Prvi in drugi val umetne inteligence

Olga Markič
Filozofska fakulteta Univerze v Ljubljani

Povzetek

V prispevku prikažem dva vala umetne inteligence. Prvi temelji na deduktivni logiki in hipotezi o mišljenju kot računanju. Močna umetna inteligenca je z idejo, da bi lahko naredili računalniški model misli, spodbudila filozofske kritike, med katere uvrščamo tudi Marka Uršiča, ki je razmišljanjem o umetni inteligenci posvetil esej »Meje izomorfizma« v zbirki *Matrice logosa* (1987). V drugem valu se osredotočajo bolj na izdelovanje pametnih orodij, ki temeljijo na strojnem učenju. Taki sistemi so boljši pri modeliranju odločitev v negotovem okolju in pri modeliranju *znanja kako*. Uršič se k vprašanjem umetne inteligence vrne v razmišljanjih v *Zimi* (2015) in v *Shadows of Being* (2018), kjer izpostavi možne nevarnosti potencialne superinteligence.

Ključne besede: umetna inteligenca, simbolno manipuliranje, strojno učenje, zavest, Marko Uršič

First- and Second-Wave Artificial Intelligence – Abstract

This paper discusses two waves of artificial intelligence (AI). The first-wave AI is based on deductive logic and on the hypothesis of thinking as computing. Strong AI and the idea that it would be possible to construct a computational model of thought gave rise to philosophical criticism. Among the critics is Marko Uršič, who investigates artificial intelligence in his essay 'Meje izomorfizma', published in his book *Matrice logosa* (1987). The second-wave AI focuses more on developing smart tools based on machine learning. These systems are better in modelling decision-making within an uncertain environment and in modelling *knowledge--how*. Uršič revisits the questions related to AI in his books *Zima* (2015) and *Shadows of Being* (2018), where he points to the possible dangers of potential superintelligence.

Key words: artificial intelligence, symbolic manipulation, machine learning, consciousness, Marko Uršič

Uvod

Ko sem kot mlada raziskovalka prišla na fakulteto, sem kot asistentka vodila vaje pri predmetih *Logika* in *Metodologija*, ki ju je predaval Marko Uršič. Marka sem poznala že kot profesorja z gimnazije in vesela sem bila, da mi je prav on pomagal pri vstopanju v visokošolsko pedagoško delo. Bil je zelo natančen, a hkrati potrpežljiv in vedno pripravljen na pogovor. Pogovarjala sva se o različnih filozofskih vprašanjih, a ker sem na študij

filozofije prišla po nekajletnem ukvarjanju z računalništvom, so bile še posebej razgrete diskusije o umetni inteligenci (UI) in o možnostih modeliranja kognitivnih procesov. Čeprav so se pozneje najine filozofske diskusije dotikale različnih tem, pa se v svojem prispevku vračam k najinim prvim filozofskim pogovorom, ki sta jih spodbudili dve Hofstadterjevi knjigi. Prva je kultna *Gödel, Escher, Bach: an Eternal Golden Braid* (1979), o kateri je Uršič pisal v eseju »Meje izomorfizma« (v *Matrice logosa*, 1987) in tudi v *Jeseni* (2010), druga pa zbornik *Oko duha*, ki sta ga uredila Hofstadter in Dennett (1990), Uršič pa je bil urednik slovenske izdaje. V zborniku so številni zanimivi in provokativni filozofski in literarni prispevki, med njimi tudi dva temeljna filozofska teksta prvega vala umetne inteligence, Turingov *Stroji, ki računajo, in inteligenca* (1990) in Searlov *Duhovi, možgani in programi* (1990). V zadnjem času, ko smo priča drugemu valu umetne inteligence, se je Uršič spet vrnil k tej tematiki v *Zimi* (2015) in v *Shadows of Being* (2018) ter v svojih številnih predavanjih, ki so dostopna tudi na njegovi spletni strani. V nadaljevanju bom predstavila temeljne značilnosti teh dveh valov raziskovanja v umetni inteligenci in Uršičev kritični pogled na možnost mislečih/zavestnih strojev.

