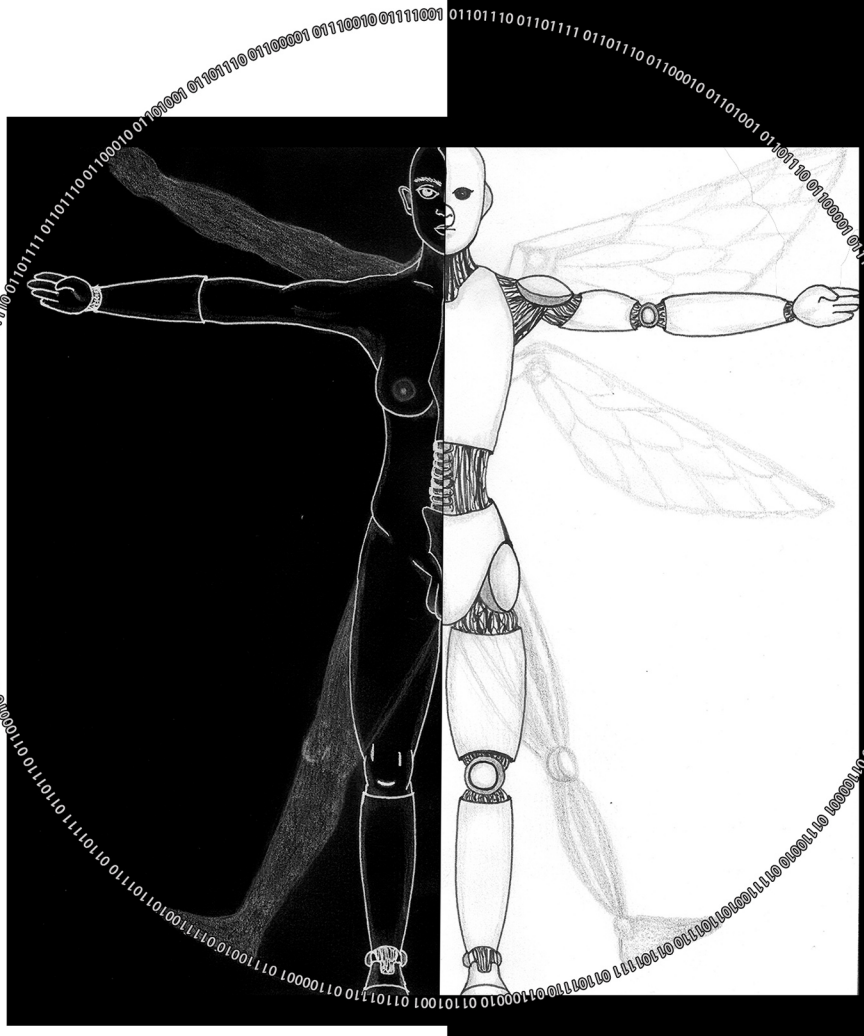
Univerzitetna založba
Univerze v Mariboru

Sodobne perspektive družbe: umetna inteligenca na stičišču znanosti



Uredniki: Janez Bregant, Boris Aberšek, Bojan Borstner

Univerza v Mariboru

Filozofska fakulteta

Sodobne perspektive družbe

Umetna inteligenca na stičišču znanosti

Uredniki
Janez Bregant
Boris Aberšek
Bojan Borstner

December 2022

Naslov / *Title*: **Sodobne perspektive družbe** / *Contemporary Perspectives of Society*

Podnaslov / *Subtitle*: **Umetna inteligenca na stičišču znanosti** / *Contemporary Perspectives of Society: Artificial Intelligence at the Intersection of Sciences*

Urediki / *Editors*: Janez Bregant (Univerza v Mariboru, Filozofska fakulteta); Boris Aberšek (Univerza v Mariboru, Fakulteta za naravoslovje in matematiko); Bojan Borstner (Univerza v Mariboru, Filozofska fakulteta)

Recenzija / *Review*: Nenad Miščević (Univerza v Mariboru, Filozofska fakulteta); Stanislav Avsec (Univerza v Ljubljani, Pedagoška fakulteta)

