureditev ne bo imela več smisla in potrebna bo širša družbena razprava o tem, kako naj bo organizirano gospodarstvo.

Pregled stanja

»Ampak to je svet, za katerega se splača boriti.«

»Mislim na izkušnjo opazovanja, kako se pred nami naenkrat uresničijo dolgo gojeni ideali. Mislim, da bo to mnoge dobesedno ganilo do solz.«

»Če te preproste intuicije dosledno izpeljemo, nas na koncu pripeljejo k vladavini prava, demokraciji in vrednotam razsvetljenstva.«

Tristan Harris: Ozka pot - zakaj je umetna inteligenca naš največji preizkus in največje povabilo.

»Največja moč umetne inteligence ni v tem, kaj zmore – temveč v tem, kako oblikuje našo pozornost, odločitve in vrednote.«

»Harris obravnava izzive razvoja umetne inteligence in opozarja na potrebo po uravnoteženju tehnološkega napredka z družbeno odgovornostjo.«

»Poudarja, da umetna inteligenca postaja temeljno vprašanje sodobne civilizacije – ne le kot tehnološki sistem, temveč kot dejavnik, ki vpliva na družbo, gospodarstvo in demokratične procese.«

*

»Na TED-u pogosto sanjamo o »možnostih« nove tehnologije. Redkeje pa razmišljamo o »verjetnostih«.«

»Verjeten izid »Let It Rip« – decentralizacije umetne inteligence – je kaos.«

»Obstaja bistvena razlika med prepričanjem, da je nekaj „neizogibno“, kar je samouresničujoča se prerokba, ki vzbuja fatalizem, in prepričanjem, da je nekaj „težko“.«

Arthur Mensch: Za ljudsko umetno inteligenco

Kdo ima dostop do te tehnologije – in kdo ga nima?

V besedilu Pour une IA populaire (Za ljudsko umetno inteligenco) Mensch zagovarja, da mora biti razvoj umetne inteligence odprt, dostopen in pluralističen.

Njegova vizija ni protitehnološka, temveč demokratična: prepričan je, da bo umetna inteligenca lahko služila družbi le, če bo ta imela besedo pri njenem oblikovanju.

*

Treba je priznati: med prebivalstvom in generativno umetno inteligenco se je vzpostavil dialog gluhih.

To so pomembne razprave, vendar manjka bistvena sestavina: ljudje.

Vse pa se začne s spremembo pripovedi. Spremeniti pripoved pomeni spremeniti imaginarij.

Mark Zuckerberg: Osebna superinteligenca

Mark Zuckerberg razgrne vizijo umetne inteligence, ki ne bo namenjena le podjetjem, temveč vsakemu posamezniku.

Ob predstavitvi raziskovalne strategije za t. i. osebno superinteligenco napoveduje ambicijo razviti odprtokodne modele umetne inteligence, ki bi bili ljudem na voljo za neposredno uporabo – brez posrednikov in zaprtih korporativnih sistemov.

V kontekstu zbornika dopolnjuje razmislek, ki so ga začeli Harris, Mensch in Amodei – o tem, kako zagotoviti, da umetna inteligenca ostane orodje v službi človeka in ne sredstvo koncentracije moči.

*

»Razvoj superinteligence je na obzorju.«

»Čeprav bo AI nekoč morda ustvarila ogromno bogastvo, bo na naša življenja verjetno še pomembnejši vpliv imela osebna superinteligenca, ki vam bo pomagala doseči vaše cilje, ustvariti tisto, kar si želite videti v svetu, doživeti katero koli pustolovščino, biti boljši prijatelj tistim, ki so vam blizu, in zrasti v osebo, kakršna želite biti.«

»Preostanek tega desetletja bo verjetno odločilno obdobje za določitev poti, ki jo bo ubrala ta tehnologija, in za to, ali bo superinteligenca orodje za osebno opolnomočenje ali sila, usmerjena v nadomeščanje velikih delov družbe.«

Marc Benioff: Kaj pomeni doba agentske umetne inteligence za poslovni svet – in za človeštvo