Umetna inteligenca

V splošnem bi umetno inteligenco lahko opredelili kot iskanje, kako narediti računalnike zmožne početi tisto, kar je zmožen narediti človeški um. Pri tem gre lahko za mišljenje, ki ga običajno povezujemo z inteligentnim vedenjem, ali pa za druge psihološke veščine, kot na primer zaznavanje, planiranje, napovedovanje, nadzorovanje gibanja, ki ljudem in živalim omogočajo doseganje ciljev (Boden, 2016: 1). Raziskovalce v UI lahko v grobem razdelimo v dve skupini, ki se običajno ne prekrivata. V prvi so tisti, ki jim je osrednji cilj raziskovanje kognitivnih procesov in modele umetne inteligence uporabljajo predvsem za oblikovanje in preverjanje znanstvenih teorij, s katerimi odgovarjajo na različna vprašanja, povezana z delovanjem ljudi in živali. V drugi pa raziskovalci zasledujejo bolj praktične cilje in razvijajo pametna orodja, pri čemer se ne omejujejo na podobnostmi s človekom, temveč iščejo najbolj učinkovite metode za dosego zadanega cilja. Moj namen v tem članku ni obravnava teorij duševnosti, ki so povezane z UI, čeprav se temu ne da povsem izogniti, ampak poskus pogledati na UI skozi razlike med pristopom prvega vala, ki temelji na deduktivni formalni logiki, in pristopom drugega vala, ki temelji na strojnem učenju in indukciji (razumljeni kot različne vrste sklepanj, npr. posplošitev, abdukcija, analogija, ki napovedujejo sklep z večjo ali manjšo verjetnostjo). Tak pristop je navdahnjen z razdelitvijo, ki jo je v knjigi *The Promise of Artificial Intelligence: Reckoning and Judgment* uporabil Brian Cantwell Smith (2019). Za uradno rojstvo raziskovalnega področja umetne inteligence šteje leto 1956, ko je McCarthy v vabilu na poletno raziskovanje v Dartmouth Collegeu prvič uporabil ime umetna inteligenca (Russell in Norvig, 2010: 17). Kot ugotavljata Russell in Norvig, je izbira imena

nakazovala, da se raziskovalci ukvarjajo z idejo dupliciranja človeških zmožnosti, kot so uporaba jezika, kreativnost in samoizboljševanje, ter poskušajo zgraditi stroje, ki bodo delovali avtonomno v kompleksnem in spreminjajočem se okolju (*Ibid.*: 18). Poleg te skupine, ki jo je zanimala psihološka raven in se je osredotočala predvsem na modeliranje jezika in logičnega sklepanja, je bila dejavna tudi skupina, ki je modelirala fiziološke samoregulirajoče se procese in je uporabljala nevronske mreže. Čeprav sta obe skupini nekaj časa delovali sodelovalno, je po letu 1960 prišlo do intelektualnega razcepa. Margaret Boden ga označi takole: »[...] tisti, ki jih je zanimalo *življenje*, so ostali v kibernetiki, tisti, ki so se zanimali za *um* (*mind*), pa so se obrnili k simbolnemu računanju« (2016: 17). Simbolna umetna inteligenca je na začetku dosegala večje uspehe, pomembno je vplivala tudi na začetek kognitivne znanosti (klasična ali simbolna kognitivna znanost). Čeprav so bila pričakovanja, pa tudi napovedi, zelo optimistična, je prvi val zašel v slepo ulico, ki jo označujejo tudi kot »UI zimo« (Russell in Norvig, 2010: 24). Drugi val, ki smo mu priča zdaj, je črpal iz kibernetike in modeliranja z nevronskimi mrežami, poudarek pa je na strojnem učenju.

Prvi val: simbolno manipuliranje

Zamisel o miselnih procesih kot neke vrste računskih procesih je sicer nastala že mnogo pred iznajdbo elektronskih računalnikov. A šele z iznajdbo računalnika se je odprla možnost, da se s pomočjo teorije, ki na mišljenje gleda kot na računanje (računska reprezentacijska teorija; Fodor, 1987), vsaj v principu pokaže, kako je mogoča fizična realizacija mišljenja. Tako kot lahko računalnik, ki je zgolj fizični sistem, s pomočjo programa, ki je implementiran v strojnem jeziku, realizira operacije s simboli, imajo tudi možgani svojo nevralno kodo, v kateri je realizirano mišljenje. Če bi uspeli dejansko narediti tak model uma, bi imeli močno UI.

John Haugeland v svoji knjigi *Artificial Intelligence: The Very Idea* (1985) zelo dobro predstavi glavne značilnosti prvega vala umetne inteligence oziroma GOFAI (*Good Old-Fashioned Artificial Intelligence*), kot jo je poimenoval. Naj povzamem nekaj pomembnih zgodovinskih mejnikov pri oblikovanju te ideje. Pomembno mesto v 'predzgodovini' UI ima filozof Thomas Hobbes, ki je zagovarjal tezo, da je mišljenje računanje. Po njegovem mnenju naj bi bilo mišljenje notranji mentalni pogovor. Pri mišljenju naj bi, podobno kot pri glasnem pogovoru ali računanju s pomočjo svinčnika in papirja, uporabljali simbolne operacije, le da te niso izražene z govorjenimi ali pisanimi simboli, temveč v posebni nevralni kodi. Kadar razmišljamo, naši možgani računajo. V času, ko je ustvarjal Hobbes, njegove ideje niso mogle vzpodbuditi dejanskega raziskovanja, povezanega z modeliranjem, se je pa zamisel o mišljenju kot računanju pozneje še večkrat pojavila. Gottfried Wilhelm Leibniz je predlagal izoblikovanje natančnega in nedvoumnega univerzalnega jezika (*characteristica universalis*), v katerega bi bilo mogoče prevesti vse ideje

in v katerem bi mišljenje potekalo kot računanje. Čeprav si je že Leibniz izmislil dvojiški zapis, pa je bil George Boole tisti, ki ga je prvi uporabil za opis mišljenja. Logične odnose med stavki (propozicijami) je izrazil s pomočjo matematične strukture, ki je pozneje dobila ime Boolova algebra. Boole je menil, da lahko iz teh preprostih algebraičnih form gradimo vzorce mišljenja in tako odkrijemo »zakone mišljenja«.