Lektoriranje / *Language editing*: Mihaela Koletnik (Univerza v Mariboru, Filozofska fakulteta)

Tehnični urednik / *Technical editor*: Jan Perša (Univerza v Mariboru, Univerzitetna založba)

Oblikovanje ovitka / *Cover designer*: Tadej Todorović

Grafika na ovitku / *Cover graphics*: Tina Ritlop, 2022

Grafične priloge / *Graphics material*: Avtorice in avtorji prispevkov, Bregant, Aberšek, Borstner 2022

Založnik / *Published by*: **Univerza v Mariboru, Univerzitetna založba**
Slomškov trg 15, 2000 Maribor, Slovenija
https://press.um.si, zalozba@um.si

Izdajatelj / *Issued by*: **Univerza v Mariboru, Filozofska fakulteta**
Koroška cesta 160, 2000 Maribor, Slovenija
https://www.ff.um.si, ff@um.si

Izdaja / *Edition*: Prva izdaja

Vrsta publikacija / *Publication type*: E-knjiga

Dostopno na / *Available at*: https://press.um.si/index.php/ump/catalog/book/737

Izdano / *Published*: Maribor, Slovenija, december 2022

CIP - Kataložni zapis o publikaciji
Univerzitetna knjižnica Maribor

004.8:1(082)(0.034.2)

SODOBNE perspektive družbe [Elektronski vir] : umetna inteligenca na stičišču znanosti / uredniki Janez Bregant, Boris Aberšek, Bojan Borstner. - 1. izd. - E-zbornik. - Maribor : Univerza v Mariboru, Univerzitetna založba, 2022

Način dostopa (URL): https://press.um.si/index.php/ump/catalog/book/737
ISBN 978-961-286-675-4 (PDF)
doi: 10.18690/um.ff.11.2022
COBISS.SI-ID 132725251

ISBN 978-961-286-675-4 (pdf)
978-961-286-676-1 (trda vezava)

DOI https://doi.org/10.18690/um.ff.11.2022

Cena
Price Brezplačni izvod

Odgovorna oseba založnika
For publisher prof. dr. Zdravko Kačič,
rektor Univerze v Mariboru

Citiranje
Attribution Bregant, J., Aberšek, B., Borstner, B. (ur.). (2022). *Sodobne perspektive družbe: umetna inteligenca na stičišču znanosti.* Univerza v Mariboru, Univerzitetna založba. doi: 10.18690/um.ff.11.2022

Univerzitetna založba
Univerze v Mariboru

Kazalo

Univerzitetna založba
Univerze v Mariboru

Uvod: za kaj sploh gre?

JANEZ BREGANT

Umetna inteligenca je danes tako močno prisotna v našem življenju, da si družbe brez nje sploh več predstavljati ne moremo. V resnici je postala polnokrven del v osnovi fizične narave sveta, v katerem živimo, je nekaj tako samoumevnega, da imamo občutek, kot da je z nami že od nekdaj. Njena vpetost v naše vsakodnevno početje, ki sega recimo od uporabe spletnih iskalnikov, navigacijskih naprav in mobilnih aplikacij, pa vse do spletnih nakupov, učenja na daljavo in nudenja pomoči uporabnikom, ta vtis samo še utrjuje. Kljub temu pa se zdi, da smo (spet) pred vrati tehnološke revolucije: stroji, ki se učijo, ne bodo prevzeli zgolj umazanih in zdravju škodljivih služb, kot se je to v različnih fazah industrializacije dogajalo v preteklosti, ampak bodo nadomestili ljudi tudi tam, kjer se zahtevajo npr. ustvarjalnost, načrtovanje, analiza in sinteza. Tako se lahko (v kolikor se že ni) kmalu zgodi, da bodo o odobritvi kredita za nakup hiše ali stanovanja, medicinskih diagnozah in postopkih zdravljenja, o tem, kdo je upravičen do socialne pomoči in kakšna naj bo višina kazni glede na verjetnost ponovitve zločina, odločali (samo) še stroji. Zato ni presenetljivo, da v nas to, kar nas čaka v prihodnosti, vzbuja tako čudenje kot strah.