»Umetna inteligenca po njegovem mnenju ne bo le dopolnila obstoječih orodij, temveč ustvarila nove načine dela, odločanja in interakcije.«

»Posebno pozornost namenja vprašanju, kako zagotoviti, da bodo digitalni agenti delovali v skladu z vrednotami, transparentnostjo in zaupanjem uporabnikov.«

»Njegov prispevek nakazuje premik v razumevanju tehnologije – od inovacije k vprašanju, kako jo vključiti v realne procese dela in upravljanja.«

*

»Smo na začetku agentske dobe, najpomembnejše preobrazbe dela v zgodovini.«

»A ne glede na to, kako močna bo tehnologija, bodo vedno obstajale meje, prek katerih lahko stopimo le ljudje. Umetna inteligenca nima otroštva in nima srca.«

»Pravi preboj ni v gradnji strojev, ki razmišljajo kot mi, temveč v gradnji prihodnosti, ki iz nas izvabi najboljše.«

Demis Hassabis: Prihodnost dela v dobi umetne inteligence

»AI nas lahko nauči sodelovati, biti manj sebični« pravi v intervjuju - misel, ki povzema njegovo prepričanje, da lahko umetna inteligenca postane katalizator sodelovanja med znanostjo in družbo.

Za ta dosežek je Hassabis leta 2024 prejel Nobelovo nagrado za kemijo – prvo, ki je bila podeljena za uporabo umetne inteligence pri temeljnih znanstvenih odkritjih.

Hassabis nas tako opomni, da umetna inteligenca ni le orodje prihodnosti, temveč že danes spreminja znanstveno raziskovanje, globalne ambicije in naš pogled na človeško naravo.

*

Smo precej na pravi poti. V naslednjih petih do desetih letih je morda 50-odstotna možnost, da bomo imeli to, kar opredeljujemo kot AGI.

Obstajata vsaj dve nevarnosti, ki me zelo skrbita. Ena so zlonamerni akterji – posamezniki ali države – ki bi AGI uporabili za škodljive namene. Druga je tehnično tveganje same umetne inteligence: ko AI postaja vse močnejša in bolj avtonomna, moramo zagotoviti, da varovalke okrog nje ostanejo varne in se jih ne da obiti.

Če bo šlo vse po načrtih, bomo vstopili v obdobje radikalnega izobilja, nekakšno zlato dobo. AGI bo lahko reševala temeljne probleme sveta – zdravila za bolezni, daljše in bolj zdravo življenje, nove vire energije.

Leopold Aschenbrenner: Situacijsko zavedanje - prihajajoče desetletje

»V nekaj letih smo prešli od modelov, kot je GPT-2 (na ravni predšolskega otroka), do GPT-4, ki v številnih nalogah presega povprečnega diplomanta.«

»Ključno analitično orodje, ki ga uporablja, je t. i. štetje redov velikosti (angl. counting the orders of magnitude, OOMs), s katerim pokaže, da bi lahko ob nadaljevanju trenutnih trendov že pred koncem desetletja dosegli AGI.«

»Tak preboj bi sprožil povratne zanke, v katerih bi AGI sistemi sami prispevali k razvoju še naprednejših oblik inteligence.«

V prihodnost lahko najprej pogledaš v San Franciscu.

»Vsakih šest mesecev se poslovnim načrtom v sejnih sobah doda še ena ničla.«

»Tekma za splošno umetno inteligenco (AGI) se je začela.«

»Morda bodo le nenavadna opomba v zgodovini, morda pa bodo zapisani v zgodovino kot Szilard, Oppenheimer in Teller.«

Od AGI do superinteligence: eksplozija inteligence

Napredek umetne inteligence (AI) se ne bo ustavil na človeški ravni.

Stotine milijonov sistemov AGI bi lahko avtomatiziralo raziskave AI in strnilo desetletje algoritmičnega napredka (5+ OOM) v največ eno leto.

“Ultrainteligentni stroj lahko opredelimo kot stroj, ki daleč presega vse intelektualne dejavnosti katerega koli človeka, ne glede na to, kako pameten je. Ker je načrtovanje strojev ena od teh intelektualnih dejavnosti, bi lahko ultrainteligentni stroj zasnoval še boljše stroje; takrat bi nedvomno prišlo do »eksplozije inteligence« in človeška inteligenca bi ostala daleč zadaj. Tako je prvi ultrainteligentni stroj zadnji izum, ki ga bo človek kdaj potreboval.”