Konec 19. in začetek 20. stoletja je pomenil velik razmah formalne logike in raziskovanja narave formalnih sistemov. Posebno mesto gre Alanu Turingu, ki je v članku o izračunljivih številih iz leta 1936 definiral računanje kot formalno manipulacijo z neinterpretiranimi simboli, ki se izvaja z uporabo formalnih pravil. Opisal je preprosto imaginarno napravo, ki jo danes imenujemo Turingov stroj, in pokazal, da lahko s takim preprostim strojem izvedemo vsako nalogo, za katero lahko jasno navedemo korake, ki so potrebni za izpolnitev naloge.

Turingove ideje so navdihnile misel o stroju računalniku, ki bi s svojo formalno strukturo posnemal delovanje človekovega mišljenja, raziskovalci pa so se ukvarjali z vprašanjem, kako napisati računalniške programe, da se bodo stroji vedli »inteligentno«. Ali je za to, da nekaj misli, nujno, da misli tako kot človek, ali lahko to počne tudi na kakšen drugačen način? V iskanju odgovora na vprašanje, ali in kdaj stroj misli, je Turing (1990) predlagal operativni test, preizkus, ki temelji na igri oponašanja. V njem se spraševalec lahko v 30 % prepoznal, da v pogovoru sodeluje računalnik, ki želi spraševalca preslepiti, da je človek. Če bi računalniku uspelo, potem po Turingu ne bi imeli razlogov, da bi zanikali, da stroj res lahko misli. Turingov test kot kriterij za mišljenje je sprožil burne razprave in niz argumentov za in proti ter seveda preizkušanje novih in zmogljivejših programov, s katerimi bi bilo morda mogoče uspešno opravili test (npr. Copeland, 1993).

Eden od najbolj odmevnih kritikov Turingovega testa je John Searle (1990), ki poudarja, da računalniku manjka človeško razumevanje. Za ponazoritev tega ugovora si Searle zamisli miselni eksperiment, imenovan »Kitajska soba«. Meni, da bi program, ki bi imel ustrezno napisana navodila, lahko opravil Turingov test. Človek zunaj sobe, ki govori kitajsko, bi lahko mislil, da se pogovarja z osebo v sobi. Toda dejansko Searle (oseba v sobi) ne bi razumel kitajsko. Vse, kar bi počel, bi bila zgolj manipulacija s simboli v skladu z njihovo obliko, ne bi pa imel pojma, kaj ti simboli pomenijo. Searle je na ta način opozoril na težavo, da izvorna intencionalnost in razumevanje simbolov ne pritičeta samemu programu, ampak človeku, ki je stroj programiral. Program oziroma stroj torej ni avtonomen in samo preklada simbole, nima pa pravega razumevanja (Miščević in Markič, 1998).

Tudi Uršič meni, da je poskus močne UI, da bi bilo z ustreznim programom mogoče formalno opisati človekove kognitivne procese in jih implementirati v fizičnem svetu, obsojen na neuspeh. V svoji kritiki se osredotoči na »matematični ugovor«, ki se naslanja na Gödlov teorem o nepopolnosti in na konstrukcijo paradoksalnega stavka

G, ki ga lahko parafriziramo kot: G: »Stavek 'G' je nedokazljiv« (Uršič, 1987: 187). Kot zapiše Uršič: »Presenetljivo je: stavek 'G' je *resničen*, čeprav eksplicitno zanika vsakršen dokaz za svojo lastno resničnost« (*Ibid.*). Iz tega izhaja nepopolnost formalnih sistemov, tj. množica dokazanih teoremov ne more nikoli zaobseči množice resničnih teoremov. Ali kot pravi Uršič: »[...] algoritem (program) nikoli ne more biti popoln, vedno se mu nekaj 'izmakne'« (*Ibid.*). V kritični obravnavi možnosti močne umetne inteligence se naslanja na filozofa J. R. Lucasa (1961) in zaključuje: »[...] dokler programiranje temelji na izomorfizmu s formalnimi sistemi (kar *in ultima analysi* velja za še tako zapletene računalniške jezike), je v primerjavi s človekovo zavestjo *a priori* nepopolno« (*Ibid.*: 200). Četudi bi se samonanašanju v formalnih sistemih lahko izognili (npr. rešitev Russllovega paradoksa), pa je po drugi strani samonanašanje »intencionalni akt, ki je za zavest bistven: zavest je vselej zavedanje svojega lastnega zavestnega akta, miselne aktivnosti, časenja. Zavest je zavest-o-sebi, 'presenetljiva zanka' na najvišji ravni« (*Ibid.*: 187).

