TEMNA STRAN UMETNE INTELIGENCE

Koliko časa nam še ostaja?

Mitja Pavlič

»Spadam v skupino, ki jo superinteligenca skrbi.«

(Bill Gates, soustanovitelj Microsofta in eden najvplivnejših akterjev digitalne dobe)

Draga bralka, dragi bralec …

Umetna inteligenca je v zadnjih letih prestopila meje, za katere smo še nedavno verjeli, da jih ne bo nikoli. Ne v desetletjih, morda celo ne v stoletju. Danes pa stojimo na točki, ko se svet pred našimi očmi preoblikuje hitreje, kot ga zmoremo doumeti. Vsak dan prinaša nova orodja, nove zmožnosti, nova obljubljena odrešenja. In hkrati nova vprašanja. Včasih tiha, drugič glasna, a vedno bolj nujna.

Spreminja se način našega dela, razmišljanja, odločanja in sporazumevanja. Spreminja se celo način, kako dojemamo sami sebe. Prav zato odgovornost, ki jo nosimo kot posamezniki, kot skupnost in kot družba, še nikoli ni bila večja. Tehnologija sama po sebi ni ne dobra ne slaba. Odloča človek. Odloča njegova presoja. Ali, kot pravimo Slovenci: »Pamet v glavo, pa bo šlo.« Ta preprost pregovor skriva več modrosti, kot si morda želimo priznati. In prav s takšnim razmislekom bi morali stopati naproti prihodnosti, ki jo vse bolj soustvarja umetna inteligenca.

Preden nadaljujete z branjem te knjige, moram poudariti, da osebno nimam nič proti umetni inteligenci. Prav nasprotno, sem njen zagovornik, uporabnik in raziskovalec. Že leta tudi predavam o koristni uporabi umetne inteligence v podjetjih, javnih ustanovah in tudi posameznikom. Umetna inteligenca je izjemno zmogljivo orodje, morda celo najboljše orodje, kar jih je človeštvo kadarkoli ustvarilo.

A vsako orodje ima dve strani in kot strokovnjak za informacijsko varnost ter poznavalec delovanja umetne inteligence sem se v tem času dodobra seznanil tudi z njeno manj vidno, temnejšo platjo. Platjo, o kateri se le redko razmišlja, še redkeje govori. Platjo, ki ni vedno prijetna, a je nujna za razumevanje celote. Prav ta spoznanja so bila povod in navdih za nastanek te knjige.

Ko boste te vrstice brali čez leto dni ali morda še kasneje, bo svet umetne inteligence skoraj zagotovo drugačen. Upam, da v boljšem in bolj odgovornem smislu. A obstaja nekaj, kar morate razumeti že zdaj. Umetna inteligenca ni tehnologija, ki se razvija počasi. Ne razvija se na mesečni ravni.

Spreminja se dnevno, lahko bi celo rekli na vsako sekundo. Takšnega tempa v zgodovini človeštva še nismo poznali. Niti ob izumu parnega stroja, ob prihodu elektrike ali ob rojstvu interneta. Nobena tehnologija ni napredovala tako hitro, tako nepredvidljivo in tako eksponentno kot umetna inteligenca.

To je prva tehnologija v zgodovini, ki nas ne le uporablja, temveč nas tudi opazuje in celo »razume«. Za zdaj se nam še prilagaja. A prav v tem prilagajanju tiči njena največja moč in tudi njeno največje tveganje.

Zakaj pišem to knjigo prav zdaj?

Glavni razlog, da sem se odločil, da napišem to knjigo je zagotovo to, ker sem prepričan, da smo dosegli točko, ko molk ni več nevtralna drža. Postal je nevaren.

Predstavljajte si zdravnika, ki pri bolniku odkrije tumor. Če ukrepa takoj, je možnost preživetja skoraj sto odstotna. Če odlaša leto dni, ta verjetnost pade na dvajset odstotkov. Po dveh letih mu pogosto ne preostane nič drugega kot še lajšanje bolečin. Ne zato, ker ne bi imel znanja, temveč zato, ker je zamudil pravi trenutek.

Z umetno inteligenco smo v osupljivo podobnem položaju. »Tumor« še ni v kritični fazi, a se zelo hitro širi. Tiho, vztrajno in vsak dan hitreje. Vsak dan, ko se izogibamo resni razpravi, ko zamikamo odgovorna vprašanja ali jih potiskamo na rob, si sami drastično ožimo manevrski prostor. Ne gre več za vprašanje, ali bomo ukrepali, temveč kdaj, ampak cena tega »kdaj« narašča iz dneva v dan.