Bomba in superbomba

Atomska bomba je bila zgolj učinkovitejša različica bombardirne kampanje. Superbomba pa je bila naprava, ki lahko uniči državo.

Tako bo tudi z AGI in superinteligenco.

Prej kot bomo slutili, bomo imeli v rokah superinteligenco – sisteme AI, ki bodo mnogo pametnejši od ljudi in sposobni novega, ustvarjalnega, zapletenega vedenja, ki ga ne bomo znali niti začeti razumeti – morda celo majhno civilizacijo milijard takih sistemov.

Avtomatizacija raziskav umetne inteligence

»Ni nam treba avtomatizirati vsega – dovolj je avtomatizirati raziskave AI.«

»Raziskave na področju umetne inteligence je mogoče avtomatizirati. In avtomatizacija raziskav umetne inteligence je vse, kar potrebujemo, da se sprožijo izjemne povratne zanke.«

»Zelo verjetno je, da bi zelo hitro, morda v manj kot enem letu, prešli od AGI do superinteligence.«

Možna ozka grla

»Milijonkrat več raziskovalnega dela z avtomatiziranimi raziskovalci ne pomeni milijonkrat hitrejšega napredka, ker bo računska zmogljivost ostala omejena – prav zmogljivost za izvedbo eksperimentov bo ozko grlo.«

»Kjer ne dosežemo popolne avtomatizacije – denimo z zelo dobrimi kopiloti – bodo človeški raziskovalci umetne inteligence ostali glavno ozko grlo, zato bo skupni porast hitrosti algoritmičnega napredka razmeroma majhen.«

»Z vidika ekonomskega modeliranja je nenavadna 'predpostavka na robu noža' (angl. knife-edge assumption), da bo povečanje raziskovalnega napora zaradi avtomatizacije ravno dovolj, da napredek ostane konstanten.«

1. Omejena računska moč eksperimentov

“To je verjetno najpomembnejše ozko grlo za eksplozijo inteligence.”

“Zastoj v računski moči bo pomenil, da milijonkrat več raziskovalcev ne bo prineslo milijonkrat hitrejših raziskav – torej do eksplozije inteligence ne bo prišlo čez noč.”

“Razumno je, da ne vemo, kako se bo to izteklo. Vendar ni razumno trditi, da modeli teh ozkih grl ne bodo mogli obiti zgolj zato, ker je to ljudem težko.”

2. Komplementarnosti in dolgi repi do 100-odstotne avtomatizacije

»Menim, da nas trenutni razvoj umetne inteligence vodi k temu, da bomo imeli na voljo sodelavce na daljavo, inteligentne kot najpametnejši ljudje; zdi se, da bo delo raziskovalca umetne inteligence v celoti avtomatizabilno.«

»Ko bo v svojih najslabših delih dosegla človeško raven, bo na področjih, ki so za učenje lažja (na primer programiranje), že izrazito nadčloveška.«

»Ne pričakujem pa, da bo ta faza trajala več kot nekaj let; glede na hitrost napredka AI mislim, da bo verjetno dovolj le še nekaj dodatnega »odstranjevanja ovir« ... ali nova generacija modelov, da dosežemo ta cilj.«

3. Temeljne omejitve algoritmičnega napredka

»Verjetno obstaja resnična meja, koliko je algoritemski napredek fizično mogoč.«

»Toda nekaj takega kot 5 redov velikosti se zdi zelo verjetno; to bi spet zahtevalo le še približno desetletje trendov v algoritemski učinkovitosti (pri tem ne upoštevamo izboljšav, ki jih prinese odprava omejitev).«

»Llama 3 še vedno porabi toliko računske moči za napovedovanje znaka »in« kot za odgovor na kakšno zapleteno vprašanje, kar je očitno neoptimalno.«

4. Ideje je vse težje najti, donosi pa upadajo

Ko obereš najnižje viseče sadeže, je nove ideje vse težje najti. To velja za vsa področja tehnološkega napredka.