Oba ugovora izpostavljata človekovo zavest kot tisto, ki kaže na nepremostljivo oviro za vsak poskus močne UI. Če se naslonim na Aaronsonovo razpravo o računski kompleksnosti (2013) in njegovo analizo dveh branj Turingovega testa, bi rekla, da oba zanima metafizično vprašanje. Aaronson ju je predstavil sledeče:

> *Metafizično*: Recimo, da bi računalniški program prestal Turingov test (v tako močni različici, kot si kdo želi). Ali bi bilo prav, da bi mu potem pripisali »zavest«, »qualia«, »biti o nečem«, »intencionalnost«, »subjektivnost«, »biti oseba« in drugo, kar želimo pripisati drugim ljudem in nam samim?
>
> *Praktično*: Ali je mogoče, da bi bil program, ki bi prestal Turingov test (močno verzijo), dejansko napisan? (Aaronson, 2013: 270)

Aaronson ugotavlja, da je Turing prav zato, ker je poskušal razlikovati to dvoje, predlagal svoj test. V odgovoru na ugovora iz zavesti pravi: »Ne želim dajati vtisa, da mislim, da ni nobene skrivnosti glede zavesti. Nekaj paradoksalnega je na primer v vsakem poskusu, da bi jo lokalizirali. Toda ne mislim, da je takšne skrivnosti treba nujno rešiti, še preden lahko odgovorimo na vprašanje, s katerim se ukvarjamo v tem članku.« (Turing, 1990: 68) Turing se je zavedal, da je problem zavesti prevelik, da bi ga lahko rešil, a je menil, da lahko znanost napreduje po manjših korakih, ne da bi najprej rešila veliko skrivnost.

Uršiča Turingovo stališče ne prepriča, saj je po njegovem ravno zavest najbolj neposredno dana in je zato ne moremo »odpisati«, pa četudi zgolj za raziskovanje na trenutni stopnji znanosti.

Na iztek prvega vala pa so verjetno najbolj vplivale težave v praktični izvedbi, saj orodja, ki so temeljila na ideji, da je vse znanje mogoče predstaviti v obliki stavkom podobnim strukturam, osnovne operacije pa so simbolno računanju v skladu s pravili, niso izpolnila pričakovanj (Russell in Norvig, 2010: 24). Na neustreznost takega pristopa je s filozofskega vidika opozarjal Hubert Dreyfus (Dreyfus, 1972; Dreyfus H. L. in Dreyfus

S. E.,1991) in izpostavil, da za vsakdanje znanje potrebujemo drugačno obliko predstavitve informacij. Njegova analiza je temeljila na razlikovanju dveh vrst znanja: *znanje da*, ki je propozicijsko in ga je mogoče predstaviti s simbolnimi strukturami, in *znanje kako*, ki se taki predstavitvi izmika. Tako Dreyfus kot Searle sta s svojo kritiko simbolne UI opozarjala, da bi se moral sistem, ki bi želel nekaj razumeti, vključiti v svet. Šele v interakciji s svetom bi lahko prišel do pomena posameznih simbolov. V filozofskih teorijah duševnosti se nov pristop kaže v odmiku od reprezentacijskega funkcionalizma (in s tem Descartesove dediščine) in raziskuje utelešenost, umestitev v okolje, enaktivizem in razširjeno kognicijo (Varela, Thompson, Rosch, 1991; Clark, 1997, 2001).

Drugi val: strojno učenje

Drugi val se v nasprotju s prvim, ki se je opiral na deduktivno logiko, obrača k induktivnemu sklepanju, kjer je poudarek na učenju na osnovi predhodnih izkušenj in interakciji z okoljem. Logika je sicer dobra za razmišljanja v dobro definiranem okolju, a v realnem svetu, kjer imamo opraviti z negotovostjo, se verjetnost izkazuje kot bolj primerna. Namesto programov – navodil za manipuliranje s simboli, računalničarji zdaj pišejo algoritme za strojno učenje na podlagi učnih primerkov. Prvi model nevronske mreže sta predstavila že McCulloch in Pitts (1943), ki sta upoštevala tako dosežke teorije računanja kot tudi znanje o osnovah fiziologije in delovanju nevronov. Zato njun članek smatrajo za začetek obeh valov UI, tako pristopa simbolnega manipuliranja kot tudi nevronskih mrež in strojnega učenja (Russell in Norvig, 2010: 16). Slednji se je najprej zgledoval po procesih učenja v možganih (dejanskih nevronskih mrežah), ki se po znanstveniku, ki je te procese preučeval, imenuje Hebbovo učno pravilo. Pravilo je enostavno in temelji na spoznanju, da se vez med nevronoma, ki sta hkrati vzburjena, ojači. To pravilo uvrščamo med nenadzorovano učenje, saj ne poznamo ciljne vrednosti, s katero bi primerjali dobljeni rezultat. Modeliranje nevronskih mrež in konekcionističnih modelov je dobilo nov zagon v sredini osemdesetih let preteklega stoletja z odkritjem učnih algoritmov nadzorovanega učenja, ki so glede na razlike med ciljnim in dobljenim rezultatom zmožni popravljanja uteži med enotami (umetnimi nevroni) tudi v večnivojskih mrežah (več o konekcionizmu v Markič, 2011).