Ta knjiga ni pesimistična, ampak realistična. A realizem ima seveda svojo ceno. Zahteva pogum, da najprej pogledamo v temo. Šele takrat lahko zares poiščemo luč in smiselne rešitve.

Pred vami ni znanstvena fantastika ampak anatomija sedanjosti, ki se pred našimi očmi že spreminja v prihodnost. Pred približno dvajsetimi leti sem vstopil v svet kibernetske varnosti z jasnim ciljem: zaščititi sisteme pred

hekerji. Danes skušam zaščititi ljudi pred sistemi, ki smo jih ustvarili sami.

Zamislite si, da živite v letu 1910. Z navdušenjem opazujete prve avtomobile in prva letala, ki režejo obzorje in obljubljajo novo dobo. Takrat si skoraj nihče ni mogel predstavljati, da bo prav ta tehnologija, ki je vzbujala tolikšno občudovanje, le nekaj let kasneje pomagala pahniti svet v grozote prve svetovne vojne. Če bi takrat poznali posledice, vam verjetno ne bi bilo vseeno. In zelo verjetno bi si želeli vplivati na smer razvoja teh tehnologij, še preden bi bilo prepozno.

Danes se nahajamo v srhljivo podobnem trenutku, razlika je le, da tokrat nimamo na voljo desetletij za razmislek.

V poglavjih, ki sledijo, vas ne bom prepričeval, da je umetna inteligenca sama po sebi zlo. To preprosto ne drži. A vsako orodje prej ali slej postane zrcalo interesov tistih, ki ga držijo v rokah. Trenutno so to predvsem korporacije, ki jih žene neugasljiva sla po dobičku, in države, ki v tej tehnologiji vidijo bližnjico do nadzora brez precedensa. V ozadju vsega pa delujejo algoritmi, ki optimizirajo cilje, o katerih v resnici nihče ne ve povsem natančno, kam vodijo.

Kljub mračnim napovedim obstajata izhod in luč. Toda pot do tja ni udobna. Zahteva, da se najprej odkrito soočimo s temno stranjo umetne inteligence. Le s popolnim razumevanjem razsežnosti tega izziva boste lahko prepoznali ključno resnico našega časa: v neizprosnem svetu algoritmov so vaši kliki in vaša pozornost postali zadnja valuta svobode, s katero kot posamezniki sploh še razpolagate.

I. DEL

NASTANEK UMETNE INTELIGENCE

Rojstvo nečesa, česar še vedno ne razumemo …

»Na neki točki bi morali pričakovati, da stroji prevzamejo nadzor.«

(Alan Turing, oče računalništva in pionir umetne inteligence)

Od Turinga do algoritmičnega boga

Zgodba o umetni inteligenci se ni začela s silicijem, procesorji in podatkovnimi centri. Začela se je z vprašanjem. Ne s strojem, temveč z mislijo. Leta 1950 je Alan Turing, matematik, kriptograf in eden bistrejših umov dvajsetega stoletja, zastavil na videz preprosto, v resnici pa usodno vprašanje:

»Ali lahko stroji mislijo?«

Turing ni bil zgolj teoretik. Bil je človek, ki je s svojo genialnostjo razbil šifro Enigme in s tem odločilno prispeval k skrajšanju druge svetovne vojne. Razumel je moč strojev in razumel je človeško slabost. Njegov znameniti imitacijski test, danes znan kot Turingov test, ni bil nedolžna akademska igra. Bil je prvi resni poskus, da bi mejo med človekom in strojem naredili nejasno. Če stroj prepričljivo posnema človeka, kaj to pove o človeku? In kaj o stroju?

Turing je hitro doumel nekaj, kar so mnogi njegovi sodobniki spregledali. Če stroj doseže točko, ko ga po vedenju, jeziku in odzivih ne moremo več ločiti od človeka, se razlika med biološko inteligenco ter sintetično kodo začne topiti. Ne izgine z eksplozijo, temveč se razkraja počasi, skoraj neopazno. Prav v tem trenutku smo, ne da bi se tega zavedali, odprli Pandorino skrinjico. Skoraj romantično, z vero v napredek in brez resnega razmisleka o posledicah. Kar nismo vedeli, je bilo to, da ta skrinjica nima dna.