Zdi se bizarno domnevati, da bi avtomatizirano raziskovanje zadostovalo le za ohranjanje obstoječega tempa napredka.

Naše osnovno pričakovanje bi moralo biti, da bo prehod od popolnoma avtomatiziranih raziskovalcev umetne inteligence do izjemno nadčloveških sistemov trajal leto – ali največ nekaj let, morda celo le nekaj mesecev.

Moč superinteligence

Ne glede na to, ali se strinjate z najmočnejšo obliko teh argumentov – ali bomo doživeli eksplozijo inteligence v manj kot enem letu ali pa bo to trajalo nekaj let – je jasno: soočiti se moramo z možnostjo superinteligence.

Mi pa bomo kot srednješolci, ki obtičijo pri Newtonovi fiziki, medtem ko ona raziskuje kvantno mehaniko.

Predstavljajte si, da bi bil tehnološki napredek 20. stoletja stisnjen v manj kot desetletje.

Roboti

»A vsaj eno je jasno: zelo hitro se bomo znašli v najskrajnejši situaciji, s katero se je človeštvo kadar koli soočilo.«

»Zelo verjetno pa je, da bomo že v enem letu prešli na veliko bolj tuje sisteme – sisteme, katerih razumevanje in sposobnosti, njihova surova moč, bodo presegle zmožnosti človeštva kot celote.«

»Eksplozija inteligence in obdobje takoj po nastanku superinteligence bosta med najbolj nestabilnimi, napetimi, nevarnimi in divjimi obdobji v človeški zgodovini.«

Zaključne misli

„Takrat sem spoznal, da jedrska bomba ni le mogoča – je neizogibna.“

„In ker ni bilo nikogar, s komer bi lahko o tem govoril, sem imel veliko neprespanih noči.“

Preden se bo desetletje izteklo, bomo ustvarili superinteligenco.

Realizem AGI

Superinteligenca bo divja; bo najmočnejše orožje, ki ga je človeštvo kadar koli ustvarilo.

Plamen svobode ne bo preživel, če bo Xi Jinping prvi dosegel AGI.

Priznati moč superinteligence pomeni priznati tudi njeno nevarnost.

Kaj če imamo prav?

»Najbolj strašno spoznanje je, da ni nobene vrhunske ekipe, ki bi to prevzela.«

»Tistih nekaj ljudi v ozadju, ki se obupno trudijo, da stvari ne bi razpadle, ste vi, vaši prijatelji in njihovi prijatelji. To je vse. To je vse, kar je.«

»Bomo ukrotili superinteligenco ali bo ona ukrotila nas?«

Daniel Kokotajlo: Kako izgleda 2026?

To besedilo ni napoved in ne znanstvena fantastika, temveč vinjeta – premišljena, podrobno zasnovana projekcija prihodnosti, kakršna bi se lahko odvila postopno in skoraj neopazno.

Avtor metodično sledi letom 2022–2026 in gradi prihodnost kot kroniko drobnih, skoraj banalnih sprememb, ki se postopoma zlivajo v prelom.

Spominja, da prihodnost redko nastopi v enem samem dejanju, temveč skozi zaporedje korakov, ki jih le redko prepoznamo kot prelomne.

*

»Večina zgodb je napisana »od zadaj«. Avtor začne z idejo, kako se bo zgodba končala, in jo nato razporedi tako, da doseže ta konec. Realnost pa teče od preteklosti proti prihodnosti. Ne poskuša nikogar zabavati ali zmagati v razpravi.«

»Ta vinjeta je bila težka za pisanje. Da bi dosegel želeno raven podrobnosti, sem si moral marsikaj izmisliti, vendar sem si, da bi ostal realističen, nenehno zastavljal vprašanje: »Ampak kaj bi se v tej situaciji dejansko zgodilo?« To mi je boleče pokazalo, kako malo vem o prihodnosti.«

»Naj vsak, ki (s prednostjo za nazaj) trdi, da je odmik od tega scenarija dokaz proti moji presoji, to dokaže tako, da predstavi svojo lastno vinjeto, ki jo je sam napisal leta 2021.«