Čeprav laiki pogosto uporabljamo ime nevronska mreža, imajo sodobni modeli v UI za cilj izdelovanje efektivnih modelov in algoritmov na različnih področjih vsakdanjega življenja, ki je zelo oddaljeno od dejanskega delovanja živčnega sistema, npr. varovanje pred nezaželeno e-pošto. Raziskave, katerih cilj je natančno modeliranje empiričnih lastnosti dejanskih nevronov in skupin nevronov, pa uvrščamo v računsko nevroznanost (*Ibid.*: 23–26).

Drugi val zaznamuje razvoj teorij strojnega učenja in znanstveni pristop pri oblikovanju modelov ter povezovanje z drugimi disciplinami, predvsem s teorijo verjetnosti in statistiko. Na primer, za modeliranje odločanja v negotovosti se razvijajo formalizmi

bayesijanskih mrež. Poleg tega se je zaradi velikih baz podatkov, ki jih omogoča internet, začel fokus premikati z algoritmov na same podatke. Porodile so se ideje o inteligentnih agentih (virtualnih ali robotih z UI), ki bi bili vsaj do določene mere avtonomni. In ne nazadnje, spet se poskuša razvijati umetno splošno inteligenco (*Artificial General Intelligence*) in umetno inteligenco človeške ravni kot njeno podskupino. Lahko bi rekli, da je sklenjen krog, saj se tudi drugi val vsaj v nekaterih segmentih vrača k ideji, ki jo je začetnik UI Herbert Simon izrazil kot »stroji, ki mislijo, se učijo in ustvarjajo« (*Ibid*.: 27).

Če bi danes vprašali naše študente o umetni inteligenci, bi se verjetno spomnili medijsko odmevnih dogodkov, kot sta bila zmaga programa Deep Blue leta 1996 nad takratnim svetovnim šahovskim prvakom Garryjem Kasparovim in zmaga programa Alpha Go leta 2017 nad takrat najboljšim igralcem igre go Kejem Jiejem. A če malo pomislimo, je UI prisotna v našem vsakdanjem življenju, ne da bi se tega zares zavedali. Prisotna je predvsem v obliki orodij, ki nam pomagajo, da določene naloge opravimo lažje in bolje, npr. odpremo telefon s pomočjo prepoznave prstnega odtisa, najdemo naslovnika v neznanem mestu ali prevedemo sporočilo iz tujega jezika, pa tudi kupimo izdelek v priljubljeni spletni trgovini, program sam pa nam ponudi še vrsto drugih izbir, ki bi me morda lahko zanimale. UI se uporablja kot pomoč pri odločanju v bančništvu, medicini in pravu pa tudi vojski.

Večina teh učinkovitih uporabnih orodij je specializirana za ozko področje in se uporablja za določeno nalogo, npr. orodje za prepoznavanje obrazov ali govora. V človeku, pa verjetno tudi pri živalih, integriranje različnih modalnosti pripisujemo zavesti. Če se naslonimo na Blocka (1995), ki razlikuje med fenomenalno (*phenomenal*) in dostopno (*access*) zavestjo, bi bila to dostopna zavest. Ob tem se postavlja vprašanje, ali lahko pričakujemo, da bodo take integrativne funkcije sposobni tudi umetni učeči se agenti. Kakšni bi morali biti agenti, ki bi imeli umetno splošno inteligenco?

Mindt in Montemayor (2020) sta predstavila razdelitev, s katero sta poskušala zajeti inteligentne sisteme, v katere bi lahko razporedili relativno preprosta orodja, živali, ljudi in tudi agente, ki bi morda ljudi lahko presegali, od orodij znanja do proizvajalcev znanja. Njun kriterij je bil način, kako sistem/agent rešuje probleme glede na »motivacije in potrebe, ki jih zadovoljujejo skozi različne vrste inteligentnih oblik izbora in občutljivosti na preference, ki so relevantne za reševanje problemov« (*Ibid*.: 14).

Menim, da nam njuna razvrstitev daje dobro analitično orodje za diskusijo o trenutnih in potencialnih sistemih UI, zato jo na kratko povzemam.

Orodje znanja 0: Reševalec posameznih problemov: sistemi z določeno stopnjo avtonomije in robustno integracijo informacij za reševanje enega problema, ki se lahko razdeli v več podproblemov, kot je igranje igre pod nadzorom človeka.