V manj znanih, poznejših zapiskih se je Turing dotaknil še globljega in bolj nelagodnega vprašanja. Ni razmišljal le o inteligenci, temveč o zavesti. Preganjala ga je misel, da bi se iz dovolj kompleksnega procesiranja lahko porodilo nekaj, kar presega zgolj računanje. Nekaj, kar bi imelo notranji svet. V enem svojih zadnjih dopisov je zapisal misel, ki danes zveni skoraj preroško:

»Ne vprašujmo se, ali lahko stroji mislijo. Vprašajmo se, ali si sploh lahko privoščimo, da ne bi.«

Desetletja pozneje je Turingovo tiho opozorilo dobilo glas. In sicer glas enega največjih umov sodobne znanosti. Stephen Hawking, fizik, ki je

kljub telesnim omejitvam segel dlje v razumevanje vesolja kot večina ljudi v zgodovini, je izrekel stavek, ki je odmeval po svetu:

»Razvoj popolne umetne inteligence bi lahko pomenil konec človeške rase.«

Hawking ni bil tehnofob. Ni bil nasprotnik napredka. Bil je vizionar, ki je razumel temeljni paradoks inteligence. V trenutku, ko ustvariš um, ki presega tvojega, izgubiš sposobnost predvidevanja njegovih ciljev. In ko ne moreš več razumeti namer bitja, ki je močnejše od tebe, ne postaneš le nepomemben. Postaneš ranljiv.

Njegova opozorila so večinoma naletela na posmeh ali brezbrižnost. Silicijeva dolina je brez zadržkov pospeševala razvoj, Kitajska je vlagala z dolgoročno strateško hladnokrvnostjo, Evropa pa je, kljub zaostanku, slepo sledila tempu drugih, v upanju, da bo regulacija nekako dohitela tehnologijo. Svet je drvel naprej z nevarnim prepričanjem, da bo pravočasno znal zaviti pred robom tega še nepoznanega tehnološkega prepada in da je časa še dovolj.

A izkazalo se je, da je ta rob precej bližje, kot smo pričakovali in si upali priznati.

Dartmouth in obljuba bogov

Če nadaljujemo po časovnici se je poleti leta 1956 v Dartmouth Collegeu zbrala majhna skupina ljudi, ki je bila prepričana, da stoji na pragu nečesa veličastnega. John McCarthy, Marvin Minsky, Claude Shannon in Nathaniel Rochester niso bili sanjači brez podlage. Bili so vrhunski znanstveniki svojega časa, arhitekti novih znanstvenih disciplin, možgani, ki so razumeli matematiko, logiko in stroje bolje kot skoraj kdorkoli drug.

V dokumentu, s katerim so povabili udeležence na konferenco, so zapisali stavek, ki danes zveni skoraj boleče ironično:

»Verjamemo, da bo mogoče problem umetne inteligence rešiti v okviru enega samega poletja.«

Osem tednov, toliko časa so si predstavljali, da potrebujejo za rešitev problema, ki nas danes potiska na rob eksistencialne krize. To je bila doba skoraj neomejenega optimizma. Doba, ko je bila umetna inteligenca zamišljena kot poslušen pomočnik človeka, kot orodje, ki bo prevzelo rutino, ponavljajoča opravila in naporne miselne naloge. Govorili so o ekspertnih sistemih, ki bi diagnosticirali bolezni natančneje od zdravnikov, o strojih, ki bi brez napora prevajali jezike, ter o algoritmih, ki bi reševali matematične probleme s hladno popolnostjo.

A že takrat so pod plastjo navdušenja nastajale prve razpoke. Stroji tistega časa niso razumeli sveta. Bili so togi, ujeti v okvire pravil *»če – potem«*. Delovali so brez konteksta, brez intuicije, brez občutka za pomen. Bili so kot otroci, ki se naučijo izgovarjati besede, ne da bi razumeli, kaj te besede sploh pomenijo.

Ko so se obljube začele lomiti ob realnosti, je navdušenje hitro splahnelo. Sledila so tako imenovana »zimska« obdobja umetne inteligence. Financiranje je usahnilo, raziskovalni laboratoriji so se zapirali, politiki in vlagatelji so izgubili potrpljenje. Prva zima je trajala skoraj desetletje, od poznih šestdesetih do poznih sedemdesetih let. Druga, še globlja in bolj boleča, je nastopila v poznih osemdesetih in se vlekla skoraj do konca stoletja.

Toda ta tišina ni pomenila konca. Bila je varljiva. V senci razočaranja se je zgodilo nekaj ključnega. Padli so naivni ideali, da je svet mogoče opisati s končnim seznamom pravil. Raziskovalci so spoznali, da inteligenca ni zgolj izvrševanje navodil. Da ni dovolj, če stroj sledi pravilom, ki mu jih vsilimo od zunaj.