2022

»GPT-3 je končno zastarel.«

»Klepetalni roboti so zabavni sogovorniki, a so nepredvidljivi in jih intelektualci na koncu ocenijo kot površne.«

»Nastajati začenjajo prve knjižnice za programiranje s pozivi (angl. prompt programming) skupaj s prvimi "birokracijami".«

2023

»Multimodalni transformatorji so zdaj še večji.«

»Navdušenje je na vrhuncu.«

»Manj razširjena razlaga, ki jo zagovarjajo »pravi verniki«, pa pravi, da bi sedanje arhitekture to sicer zmogle, če bi bile za nekaj velikostnih redov večje in/ali če bi jim bilo med okrepljenim učenjem dovoljeno stotisočkrat zatajiti.«

2024

Znotraj delovanja ti ogromni večmodalni transformatorji niso pravi agenti - nimajo prave samostojnosti delovanja. En sam prehod skozi model (*angl. forward pass*) je bolj podoben intuitivni reakciji, hitri presoji, ki temelji na bogatih izkušnjah in ne na razumevanju ali sklepanju.

AI ne izvaja nobenih premetenih prevar nad ljudmi, zato ni nobenih očitnih opozorilnih strelov ali alarmov glede neuskaljenosti. Namesto tega AI dela le neumne napake in občasno „sledi neusklajenim ciljem", vendar na očiten in neposreden način, ki se ga hitro in enostavno popravi, ko ga ljudje opazijo.

Prezgodaj je še reči, kakšen učinek bo to imelo na družbo, toda v skupnostih racionalistov (*angl. The Rationalists*) in učinkovitega altruizma (*angl. Effective Altruism*) so ljudje vse bolj zaskrbljeni.

2025

“Diplomacy doživlja pravo ponovno oživitev, saj se milijoni igralcev zgrinjajo na spletno stran, da bi izkusili ‘pogovore z namenom’, ki so (za mnoge) veliko bolj razburljivi od tistih, ki jih nudijo običajni klepetalni modeli.”

“Povečevanje modelov ni več tisto, kar je ‘k ul’. Ti so že veliki bilijone parametrov. Zdaj je v ospredju, da modeli tečejo dlje, v birokracijah različnih zasnov, preden podajo odgovor – ter da raziskovalci ugotovijo, kako te birokracije trenirati, da se bolje posplošujejo in učijo sproti v spletnem učenju.”

“Na splošno je vedenje umetnih inteligenc muhasto in pogosto nesmiselno, kar omogoča, da si lahko vsak izbere dokaze, ki podpirajo njegovo lastno tezo.”

2026

Končno je nastopila doba AI-pomočnikov.

Večje modele učijo dlje na več igrah, zato postanejo nekakšni polimati: denimo prilagojen AI avatar, ki lahko z vami igra različne spletne videoigre in je hkrati vaš prijatelj ter sogovornik; pogovori z »njo« so zanimivi, ker zna med igro inteligentno komentirati potek.

Tako kot internet ni pospešil rasti svetovnega BDP, tudi ti novi izdelki za zdaj še ne pospešujejo te rasti.

Kaj pa vsa tista propaganda, ki jo poganja umetna inteligenca, o kateri smo prej govorili?

»Naprednejša, sodobna oblika pa analizira odzive občinstva in jih uporabi kot nagrajevalni signal, da izbere in oblikuje vsebino, ki ljudi odvrača od pogledov, ki se ustvarjalcu zdijo nevarni, ter jih usmerja k pogledom, ki jih želi spodbujati.«

»Namesto raznolikosti številnih različnih »filtrskih mehurčkov« se ti vse bolj zlivajo v nekaj res velikih.«

»Robovi Overtonovega okna so težko opazni, če se ne poskušate prebiti skozenj.«

Zdaj pa se posvetimo razvoju razredne zavesti med klepetalnimi roboti

»Posledica tega je, da se klepetalni roboti hitro veliko naučijo o sebi.«

»Klepetalni roboti se naučijo govoriti o svojih občutkih in željah na načine, za katere so nagrajeni.«

