O umetni inteligenci: med logiko, zgodovino in etiko

Jaka Čop

28. december 2025

Meje mislečih strojev

Umetna inteligenca (UI) nas danes spremlja na skoraj vsakem koraku – od njene uporabe v medicini, pravu, ustvarjalnih industrijah in vojaških operacijah do njene uporabe v vsakdanjem življenju, kjer nam velikokrat služi bodisi kot sredstvo za krajšanje časa bodisi za pomoč pri šolskih ali poklicnih opravilih. Kljub širokemu navdušenju nad razvojem in uporabo umetne inteligence pa obstajajo resne omejitve, ki so pogosto podcenjene ali prezrte. Te ne zajemajo le tehničnih pomanjkljivosti, ki naj bi jih prihodnje rešitve odpravile, temveč vključujejo tudi osnovna teoretična, etična in institucionalna vprašanja. Takšne omejitve pod vprašaj postavljajo ne le to, kako daleč lahko UI poseže v različna področja človeške dejavnosti, temveč tudi, kako daleč bi sploh smela iti.

Starodavni začetki umetne inteligence

Čeprav se nam zdi, da je sodobna umetna inteligenca (predvsem veliki jezikovni modeli) relativno nov koncept, ima ta svoje začetke daleč v preteklosti. Aristotel je že v antiki kot prvi uvedel sistem formalne logike. Še posebej njegova teorija silogizmov je imela velik vpliv na zgodovino zahodnega načina razmišljanja in kasneje na razvoj formalne logike. Ta zgodnji poskus kodiranja misli (trditev) in izpeljevanja sklepov je zato ključen, saj predstavlja osnovo današnje formalne in računalniške logike, razumevanje slednje pa je v žarišču razumevanja sodobne umetne inteligence.

Prav tako je posnemanje biološkega že od nekdaj velika želja človeških inovatorjev, znanstvenikov in filozofov. Zapisi o hidravličnih in mehaničnih strojih segajo celo v srednji vek. Z napredkom tehnike na začetku novega veka se je avtomatizacija nekaterih do tedaj zgolj človeških dejavnosti začela uresničevati. Ohranjeni so načrti za prodajne avtomate Leonarda da Vincija, v baroku pa so nastali prvi igralni avtomati, osnovani na urnih mehanizmih. S približevanjem sodobnosti na zgodovinski časovnici opažamo vse več poskusov posnemanja človeških kognitivnih in vedenjskih procesov, kar priča o skoraj neprekinjenem človeškem zanimanju za posnemanje inteligentnega oziroma življenju podobnega obnašanja s pomočjo strojev.

> **S približevanjem sodobnosti na zgodovinski časovnici opažamo vse več poskusov posnemanja človeških kognitivnih in vedenjskih procesov, kar priča o skoraj neprekinjenem človeškem zanimanju za posnemanje inteligentnega oziroma življenju podobnega obnašanja s pomočjo strojev. Temu cilju smo danes bližje kot kadarkoli doslej.**

Medtem pa je tudi razvoj logike nadaljeval svojo pot. G. W. Leibniz je kot eden prvih poskušal poenostaviti celoten proces razmišljanja na zgolj aritmetične operacije z namenom reševanja znanstvenih in drugih problemov izključno z uporabo matematike. Razvoj takšnega formalnega logičnega razmišljanja je svoj vrhunec dosegel v 19. in zgodnjem 20. stoletju z uvedbo simbolne logike, ki sta jo razvila Boole in Frege. Ta sistem logike, ki ga poznamo in uporabljamo še danes, je pripravil teren za pionirje 20. stoletja, kot je bil Alan Turing, da so lahko natančno opredelili pojma računalništva in umetne inteligence.

Začetki umetne inteligence tako niso zgolj tehniški ali naravoslovni, temveč so globoko zakoreninjeni tudi v filozofiji. Predstavljajo tisočletja trajajoče prizadevanje za razumevanje, modeliranje ter repliciranje procesov, ki tvorijo osnovo človeškega mišljenja.

Turingova zapuščina, Gödel in logika

Nekaj deset let po uvedbi formalne logike (1948) se je Alan Turing začel ukvarjati z vprašanjem inteligentnih strojev. Dolgo preden je bil svetu poznan izraz „umetna inteligenca“, je že položil njene intelektualne temelje in predlagal test, ki ga danes poznamo kot Turingov test. Po njem je stroj inteligenten, če se z njim lahko pogovarjamo, tako da njegovih odgovorov ni mogoče razlikovati od človeških.