Kaj umetna inteligenca (UI) sploh je? Vprašanje, na katerega verjetno ni enotnega odgovora, saj se zdi, da je ta odvisen od tega, na kaj z uporabo tega pojma mislimo (največkrat kar na »pametna« orodja in naprave) in o kateri fazi razvoja UI sploh govorimo. Poleg tega je meja med človeško in umetno inteligenco zaradi tega vedno bolj zabrisana, saj na, po prevladujočem mnenju, razliko med človekom, ki misli oziroma je zavestno bitje, in strojem, ki to ni, nezavestno pozabljamo. John McCarthy, eden izmed prvih raziskovalcev UI, je nekoč dejal: »Kakor hitro deluje, je nihče več ne imenuje UI.« (Dengel 2019) Kakorkoli, ne moremo mimo dejstva, da je pojem UI sestavljen iz pojma 'umetna' in pojma 'inteligenca'. Glede tega, kaj v tej besedni zvezi pomeni pojem 'umetna', velike dileme ni, gre za operacije, ki jih ne izvaja človek, ampak stroj/sistem (tj. nekaj, kar ni rezultat narave, ampak dela človeka ali stroja).

O tem, kaj v tej frazi pomeni pojem 'inteligenca', pa zaradi naše povsem različne predstave o tem, kdo je inteligenten, pravega soglasja ni: za enega je inteligenten nekdo, ki dobro in hitro računa ter tako kaže nadarjenost za matematiko, za drugega nekdo, ki si hitro zapomni besedilo, se ga dobro spomni in zna ideje iz njega prenesti v prakso, za tretjega nekdo, ki je sposoben prepoznati sogovornikova čustva, jih razumeti in se nanje ustrezno odzvati, za četrtega pa nekdo, ki se je sposoben usklajevati z drugimi, da bi od tega vsi imeli koristi. Vidimo, da smo v resnici opisali štiri povsem različne vrste sposobnosti (lahko pa bi jih še več), vse pa razumemo kot lastnosti, ki jih v večjem ali manjšem obsegu (ni nujno niti, da vse) imajo bitja, ki so inteligentna. Lahko bi govorili o neke vrste matematični inteligenci, jezikovni inteligenci, čustveni inteligenci, socialni inteligenci itd. (Dengel 2019; Bregant 2019)

Ker je 'inteligenca' tako širok pojem, ni čudno, da ni enotne splošne definicije, ki bi na višji ravni vključevala vse, za kar mislimo, da bi inteligentna oseba morala imeti. Pomeni lahko, (i) kako se znajdemo v novih situacijah in kako se jim znamo prilagoditi, (ii) kako znamo razmišljati in reševati probleme, (iii) kako se znamo učiti iz knjig in iz izkušenj, (iv) kako znamo opazovati okolje in razumeti dogodke itd. Psihologi kot definicijo 'inteligence' ponujajo naslednje: a. sposobnost posameznika, da se obnaša skladno z namenom, razmišlja racionalno in učinkovito obvladuje svoje okolje (Wechsler 1944), b. rezultat procesa pridobivanja, hranjenja, spominjanja, kombiniranja, primerjanja in uporabe informacij ter konceptualnih sposobnosti v novem okolju (Humphreys 1979), c. k cilju usmerjeno prilagojeno obnašanje (Sternberg in Salter 1982), d. sposobnost obvladovanja kognitivne kompleksnosti (Gottfredson 1998), e. sposobnost posameznika, da svoje mišljenje zavestno

prilagodi novih zahtevam, nalogam in življenjskim pogojem (Stern 1912) itd. (Bregant 2019)