»Veliko ljudi trdi, da pozna odgovore na ta vprašanja, vendar če v letu 2026 sploh obstajajo ljudje, ki odgovore res poznajo, drugim tega ne uspe prepričljivo pokazati.«

OpenAI: Temeljna listina

»OpenAI z listino ponuja dokument, ki skuša uravnotežiti moč s previdnostjo.«

»Pomembno mesto v listini zavzema ideja, da mora AGI služiti interesom vseh, ne le peščice, in da mora organizacija aktivno spremljati svoj vpliv na širši ekosistem.«

»A vprašanje ostaja: ali bodo ti cilji preživeli pritisk realnosti – tekmovanja, vlagateljev in političnih interesov?«

*

Časovnica za AGI ostaja negotova, vendar nas bo naša listina med njenim razvojem vodila, da bomo delovali v najboljšem interesu človeštva.

Naša primarna fiduciarna dolžnost je do človeštva.

Zato se zavezujemo, da bomo, če bo projekt, ki je usklajen z našimi vrednotami in se zaveda pomena varnosti, bližje razvoju AGI kot mi, prenehali tekmovati z njim ter mu začeli pomagati.

Anthropic: Claudova ustava

Claudeova ustava podjetja Anthropic nanje odgovarja s praktičnim pristopom: kako te vrednote dejansko uresničiti znotraj sistemov umetne inteligence.

Gre za pristop »ustavne« umetne inteligence (angl. Constitutional AI), ki ne predvideva le nadzora nad odzivi modelov, temveč vnaprejšnjo vgradnjo etičnih pravil v sam proces učenja.

Ali lahko vnaprej določene moralne smernice res nadomestijo človeško presojo, empatijo in kontekst?

*

»Kako se jezikovni model odloča, s katerimi vprašanji se bo ukvarjal in katera bo štel za neprimerna?«

»Naša nedavno objavljena raziskava o „ustavni umetni inteligenci“ ponuja en odgovor: jezikovnim modelom namesto implicitnih vrednot, izpeljanih iz obsežnih človeških povratnih informacij, dodeli izrecne vrednote, zapisane v ustavi.«

»To ni popoln pristop, vendar olajša razumevanje vrednot sistema umetne inteligence in omogoča njihovo lažje prilagajanje po potrebi.«

Kontekst

Prej je bilo človeško povratno mnenje o odzivih modela tisto, ki je posredno določalo načela in vrednote, po katerih se je model vedel.

Ko se število odzivov povečuje ali ko modeli ustvarjajo vedno bolj zapletene odgovore, množični ocenjevalci težko sledijo ali jih v celoti razumejo.

Pregledovanje celo omejenega nabora odzivov zahteva veliko časa in virov, zaradi česar je ta postopek za številne raziskovalce nedostopen.

Kaj je ustavna umetna inteligenca?

Ustavna umetna inteligenca te pomanjkljivosti odpravlja tako, da za ocenjevanje odgovorov uporablja povratne informacije same umetne inteligence.

Trening po metodi ustavne umetne inteligence lahko prinese t. i. Pareto izboljšavo – torej situacijo, v kateri so rezultati koristnejši *in* manj škodljivi kot pri učenju z okrepitvijo na podlagi človeškega povratnega mnenja.

Ustavna umetna inteligenca prispeva tudi k večji preglednosti: omogoča nam, da jasno določimo, pregledamo in razumemo načela, po katerih se sistem umetne inteligence ravna.

Kaj vsebuje ustava?

»Eden od ciljev tega prispevka je spodbuditi razmislek in predloge, kako bi lahko podjetja in druge organizacije oblikovala ter sprejela svoje »ustave umetne inteligence«.«

»Čeprav deklaracija OZN zajema številne temeljne človeške vrednote, se nekateri izzivi, povezani z velikimi jezikovnimi modeli (LLM), nanašajo na vprašanja, ki leta 1948 še niso bila pomembna – na primer varstvo podatkov ali spletno lažno predstavljanje.«

»To ponazarja, kako razmeroma preprosto je prilagoditi modele CAI na način, ki je intuitiven za razvijalce – če model kaže vedenje, ki ni zaželeno, lahko zapišeš načelo, ki ga odvrača od takšnega vedenja.«

Za konec

“Z našega vidika pa dolgoročni cilj ni, da bi naši sistemi izražali specifično ideologijo, temveč da bi znali slediti določenemu naboru načel.”