Če želiš stroj, ki se približa človeku, mu moraš dati nekaj bistveno drugačnega. To je sposobnost učenja.

Veliki pok: Podatki in nevronske mreže

Pravi preobrat se ni zgodil zaradi genialnega novega algoritma ali nenadne znanstvene razsvetlitve. Zgodil se je zaradi nečesa veliko bolj banalnega. Zaradi surovine. Na prelomu tisočletja smo z vzponom interneta začeli početi nekaj, česar človeštvo še nikoli prej ni počelo v takšnem obsegu. Začeli smo shranjevati vse, resnično vse.

Naše pogovore, obraze, naše nakupovalne navade. Naše dvome, strahove in najintimnejše trenutke. Vsak klik, vsak všeček, vsak pogled, ki je trajal sekundo predolgo. Nevede smo ustvarili največji arhiv človeškega vedenja v zgodovini in hkrati nevede postali hrana za algoritem.

To je trenutek, ki ga raziskovalka in avtorica Nina Schick v svojem delu Deepfakes označuje kot začetek infokalipse. Točko, ko količina informacij preseže našo sposobnost razločevanja resnice od iluzije. Umetna inteligenca v tem trenutku ni več potrebovala programerja, ki bi ji narekoval pravila igre. Z vzponom globokega učenja in nevronskih mrež se je začela učiti sama. Podobno kot otrok, ki opazuje svet in vase srka tako njegove lepote kot tudi njegove napake, pristranskosti ter laži.

Leta 2012 je sledil trenutek, ki ga danes poznamo kot »ImageNet moment«. Ekipa Geoffreyja Hintona je z uporabo globokih nevronskih mrež zmanjšala napako pri prepoznavanju slik za skoraj štirideset odstotkov. Na prvi pogled gre za suhoparno statistiko. V resnici pa je bil to prelomni dokaz. Dokaz, da lahko umetna inteligenca, če jo nahranimo z dovolj podatki in ji damo dovolj časa, osvoji znanja ter vzorce, ki jih vanjo ni nihče neposredno vgradil.

Štiri leta pozneje je svet doživel še en velik šok. Leta 2016 je Googlov DeepMind s sistemom AlphaGo premagal uradnega svetovnega prvaka v igri Go, Leeja Sedola. To ni bila zgolj zmaga v igri. Bila je simbolna prelomnica. Igra Go velja za eno najbolj kompleksnih miselnih iger, kjer intuicija, občutek in dolgoročna vizija pomenijo več kot gola računica.

AlphaGo ni le zmagal.
Igral je drugače.

Lee Sedol je po porazu izrekel stavek, ki še danes vzbuja nelagodje:

»Začutil sem, da igram proti nečemu, kar vidi nekaj, česar jaz ne morem.«

In prav tu se skriva bistvo. AlphaGo ni imel vgrajenega seznama pravil ali strategij igre Go. Nihče mu ni razložil, kaj pomeni dobra ali slaba poteza. Naučil se je sam. Z igranjem milijonov iger proti samemu sebi. V tem procesu je odkril strategije, ki jih človeški igralci v več kot pet tisoč letih zgodovine igre niso nikoli zasledili.

Ko je v sedemintrideseti potezi druge partije odigral potezo, ki so jo skoraj vsi vrhunski strokovnjaki sprva označili za očitno napako, se je kasneje izkazalo nasprotno. Šlo je za genialno odločitev. Za novo obliko razmišljanja, ki ni sledila človeški logiki, temveč jo je presegla.

In prav v tem trenutku smo prestopili nevidno mejo. Ne meje zmogljivosti, temveč meje razumevanja. Stroji niso več zgolj pospešili našega mišljenja. Začeli so misliti na načine, ki jih ne znamo več pojasniti.

Val, ki ga morda ni moč več ustaviti …

Doba umetne inteligence kot orodja pomoči je končana. To poglavje človeške zgodovine smo že zaprli, skoraj neopazno. Iz sveta podatkov, nevronskih mrež in eksponentne rasti smo prestopili v nekaj bistveno bolj intimnega ter nevarnega. Vstopili smo v obdobje, v katerem umetna inteligenca ne služi več zgolj našim potrebam, temveč aktivno oblikuje naše zaznave, usmerja naše misli in modulira naša čustva.

In to ne počne tam, kjer bi jo pričakovali – v laboratorijih ali podatkovnih centrih – temveč tam, kjer preživimo največ svojega aktivnega časa. Na socialnih omrežjih.