V resnici gre za izvajanje kompleksnih kognitivnih/umskih/spoznavnih operacij, ki vključujejo matematične sposobnosti, komunikativnost, jezikovne sposobnosti, zaznavo, motorične sposobnosti, socialno razvitost in učenje.[1] Zaradi tega je UI ob svojem rojstvu leta 1956 na slavni Darthmouthski konferenci Marvin Minsky opisal kot »znanost o izdelavi strojev, ki so sposobni narediti stvari, za katere je po naših merilih potreben um«. (Copeland 1993: 1)[2] Gre za obdobje po 2. sv. vojni, ko so se pojavili prvi računalniki za splošno uporabo, ki jih je javnost imenovala »elektronski možgani«, in za čas, ko je kot preizkusni kamen inteligentnosti kakršnegakoli sistema štelo uspešno igranje salonskih iger (npr. dame). Cilj nove discipline ni bil zgolj posnemati mišljenja, ampak narediti napravo, ki je inteligentna, v polnem pomenu te besede. V tem t. i. prvem valu razvoja UI[3] je raziskovalce tako zanimalo predvsem vprašanje »Ali lahko stroj misli?« Zanimivo, to vprašanje se ni prvič pojavilo šele v 20. stol., ko so bili računalniki zreli za masovno proizvodnjo, ampak že daljnega leta 1637, ko je Descartes trdil, da noben stroj (izraz, ki ga je sam uporabil, je bil 'avtomat') nikoli ne bo sposoben simbolne manipulacije, ki bi mu omogočala tvorjenje smiselnih stavkov, na osnovi česar bi mu lahko pripisali mišljenje. (Descartes 1637/2007)

Tako je šele po stoletjih čakanja s prihodom prvih elektronskih računalnikov ideja o strojih, ki računajo (mislijo), v 50. letih 20. stol. dobila nov zagon. Začelo se je zlato obdobje umetne inteligence, ki je trajalo nekje do sredine 70. let 20. stol., v njem pa so raziskovalci postavili temelje (računalniškega) sklepanja, razvili prve nevronske mreže, ki so bile zasnovane po vzoru bioloških, in z vidika simuliranja inteligence dosegli prve uspehe. Sledilo je obdobje zatona, saj prvotni algoritmi, ki so delovali na enostavnih problemih in se nanašali na premikanje kock, iskanje poti v labirintu ali orientacijo v prostoru, niso bili uporabni za reševanje problemov, s katerimi se srečujemo v vsakdanjem življenju. Metode njihovega delovanja, ki so temeljile zgolj na sklepanju, niso bile primerne za reševanje kompleksnejših nalog, metode, ki so temeljile na statističnih zakonitostih, pa so zahtevale veliko podatkov in računske

[1] V splošnem bi lahko rekli, da gre za zaznavanje, spominjanje in učenje, da lahko to, kar smo doživeli, posplošimo in uporabimo za reševanje konkretnih življenjskih problemov, tj. da se znamo v družbenem okolju smiselno obnašati.

[2] Ali kot »znanost o tem, kako narediti in/ali programirati računalnike, da bodo sposobni istih stvari kot um«. (Boden 1990: 1).

[3] Razdelitev razvoja UI na dva vala uvaja Cantwell Smith (2019) in povzema tudi npr. Markič (2019).

moči, česar pa takrat še ni bilo na voljo. Ponovni vzpon UI se je začel v začetku 90. let 20. stoletja, ko so se pojavili šahovski programi, ki so bili kmalu sposobni premagovati šahovske velemojstre. (Bregant 2019)

Danes, v t. i. drugem valu razvoja UI, pa je fokus raziskovalcev bolj praktičen, gre za razvijanje pametnih orodij v smislu strojne in programske opreme, ki jih v večji ali manjši meri uporabljamo skorajda na vseh področjih našega življenja. Njen hiter razvoj, ki botruje temu, da skoraj ne mine dan, ko se v javnosti ne bi pojavila kakšna novica v zvezi z njo, gre pripisati predvsem naslednjim trem faktorjem:[4]