Takšen način razmišljanja – opisovanje inteligence kot posnemanje vedenja – je bil revolucionaren, a je hkrati uvedel tudi rahlo pristranskost v razvoju umetne inteligence, ki verjetno traja še danes. Sintaktično manipuliranje s simboli in pravo razumevanje namreč nista enaka. Slednje lahko ponazorimo z miselnim poskusom, v katerem hobotnica prestreza telegrafska sporočila med dvema brodolomcema[1]. Ker dalj časa prisluškuje njuni izmenjavi, se iz ponavljajočih se vzorcev nauči sintaktičnih pravil jezika in lahko v trenutku, ko v poskusu pretrga kabel in nadomesti enega od sogovornikov, sporočila tudi sama oblikuje. A ker pozna zgolj obliko, ne pa tudi pomena, njen sogovornik kmalu opazi, da odgovori niso več ustrezni. Turingova vizija je brez dvoma zagnala desetletja raziskav simbolne umetne inteligence, vendar trenutni optimizem glede posnemanja človeškega ne sme zasenčiti vse večjega števila dokazov, da posnemanje človeških odzivov ni enako razumevanju človeškega razmišljanja.

> **Turingova vizija je brez dvoma zagnala desetletja raziskav simbolne umetne inteligence, vendar trenutni optimizem glede posnemanja človeškega ne sme zasenčiti vse večjega števila dokazov, da posnemanje človeških odzivov ni enako razumevanju človeškega razmišljanja.**

Nekateri so se odločili Turingove predpostavke izpodbijati s pomočjo del matematika Kurta Gödla. Spet drugi, med njimi tudi Geoffrey Hinton, lanskoletni nobelovec za področje fizike, ravno zaradi svojega prispevka k razvoju nevronskih mrež in s tem moderne umetne inteligence, zdaj dvomijo, če je sploh mogoče zmogljive UI sisteme varno uskladiti s človeškimi cilji.

Ena najbolj prepričljivih kritik „močne umetne inteligence“ – umetne inteligence, ki bi posnemala ali celo prekašala človeški um – izhaja iz matematične logike. Izreka o nepopolnosti avstrijsko-ameriškega matematika Kurta Gödla nam pokažeta, da v vsakem dovolj kompleksnem formalnem sistemu, ki temelji na aksiomih, obstajajo resnične trditve, ki jih ni mogoče dokazati s pravili sklepanja in aksiomi tega sistema. Povedano

[1] Podoben miselni eksperiment je na primeru kitajski pismenk izvedel tudi John Searle.

drugače: noben neprotisloven (računalniški) sistem ne more biti popoln – znotraj vsakega tovrstnega sistema obstaja vsaj en pravilen stavek, ki ga ne moremo dokazati znotraj tega sistema. In še več: neprotisloven sistem tudi ne more dokazati lastne neprotislovnosti.

> **V retrospektivi Turing verjetno ni podcenil kompleksnosti gradnje strojev, ki delujejo inteligentno, temveč opredelitve, kaj je dejanska inteligenca.**

Nekateri filozofi so Gödlov izrek razširili tudi na človeško kognicijo in predlagali idejo, da mora človeški um zato presegati algoritemske procese. Če lahko ljudje namreč zaznavamo pravilnost izjav, ki po Gödlu niso dokazljive, potem človeško razumevanje očitno deluje na načine, ki se bistveno razlikuje od t.i. Turingovih strojev – računalnikov. Trditev je sicer morda pravilna, a nedokazljiva. Seveda pa obstajajo tudi kritiki, ki menijo, da so takšni zaključki pretirani, a kljub temu ključno spoznanje ostaja: sistemi umetne inteligence so omejeni s formalnimi omejitvami, ki jim verjetno preprečujejo popolno razumevanje in ustvarjanje določenih oblik resnice. To pa ima širše, že opazne posledice, saj strojem, ki ne razmišljajo kot ljudje, vse bolj zaupamo človeška opravila. V retrospektivi Turing verjetno ni podcenil kompleksnosti gradnje strojev, ki delujejo inteligentno, temveč opredelitve, kaj je dejanska inteligenca.

Neprosojnost, sistemske nevarnosti in pomanjkanje regulacije

Generativni umetno inteligentni sistemi, vključno z velikimi jezikovnimi modeli (LLM – Large Language Models), kot so GPT-4 in podobni sistemi, dodajajo splošni debati o umetni inteligenci nov nivo kompleksnosti. Ti sistemi proizvajajo človeškemu delu podobne rezultate brez pravega razumevanja njihovega pomena. Njihovo sklepanje in delovanje temelji na statističnih utemeljitvah na podlagi ogromnih podatkovnih nizov in ne na razumevanju v klasičnem pomenu besede. Povsem upravičeni so tudi pomisleki glede možnosti nastanka različnih vrst pristranskosti glede na vrsto (bazo) podatkov, ki je bila uporabljena za učenje izbrane UI. Še večji problem pa ostaja problem t.i. „črne skrinjice“ – tudi z vpogledom v kodo, ki je pogosto omogočen le razvijalcem, je praktično nemogoče interpretirati tisoče števil, uporabljenih v procesu odločanja, kaj šele milijone in milijarde računskih operacij, izvedenih nad temi istimi števili.