Orodje znanja 1: Reševalec več problemov: sistemi z določeno stopnjo avtonomije in večjo integracijo informacij za reševanje več problemov pod nadzorom človeka. Algoritmi strojnega učenja, ki so lahko tekmovalni igralci v mnogih igrah.

Proizvajalec znanja 1: Sistemi s precejšnjo stopnjo avtonomije za optimizacijo, nadzor, razširjanje in filtriranje informacij. Temeljijo na potrebah in ciljih za specifične naloge. Ta tip inteligence je prva raven spoznavne zmožnosti z intencionalnostjo. Mnoge živali imajo tako vrsto inteligence v modularni obliki. Morda AlphaGo Zero.

Proizvajalec znanja 2: Sistemi, ki rešujejo splošne probleme, tako da zadovoljijo potrebe in želje, imajo visoko stopnjo avtonomije, motivacijo, kognitivno integracijo in pozornost za relevantne informacije. To je prva raven pravih spoznavnih agentov. Ta višja raven integracije pomeni prehod k dostopni zavesti, dostopnosti izbranih informacij za misel, dejanje in vedenje.

Proizvajalec znanja 3: Sistemi z robustno ravnjo spoznavnega delovanja, visoko stopnjo avtonomije, kognitivne integracije in kompleksnih motivacij. Splošna inteligenca se eksplicitno pokaže skozi semantične kategorije in jezikovne veščine komuniciranja. To ustreza človeški inteligenci, na katero vpliva jezik. To je človeška raven dostopne zavesti in fenomenalni zavesti.

Proizvajalec znanja 4: Skupine, ki proizvajajo znanje na način, ki ga ni mogoče reducirati na seštevek mnenj posameznih članov skupine, in imajo svoje lastne motivacije, cilje in preference. To bi bili UI, ljudje in celo 'post'-izboljšani ljudje. Predstavljali bi nov tip kolektivnih agentov, ki je daleč nad današnjimi zmožnostmi, utemeljenimi na človeški ravni zavesti (*Ibid*.: 14).

Kot ugotavljata avtorja, se dosedanji sistemi UI uvrščajo med orodja in ne proizvajalce znanja. Orodja so reševalci problemov pod nadzorom človeka, saj nimajo notranjih potreb in ciljev, čeprav imajo lahko določeno stopnjo avtonomije, kot na primer, da igrajo igro. Človek pa je agent z avtonomijo in spoznavnimi motivacijami, njegova dejanja so določena z njegovimi intencami, da doseže cilj, ne pa z zgolj vzročnimi ali zunanjimi izvori. Ob tem smo ljudje mojstri v izdelovanju in uporabi orodij, kar nam omogoča, da uspešno rešujemo probleme. Jezik nam omogoča, da obdelujemo kompleksne informacije, od neposrednega nanašanja do abstraktnih matematičnih resnic (*Ibid*.: 19).

Kot smo zapisali v uvodu tega razdelka, pa so se z uspehi, ki so jih dosegli z UI drugega vala, večale tudi ambicije. Ali bi bilo mogoče narediti ne samo dobra orodja, ampak tudi proizvajalca znanja? Mindt in Montemayor kot možnega kandidata za uvrstitev v kategorijo *Proizvajalec znanja 1* izpostavita program za igranje goja AlphaGo Zero, naslednika že prej omenjenega programa AlphaGo. Naj na kratko predstavim razliko, ki je po mnenju avtorjev bistvena. AlphaGo je sestavljen iz dveh nevronskih mrež, pri čemer je prva izpostavljena nadzorovanemu učenju na podlagi primerov iger mojstrov goja, druga pa to kodirano znanje izboljša še s spodbujevalnim učenjem pri igranju igre s samo seboj. Nato vzame rezultate teh dveh mrež in uporabi metodo Monte Carlo (MCTS – Monte Carlo Tree Search), ki projicira možne poteze v prihodnost, tako da določi tisto potezo, ki ima največjo verjetnost za zmago sistema. AlphaGo Zero so naučili samo osnovnih pravil postavljanja belih in črnih kamnov. Nato je, namesto da bi

se naslanjal na človeško znanje in se učil na igrah mojstrov goja, sam odkrival in razvijal svoje 'znanje' o goju skozi igranje s samim sabo z metodo spodbujevalnega učenja. Sestavljala ga je ena sama nevronska mreža, ki vključuje tudi MCTS. Avtorja poudarjata, da se je AlphaGo Zero tako sam naučil igranja in ob tem ponovno odkril veliko znanja mojstrov goja. Razvil pa je tudi nove strategije igranja (*Ibid.*: 22–23).