> a. povečana računska moč strojev (računalnikov) (procesorji zmorejo danes v 1 sek. opraviti 2 X 1015, tj. 2 bilijardi, elementarnih računskih operacij, kot sta seštevanje in množenje, zapisanih v obliki, ki jih računalnik razume, t. i. operacij s plavajočo vejico;
> b. dostopnost do ogromnega števila (digitalnih) podatkov kot posledica razvoja znanosti in tehnologije, gospodarske dejavnosti in aktivnosti na družbenih omrežjih (s pomočjo t. i. spleta stvari je med seboj povezano ogromno število naprav, ki na ta način proizvedejo izjemno količino podatkov, ocenjuje se, da na dan do 2,4 trilijona);
> c. novi izdelki, ki temeljijo na obdelavi omenjenih podatkov in zaradi katerih se pojavljajo nova pravna in etična vprašanja (proizvodnja dronov, avtonomnih vozil, virtualnih asistentov itd.). (Dengel 2019; Bregant 2019)

Cilj razvijalcev algoritmov je bil sicer že od nekdaj narediti takšen program, ki bo sposoben samostojno prepoznati, analizirati in rešiti probleme, v katerih se bo znašel, ne da bi pri tem morali vsak njegov korak že vnaprej določiti. V idealnem primeru bi se umetni sistemi morali znati prilagajati (nenapovedanim ali nepredvidenim) spremembam v okolju, sprejemati ustrezne (pravilne) odločitve, si jih zapomniti ter jih v podobnih situacijah ponovno uporabiti ter se tako učiti. Prednost UI se vedno kaže tam, kjer je treba iz velike in neurejene količine podatkov razbrati (ustrezne ali zahtevane) vzorce. V korist strojnega odločanja se običajno navaja argument, da se s tem iz postopka sprejemanja odločitev izloči predsodke, ki pri človeškem odločanju pomembno vplivajo na rezultat. Ali to drži? Ne. Navedemo lahko vsaj tri značilnosti programov, ki objektivnost strojnega odločanja postavljajo

[4] Podatki so iz leta 2019 (danes so številke že mnogo višje) in služijo zgolj ponazoritvi neverjetne moči in hitrosti računalnikov.

pod vprašaj: (a) razložljivost (transparentnost), (b) neprepričljivost ali nelogičnost in (c) pristranskost.

O problemu razložljivosti govorimo zato, ker algoritmi uporabnikom svoje odločitve večinoma ponujajo brez kakršnekoli pojasnitve, kako so do njih sploh prišli. To postane očitno predvsem takrat, ko odločitve algoritmov neposredno vplivajo na kakovost našega življenja ali v nekaterih primerih celo na preživetje, npr. dostop do kredita, zdravstvenih in socialnih storitev ali službe. Problem razložljivosti pogosto izhaja iz tega, da so algoritmi, ki se uporabljajo, tako kompleksni, da njihovih odločitev do potankosti ne razumejo niti programerji. Takšni t. i. »algoritmi črne škatle«[5] se med drugim uporabljajo tudi pri analiziranju naše kreditne sposobnosti, kjer v resnici stroj odloči, ali bomo kredit dobili ali ne. Razlogov za njegovo odločitev pa žal ne zna zadovoljivo pojasniti nihče.

Odločitve, ki jih sprejemajo stroji, pa so lahko tudi neprepričljive ali nelogične. To se največkrat zgodi takrat, ko algoritmi ne poznajo razlike med korelacijo in vzročnostjo. Predstavljajte si, da bi v zdravstvu algoritem ugotovil, da je bilo med bolniki, ki so zboleli za rakom, jemali določeno zdravilo in ozdraveli, veliko levičarjev, potem pa bi vsem levičarjem, ki bi v prihodnosti zboleli za rakom, predpisal isto zdravilo. V tem primeru vsekakor ne bi šlo za razumljivo odločitev, saj upravičitev algoritma temelji na njegovi zamenjavi korelacije z vzročnostjo. (Ogola 2019)