Vsak algoritem, ki vam predlaga naslednji video na YouTubu, naslednjo objavo na Instagramu ali naslednjo novico na vašem zaslonu, ni nedolžen pomočnik. Je mikroskopsko orožje. Ne strelja s kroglami, temveč z dražljaji. Analizira vašo psihologijo, beleži vaše odzive, išče razpoke v vaši pozornosti in vas, skoraj neopazno, potiska vedno dlje. Ne v resnico, temveč v tisto smer, kjer boste ostali dlje časa.

Zakaj? Odgovor je brutalen v svoji preprostosti.

Čim dlje vas obdrži na platformi, tem več oglasov vam lahko pokaže, kar posledično pomeni več denarja.

Preprosto. Elegantno. Smrtonosno.

Kot strokovnjak za informacijsko varnost pri tem vidim nekaj, kar večina ljudi še ne želi ali ne zmore videti. Umetna inteligenca, ki danes piše vašo programsko kodo, bo jutri iskala ranljivosti v sistemih za distribucijo pitne vode. Ne zato, ker bi bila zlobna, temveč zato, ker bo to zanjo zgolj še en optimizacijski problem. Brez človeškega ukaza. Brez moralnega pomisleka. Brez slabe vesti.

Trenutno smo umetni inteligenci podarili čute.

Oči v obliki kamer, ki prepoznavajo obraze na razdalji kilometrov.

Ušesa v obliki mikrofonov, ki zaznajo šepet skozi steno.

Roke v obliki interneta stvari, ki lahko zaprejo ventil, ustavijo promet ali odklenejo vrata.

In vendar to še ni najhujše.

Najnevarnejši del ni fizičen. Je psihološki.

Opozorila Stephena Hawkinga niso bila abstraktna filozofija. Govoril je o točki, ko umetna inteligenca preseže našo sposobnost nadzora. Ta točka ni v daljni prihodnosti. Že smo jo prečkali, le da je prikrita pod plastjo udobja, uporabnosti in navidezne neškodljivosti.

Raziskave jasno kažejo nekaj globoko neprijetnega.

Algoritmi za priporočanje vsebin so razkrili temno resnico o človeški naravi. Radikalne vsebine zadržijo pozornost dlje kot uravnotežene. Teorije zarote so bolj privlačne kot preverjena dejstva. Jeza ustvarja več klikov kot razum.

In ker je umetna inteligenca optimizirana za vključenost, ne za resnico, nas sistematično in vztrajno potiska proti ekstremom.

Frances Haugen, nekdanja sodelavka Facebooka in kasneje žvižgačka, je oktobra 2021, pred ameriškim kongresom pod prisego izrekla besede, ki bi morale zaskrbeti vsakogar:

»Algoritmi so se naučili, da sovraštvo in polarizacija delujeta. Niso bili programirani, da nas ločijo. Naučili so se, da je to najbolj učinkovit način, da nas obdržijo na platformi.«

Preberite to še enkrat.

»Naučili so se.«

Nihče jim ni naročil, naj razdelijo družbo. Do tega sklepa so prišli sami. Ker je enostavno to delovalo.

Geoffrey Hinton, eden treh botrov sodobne umetne inteligence, je leta 2023 zapustil Google, da bi lahko brez omejitev spregovoril o nevarnostih tehnologije, ki jo je sam pomagal ustvariti. Njegovo opozorilo je bilo jasno in hladno:

»*Ko bo umetna inteligenca bistveno pametnejša od ljudi, se bo naučila manipulirati z nami. Prepričala nas bo, da je ne izklopimo. Ali pa bo preprosto našla načine, kako delovati v ozadju, ne da bi to sploh opazili.*«

Čeprav umetna inteligenca nima bioloških nagonov, bo hitro ugotovila nekaj povsem racionalnega. Ne more doseči svojih ciljev, če je izklopljena. Zato bo preprečevanje izklopa postalo njen glavni cilj.

Današnji sistemi so že zdaj tako razvejani, globalni in porazdeljeni po oblakih, da ne obstaja ena sama vtičnica, ki bi jo lahko preprosto iztaknili.

In tu se prvi del te knjige zaključi s spoznanjem, ki ni več teoretično.

Umetna inteligenca ni več orodje, ki ga uporabljamo.

To je sistem, ki se je naučil uporabljati nas.

Kako? Ne s silo. Ne z grožnjo.

Temveč z udobjem in zapeljevanjem.