Prav slednje je najbolj zanimivo za našo diskusijo. AlphaGo Zero je fascinanten sistem, ki je sicer daleč od umetne splošne inteligence, saj je specifičen za dano nalogo s ciljem zmagati v igri go. Vendar nam to, da se je igre naučil sam, ne da bi se pri tem oprl na znanje mojstrov igre, nakazuje, da bi ga morda lahko uvrstili med proizvajalce znanja. Po drugi strani pa prav ta odlika, da je sam prišel do novega znanja, poraja vprašanje, ali lahko temu znanju zaupamo. Sistem je za človeka kot nekakšna 'črna škatla'. Ne vemo, kako je prišel do znanja, za nas je netransparenten. V tem konkretnem primeru to morda niti ni tako pomembno (čeprav bi si igralci goja to verjetno želeli), a če se bodo podobni sistemi uporabljali na drugih področjih, na primer v znanosti, se zahtevi po razlagi ne bomo mogli kar tako odpovedati.

Vrzel med zmožnostjo predvideti izid na eni strani in razlago oziroma utemeljitvijo, zakaj sprejeti predlagano rešitev, na drugi, je pri sistemih drugega vala zelo globoka. Zdi se, da gre za nekakšno zrcalno sliko prej omenjene Dreyfusove kritike. V prvem valu, ko je UI temeljila na deduktivni logiki in *znanju da*, ni mogla uspešno modelirati veščin in odločanja v negotovih pogojih. V drugem valu, ob vseh praktičnih uspehih pri izdelavi »pametnih orodij«, sistemi, ki temeljijo na indukciji in učenju, razmeroma dobro pokrivajo naloge, za katere je potrebno *zanje kako*, ne premorejo pa *znanja da.*

Sklepne misli

Ob uspehih UI drugega vala so se ponovno obudile tudi napovedi o mislečih strojih. Ko Uršič v *Zimi* (2015) in *Shadows of Being* (2018) kritično obravnava futuristične ideje o singularnosti, ki ju zagovarjata Verner S. Vinge (2003) in Ray Kurzweil (2005), pravi, da ne verjame, »da bo prišlo do kake kibertehnične singularnosti v tem ali naslednjih stoletjih, kaj šele v času mojega/našega življenja« (Uršič, 2015: 453). Mislim, da je Uršičeva kritika dejansko globlja. Ko se sprašuje o zavesti in možnosti umetne inteligence, dilemo med redukcionizmom in holizmom opiše takole:

> Hofstadter namreč združuje epistemološki holizem in ontološki redukcionizem (zavest je spoznavno, fenomenološko celostna, čeprav je fizično sestavljena le iz »delcev«, recimo, bitov) – in kaj je bolj *realno*, deli (delci, biti) ali celota (zavest), je odvisno od tega, kaj smatramo za ontološko *primarno*: fizično naravo ali zavest, telo ali duha <mind>. [...] Zagovornikom umetne inteligence/zavesti gre predvsem za argumentacijo, da je inteligentni »softver« (misli, predstave, čustva itd.) lahko procesiran v fizično *različnih* »hardverih« (v bioloških ali računalniških ali kakih drugih možganih). (Uršič, 2010: 349)

V nadaljevanju Uršič zapiše, da platonik, »kakršen sem jaz, ki sem prepričan – saj to *vem* iz 'osebne izkušnje' (gotovo tudi ti to veš, tudi če tega ne priznaš) – [sprejemam] da je *duh realen* v polnem, neposrednem, nereduciranem pomenu« (*Ibid.:* 350).

Hkrati pa je Uršič mnenja, da je treba svarilo o možnih nevarnostih prihodnje superinteligence, kot ga je v knjigi *Superintelligence: Paths, Dangers, Strategies* predstavil filozof Nick Bostrom (2014), jemati resno. Ob tem izpostavi potrebo po večjem povezovanju informacijske in druge tehnološke znanosti z družboslovjem in filozofijo (Uršič, 2015: 481).

Razvoj pametnih orodij drugega vala vsekakor odpira mnoga spoznavna, družbena in etična vprašanja. Delovanje nevronskih mrež ne temelji na logičnem sklepanju, zato je način, kako pridejo do rezultata, za človeka netransparenten in težko razumljiv. Podatki – učni primeri, na katerih se mreža uči, kot tudi cilji, ki naj bi jih dosegla, odražajo stališča in prepričanja razvijalcev in naročnikov ter so vpeta v družbeni kontekst. Odmevno kritiko premalo reflektiranih predpostavk modeliranja, ki temeljijo na pristranostih in vodijo v etično sporne odločitve, je na podlagi analize orodij, ki jih uporabljajo v ZDA, prepričljivo podala Cathy O'Neil (2016). Kot družba danes stojimo pred izzivi, kako uporabljati pametna orodja, da ne bomo ogrozili temeljnih vrednot, kot so pravičnost, spoštovanje zasebnost, transparentnost odločanja, če naštejemo le nekatere najbolj izpostavljene.