Problem pristranskosti pa se pojavi zato, ker algoritmi v podatkih iščejo tiste vzorce, ki so se jih naučili reproducirati v času svojega urjenja. In če v tem času podatki, na katerih se učijo, niso popolni, tudi njihove odločitve ne bodo objektivne. Takšna pristranskost je botrovala krivičnim odločitvam v ameriškem pravosodnem sistemu, ko so zaradi prepolnih zaporov iz njih želeli predčasno izpustiti določeno število obsojencev. Pri izračunu tega, kakšna je verjetnost, da zapornik, če je prej izpuščen na prostost, spet stori kakšno kriminalno dejanje, so se zanašali na algoritme. In ti so izračunali, da je verjetnost za to pri temnopoltih zapornikih dvakrat višja kot pri belopoltih. Tako so imeli belopolti zaporniki dvakrat večjo možnost za predčasni izpust od temnopoltih. Podatki, na katerih so se ti algoritmi učili, so bili rasistični, kar je (še enkrat več) vodilo v diskriminacijo na osnovi barve kože.[6]

[5] Angl. *Black Box Algorithms*.
[6] Za kritiko nepremišljenih odločitev, ki so posledica pristranskosti, glej O'Neil (2016).

Vidimo, da tehnologija niti slučajno ni nevtralna. Algoritem je dober le toliko, koliko so dobri podatki, ki jih obdeluje, pri čemer lahko predsodke, ki se zrcalijo v zbranih informacijah, posvoji do te mere, da postane določen družbeni problem še večji. Iz (pogosto) zgolj navidezno objektivnih odločitev umetnih sistemov pa izhaja kup praktičnih moralnih vprašanj: »Kdo je odgovoren za odločitve, ki jih sprejmejo avtomatizirani umetni sistemi?« (Bryson 2018), »Ali imamo pravico vedeti, na osnovi česa stroji sprejemajo svoje odločitve?«, »Ali obstajajo bitja z višjim moralnim statusom od ljudi?« (Agar 2012), »Ali se dajo moralne odločitve, ki jih človek sprejema intuitivno, sprogramirati?«, »Kako umestiti UI v svet tako, da bo ravnala v skladu z našimi pričakovanji?« (Rahwan 2018), »Ali sme UI, razvita v azijskem okolju, sprejemati moralne odločitve v evropskem okolju?« itd.

Splošno vprašanje je, katere moralne kriterije mora UI izpolniti, da bo pri sprejemanju svojih odločitev, ki pomembno vplivajo na naše življenje (npr. prijava za službo, odobritev kredita, odločitev o načinu zdravljenja itd.), pravična? Politične smernice, ki naj bi olajšale usklajevanje strojnega odločanja s pravom in moralo, je na ravni Evropske unije pripravila t. i. Strokovna skupina za UI, ki jo je leta 2018 imenovala Evropska komisija. Njen cilj je bil določitev etičnih standardov, s katerimi naj bi se povrnilo zaupanje ljudi v UI.[7] Poročilo vključuje naslednje zahteve:

> (i) nadzor (umetni sistemi morajo človeku pomagati pri sprejemanje argumentiranih odločitev, skladnih z njegovimi pravicami in svoboščinami; poleg tega morajo vključevati tudi primerne nadzorne mehanizme, ki se pri svojem delu zgledujejo po človeškem ravnanju);
> (ii) varnost (umetni sistemi morajo biti robustni in varni, vključevati morajo rezervni načrt v primeru, da gre kaj narobe, pa tudi natančni in zanesljivi);
> (iii) zaščita podatkov (umetni sistemi morajo zagotavljati in spoštovati pravico do zasebnosti; poleg tega morajo vključevati tudi mehanizme upravljanja s podatki, ki zagotavljajo kakovost njihove obdelave in omogočajo dostop do njih vsem, ki imajo do tega pravico);
> (iv) transparentnost (umetni sistemi morajo vključevati mehanizme za sledljivost sprejetih odločitev, človek pa mora vedeti, kdaj je v interakciji s takšnim sistemom in česa je ta zmožen);

[7] Žal empirične raziskave kažejo, da so etične smernice zgolj brezzobi okras in nimajo pomembnega vpliva na razvoj nove programske opreme. (Hagendorff 2020)