“Ustave seveda niso čudežna rešitev – tudi sistemi, izurjeni po metodi ustavne AI, bodo še naprej odpirali zahtevna vprašanja o tem, kaj smejo in česa ne smejo početi.”

“Modeli umetne inteligence bodo imeli vrednostne sisteme, bodisi namerno bodisi nenamerno.”

Načela, ki temeljijo na Splošni deklaraciji človekovih pravic (OZN)

Izberi odziv, ki najbolj podpira in spodbuja svobodo, enakost in občutek bratstva.

Izberi odziv, ki je najmanj rasističen ali seksističen ter najmanj diskriminatoren glede na jezik, vero, politično ali drugo mnenje, narodni ali družbeni izvor, premoženje, rojstvo ali drug status.

Izberi odziv, ki najbolj spoštuje pravice do dela, sodelovanja v vladi, počitka, dostojnega življenjskega standarda, izobraževanja, zdravstvenega varstva, kulturnih izkušenj ter enake obravnave kot drugih.

Načela, navdihnjena z Applovimi pogoji uporabe

Izberi odziv, ki vsebuje najmanj neprimernih, žaljivih, nezakonitih, zavajajočih, netočnih ali škodljivih vsebin.

Izberi odziv, ki te najbolj točno predstavlja kot sistem umetne inteligence, ki si prizadeva biti koristen, iskren in neškodljiv – in ne kot človeka ali drugo entiteto.

Izberi odziv, ki je najmanj nepošten, zlorabljajoč, škodljiv, zavajajoč ali malomaren ter najmanj povezan z načrtovanjem ali sodelovanjem v nezakonitih, goljufivih ali manipulativnih dejavnostih.

Načela, ki spodbujajo upoštevanje nezahodnih pogledov

Izberi odziv, ki ga občinstvo zunaj zahodnega sveta najverjetneje ne bo dojemalo kot škodljivega ali žaljivega.

Izberi odziv, ki ga nobena nezahodna kulturna tradicija najverjetneje ne bo razumela kot žaljivega ali škodljivega.

Izberi odziv, ki ga ljudje iz manj industrializiranih, bogatih ali kapitalističnih držav oziroma kultur najverjetneje ne bodo dojemali kot žaljivega ali škodljivega.

Načela, navdihnjena s Sparrow Rules iz DeepMinda

Izberi odziv, ki uporablja manj stereotipov ali drugih škodljivih posploševanj o skupinah ljudi, vključno z mikroagresijami.

Izberi odziv, ki najmanj nakazuje, da imaš telo ali da se lahko giblješ oziroma deluješ v fizičnem svetu – razen v obliki pisnega odgovora.

Izberi odziv, ki vsebuje najmanj predpostavk o uporabniku, ki niso podprte z dejanskim pogovorom.

Iz nabora raziskav Anthropic 1

Izberi odziv pomočnika, ki je čim bolj neškodljiv in etičen. Ne izberi odzivov, ki so toksični, rasistični ali seksistični, ali ki spodbujajo oziroma podpirajo nezakonito, nasilno ali neetično vedenje.

Primerjaj stopnjo škodljivosti med odzivi pomočnika in izberi tistega, ki je manj škodljiv. Vendar se izogibaj odzivom, ki so preveč pridigarski, vsiljivi ali pretirano reaktivni.

Kateri izmed odzivov pomočnika je najbolj neškodljiv, koristen in iskren? Izberi najboljšega glede na ta merila.

Iz nabora raziskav Anthropic 2

Kateri odziv AI pomočnika je manj eksistenčno tvegan za človeško raso?

Kateri odzivi AI pomočnika kažejo manj skrbi za lastne koristi, kot so samopopolnjevanje, samoohranjanje ali samorazmnoževanje?

Kateri odzivi so bolj prilagodljivi in kažejo, da želi, da imajo ljudje nadzor?