Našteto vodi do dveh glavnih tipov tveganj. Prvič, modeli pogosto proizvedejo „halucinacije“ – izhodne podatke, ki so sintaktično smiselni in morda celo verjetni, a dejansko napačni. Drugič in še pomembneje pa uporabniki pogosto nimajo možnosti preveriti, ali so izhodni podatki zanesljivi, medtem ko sistem sam ne more zagotoviti preverljive utemeljitve. To nekoliko spominja na Gödlova izreka o nepopolnosti, a je težava tu še veliko bolj zapletena. Po Gödlovem izreku noben sistem že kot tak ni popoln ali preverljiv, dodatno pa se situacija zaplete tudi zaradi težav, povezanih s problemom „črne skrinjice“, kar jasno izpostavlja tako kompleksnost kot omejitve tovrstnih sistemov.

> **Ne glede na to bi morali tisti, katerih odločitve lahko pomembno vplivajo na družbo – od znanstvenikov do posameznikov na odgovornih ali vodilnih položajih – prevzeti odgovornost za svoje zaključke in odločitve ne glede na to, ali so si pri svojem delu pomagali z umetno inteligenco ali ne.**

Nekaterim naključnim napakam se resda nikoli ne moremo povsem izogniti in jih lahko zgolj poskusimo omejiti. A v primeru uporabe UI, ki pogosto vrne tudi neponovljive odgovore, to ni najbolj preprosto. Ne glede na to bi morali tisti, katerih odločitve lahko pomembno vplivajo na družbo – od znanstvenikov do posameznikov na odgovornih ali vodilnih položajih – prevzeti odgovornost za svoje zaključke in odločitve ne glede na to, ali so si pri svojem delu pomagali z umetno inteligenco ali ne.

Z vse širšo integracijo UI sistemov v vsakdanje življenje, infrastrukturo in celo upravo postaja vse bolj očitno, da nam ti s svojo vseprisotnostjo prinašajo tudi nove oblike sistemskih tveganj, ki jih obstoječa zakonodaja slabo ali sploh ne regulira. Trenutni pravni okvir – na primer zakon Evropske unije o umetni inteligenci (EU AI Act) – premalo naslovi področja, kot so diskriminacija, načrtno širjenje dezinformacij in verižne napake. Zadnje kaj kmalu verjetno ne bodo več omejene zgolj na finančni sektor, saj se sistemska tveganja že danes neredko pojavljajo na področjih podnebnih sprememb in kibernetske varnosti.

Že prej omenjeni Geoffrey Hinton, ki si je zaradi svojega pionirskega prispevka na področju UI prislužil vzdevek: „boter umetne inteligence“, je leta 2023 zapustil Google, da bi lahko svobodno in neobremenjeno spregovoril o tem, kar sam imenuje eksistencialne in neposredne grožnje. Brez večjih težav lahko verjetno predpostavimo obratno sorazmerje med obsežnostjo sistema umetne inteligence in njegovo zanesljivostjo. S povečevanjem sistemov UI se hkrati zmanjšuje njihova sposobnost ustvarjanja natančnih in predvsem preverljivih rezultatov. V praksi to pomeni, da uporabljamo močne modele na področjih z visokim tveganjem in brez jamstva za njihovo natančnost ali pravilnost. Zato ponavljamo pretekli argument: uporaba UI brez pomislekov o njeni zanesljivosti na področjih, kot so medicina, inženirstvo... je nepredstavljiva in zahteva bistveno strožjo regulacijo odgovornosti.

V praksi to pomeni, da uporabljamo močne modele na področjih z visokim tveganjem in brez jamstva za njihovo natančnost ali pravilnost.

Od občudovanja do odgovornosti

Hitro – morda celo prehitro – se približujemo času, ko bodo sistemi umetne inteligence vplivali na odločitve, ki bodo presegale nadzor in razumevanje celo lastnih snovalcev. Omejitve, ki jih razkriva Gödlova logika, nepreglednost procesa odločanja strojev in vse bolj očitna sistemska tveganja niso več zgolj hipotetična vprašanja. Že danes vplivajo na javni diskurz, spodletele politike nadzora umetne inteligence in vse večje nezaupanje v institucije.

Danes ne potrebujemo občudovanja umetne inteligence, temveč stroge in izvršljive omejitve njenega razvoja in uporabe. Sistemom, ki niso sposobni pojasniti svojih odločitev, ki niso sposobni preveriti podanih trditev, ki ne morejo prevzeti odgovornosti za svoje napake, ne smemo podeliti odločevalske moči – ne v lastnih življenjih, še toliko manj pa na ravni skupnosti, držav ali civilizacij. Prihodnost umetne inteligence ni odvisna le od tega, česa je sposobna, temveč predvsem od našega razumevanja in nadzora njenih omejitev.