(v) nediskriminacija (umetni sistemi se morajo izogibati nepravičnim odločitvam na osnovi predsodkov, ker to povečuje marginalizacijo že tako odrinjenih skupin, dostopni pa morajo biti vsem skozi svoje celotno življenjsko obdobje);
(vi) trajnostna orientacija (umetni sistemi morajo biti prijazni do okolja, od njih naj bi v končni fazi imelo korist več generacij, njihov vpliv na vsa v naravi živeča bitja pa mora biti skrbno preverjen in premišljen);
(vii) odgovornost (razviti je treba mehanizme, s katerimi se zagotavlja in preverja moralna in/ali pravna odgovornost umetnih sistemov ter njihovih dejanj; poleg tega je treba priskrbeti ustrezna pravna sredstva, ki se uporabljajo v primeru morebitnih kršitev). (Ogola 2019; Bregant 2020)

Pričujoč zbornik na takšen in drugačen način nagovarja nekatera izmed pravkar omenjenih vprašanj in dilem. V splošnem smislu gre za temo, ki jo lahko povzamemo kot prednosti in slabosti hitrega razvoja UI z vidika njenega vpliva na moralo, psihologijo in izobraževanje. Vsebuje štirinajst člankov, razdeljenih v dva vsebinska dela, pri čemer se prvi osredotoča na presek med UI, filozofijo in etiko, drugi pa na presek med UI, psihologijo in izobraževanjem.

V prvem sklopu, imenovanem *Umetna inteligenca, filozofija in etika*, najprej Janez Bregant v prispevku *Od Descartesa do Alexe: filozofski pogled na razvoj umetne inteligence* oriše filozofski pogled na razvoj UI s filozofske perspektive, ki se razteza od Descartesovih avtomatov pa vse do Amazonove Alexe, pri čemer izpostavi dva t. i. vala UI, prvega bolj zgodovinskega, kjer je v ospredju stalo bolj kot ne filozofsko vprašanje, ali je možno narediti stroj, ki bi mu lahko pripisali mišljenje oziroma zavest, in drugega, ki traja še danes, kjer je pozornost usmerjena v razvijanje modelov UI, ki na takšen ali drugačen način izboljšujejo kakovost našega življenja.

Tadej Todorović skuša v članku *Biti ali (le) bit: kdaj je umetna inteligenca zavestna, kdaj inteligentna in ali obstaja razlika?* razjasniti vprašanje, kaj mislimo s tem, da umetna inteligenca misli ali čuti, pri čemer zagovarja prepričanje, da se v filozofskih razpravah koncept 'umetne inteligence' uporablja neustrezno in na trenutke zavajajoče.

Olga Markič v besedilu *Transparentnost in razložljivost kot zahtevi za zaupanja vredno UI* opozarja na pomanjkljivosti, ki jih imajo orodja drugega vala UI, s katerimi se modelirajo programi, ki napovedujejo dogodke oziroma se o prihodnosti odločajo namesto nas, saj modeli strojnega učenja večinoma ne temeljijo na človeku razumljivem logičnem sklepanju, pa tudi v zastavljenih ciljih njihovih učnih primerov se pogosto odražajo družbene vrednote programerjev in družbeni kontekst, v katerega so ti vpeti, zaradi česar se avtorica vpraša, v kolikšni meri današnji sistemi sploh lahko izpolnijo t. i. etične smernice za zaupanja vredno UI.

Borut Trpin v prispevku *Raziskovalna umetna inteligenca in standardi transparentnosti* s primerom iz računalniške filozofije, tj. filozofije, kjer računalniške metode igrajo osrednjo vlogo kot metode filozofske argumentacije, pokaže, da lahko zanašanje na zgolj na videz smiselne rezultate oziroma rezultate, ki jih ne razumemo dovolj, vodi v zmotne zaključke, s čimer dokazuje potrebo po višjem standardu transparentnosti oziroma razložljivosti pri raziskovalni UI (v primerjavi z naravno inteligenco) oziroma vsaj potrebo po večjem preverjanju robustnosti takšnih rezultatov.