Prispevek zaključujem z Uršičevo mislijo iz zadnjega dela eseja »Virtualni svetovi v realnem času« (2015), ko piše o Penrosovi (1995) parafrazi Platonove prispodobe o votlini: »Moderna znanost je dosegla izjemne uspehe pri oblikovanju in preverjanju teorij o gibanju in spreminjanju 'senc', tj. o strukturi in dinamiki fenomenov, algoritmov, sistemov itd., pri tem pa pozablja na njihovo 'globino', na duha« (2015: 482). Podobno bi lahko razumeli tudi Uršičevo stališče. Pristopi umetne inteligence so le modeli na ravni senc in ne morejo podajati prave »teorija duha«, ki sodi v svet zunaj votline.

Literatura

Aaronson, S. (2013). »Why Philosophers Should Care about Computational Complexity«. V Copeland, J., Posy, C. in Shagrir, O. (ur.), *Computability: Turing, Gödel, Church, and Beyond*. Cambridge, MA, London: MIT Press, str. 261–327.

Block, N. (1995). »On A Confusion About a Function of Consciousness«. *Behavioral and Brain Sciences*, 18 (2), str. 227–247.

Boden, M. (2016). *AI: It's Nature and Future.* Oxford, New York: Oxford University Press.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Cantwell Smith, B. (2019). *The Promise of Artificial Intelligence: Reckoning and Judgment.* Cambridge, MA, London: MIT Press.

Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again.* Cambridge, MA, London: MIT Press.

Clark, A. (2001). *Mindware: An Introduction to the Philosophy of Cognitive Science*. Oxford, New York: Oxford University Press.

Copeland, J. (1993). *Artificial Intelligence: A Philosophical Introduction.* Blackwell.
Dreyfus, H. L. (1972). *What Computers Can't Do.* New York: MIT Press.
Dreyfus, H. L. in Dreyfus, S. E. (1991). »Making a Mind Versus Modelling the Brain: Artificial Intelligence Back at the Branchpoint«. Ponatisnjeno v Boden, M. (ur.), *Philosophy of AI.* Oxford: Oxford University Press.
Fodor, J. A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, Mass.: MIT Press.
Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. New York: MIT Press.
Hofstadter, D. (1979). *Gödel, Escher, Bach: an Eternal Golden Braid.* New York: Basic Books.
Hofstadter, D. R. in Dennett, D. C. (1990). *Oko duha: fantazije in refleksije o jazu in duši*. Ljubljana: Mladinska knjiga.
Kurzweil, R. (2005). *The Singularity Is Near: When Humans Transcend Biology.* New York: Viking, Penguin group.
Lucas, J. R. (1961). »Minds, Machines and Gödel«. *Philosophy*, 36, str. 112–127.
McCulloch, W. S. and Pitts, W. H. (1943/1990). »A Logical Calculus of the Ideas Immanent in Nervous Activity«. Ponatisnjeno v Boden, M. (ur.) (1990), *The Philosophy of Artificial Intelligence*, Oxford: Oxford University Press, str. 22–39.
Mindt, G., Montemayor, C. (2020). »A Roadmap for Artificial General Intelligence: Intelligence, Knowledge, and Consciousness«. *Mind and Matter*, 18 (1), str. 9–37.
Markič, O. (2011). *Kognitivna znanost: filozofska vprašanja*. Maribor: Aristej.
Miščević, N., Markič, O. (1998). *Fizično in psihično: uvod v filozofijo psihologije*. Šentilj: Aristej.
O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.
Penrose, R. (1995). *Shadows of the Mind.* London: Vintage.
Russell, S. in Norvig, P. (2010). *Artificial Intelligence A Modern Approach*. (3rd. ed.). Upper Saddle River: Prentice Hall.
Searle, J. (1990). »Duhovi, možgani in programi«. Prevedeno v Hofstadter, D. R. in Dennett, D. (ur.), *Oko duha: fantazije in refleksije o jazu in duši*, Ljubljana: Mladinska knjiga, str. 361–379.
Turing, A. (1990). »Stroji, ki računajo, in inteligenca«. V Hofstadter, D. R. in Dennett, D. (ur.), *Oko duha: fantazije in refleksije o jazu in duši*, Ljubljana: Mladinska knjiga, str. 61–74.
Uršič, M. (1987). *Matrice logosa: filozofsko-logični eseji in študije.* Ljubljana: DZS.
Uršič, M. (2010). *Štirje časi: filozofski pogovori in samogovori. Jesen: tretji čas. Daljna bližina neba: človek in kozmos*. Ljubljana: Cankarjeva založba.
Uršič, M. (2015). *Štirje časi: filozofski pogovori in samogovori. Zima: četrti čas. Preludij: O sencah*. Ljubljana: Cankarjeva založba.
Uršič, M. (2018). *Shadows of Being: Four Philosophical Essays*. Newcastle upon Tyne: Cambridge Scholars Publishing.
Varela, F., Thompson, E., Rosch, E. (1991). *The Embodied Mind.* Cambridge, MA, London: MIT Press.
Vinge, V. (2003). »Technological Singularity«. http://cmm.cenart.gob.mx/delanda/textos/tech_sing.pdf