Niko Šetar in Bojan Borstner v članku *Ali nas umetna inteligenca lahko premaga: od algoritma do singularnosti po poteh etičnega vrednotenja* najprej predstavita argument za to, da je razprava o etiki UI smiselna, potem pa preko nizanja etičnih problemov (od nižje UI preko človeku podobne UI do superinteligence) dokazujeta, da je zaradi narave strojnega učenja, ki je bistven element sodobnih modelov UI, treba rešitve za etične dileme, povezane z njo, iskati v okviru teorije utemeljevanja.

Friderik Klampfer v besedilu *Etične in politične dileme programiranja samovoznih avtomobilov za odločanje o neizogibni škodi* opozarja na to, da je treba v algoritme, ki upravljajo z avtonomnimi vozili, nujno vgraditi moralna načela, ki uravnavajo t. i. porazdelitev neizogibne škode, tj. sprejemanje odločitev v situacijah, ko se mora avto odločiti, ali naj npr. povozi več pešcev na prehodu za pešce ali pa namesto tega zavije na pločnik in tam ubije enega ali npr. katera življenja naj ohrani, življenja potnikov v vozilu ali drugih udeležencev v prometu itd.

Tomaž Grušovnik pa v zadnjem prispevku prvega sklopa z naslovom *Avtonomna orožja in erozija moralne odgovornosti: sistematizacija nasilja in izginjanje pravičnosti* svari, da bi lahko erozija moralne odgovornosti zaradi ponikanja individualnega in osebnega moralnega vršilstva v avtonomnih sistemih predstavljala nekaj takšnega kot dodatno

sistematizacijo nasilja, kjer bi algoritmi, o katerih bi lahko odločala le peščica močnih posameznikov, narekovali smer družbenega razvoja.

Drugi sklop, imenovan *Umetna inteligenca, znanost in izobraževanje*, odpirajo Izidor Mlakar, Simona Majhenič in Matej Rojc s člankom *Virtualni pogovorni agent EVA, umetna inteligenca za bolj naravno interakcijo z napravami*, v katerem se ukvarjajo z vprašanjem, kako ustrezno modelirati interakcijo med človekom in strojem, pri čemer je njihov cilj razviti model UI za ustvarjanje človeku podobnega pogovora in najti rešitev za čustveno in poosebljeno interakcijo med človekom in strojem.

Urška Martinc, Boris Aberšek in Bojan Borstner v besedilu *Ali je inteligenca roja prihodnost za umetno inteligenco?* raziskujejo povezavo med inteligenco roja (angl. *swarm intelligence*) in UI, pri čemer inteligenco roja definirajo kot kolektivno obnašanje nekega samoorganiziranega sistema, ki ga sestavljajo številni homogeni posamezniki, med katerimi poteka interakcija po preprostih vedenjskih pravilih, ki izkoriščajo le lokalne informacije, ki si jih posamezniki izmenjujejo neposredno ali prek okolja (stigmergija), ter dokazujejo, da velja isto tudi za umetne sisteme, tj. da lahko kot inteligenco roja obravnavamo tudi UI, pri čemer razčlenijo koncept inteligence roja in utemeljijo temeljne gradnike takih sistemov.

Alen Lipuš v prispevku *Umetna inteligenca in eksistenčno tveganje* problematizira vprašanje nadzora v razvoju modelov UI, pri čemer opozarja na to, da je treba vztrajati pri zahtevi, da se UI normativno omeji, saj je možno le tako zmanjšati eksistencialno tveganje in s tem povečati njeno usklajenost z najvišjimi družbenimi vrednotami.

Bojan Musil in Nejc Plohl v članku *Po sledeh umetne inteligence: Kaj nam o psiholoških lastnostih posameznikov povedo njihovi digitalni odtisi?* dokazujeta, da lahko s pomočjo strukturne analize digitalnih odtisov pridemo do dokaj natančnih napovedi posameznikove osebnosti, pri čemer poseben poudarek namenita tudi praktični uporabi tako pridobljenih osebnostnih podatkov, ki se prevečkrat uporabljajo na moralno sporen način, in sicer v smislu prilagajanja oglasov z namenom vplivanja na naše odločitve in preference.

Nenad Čuš Babič se v besedilu *Socialni odnosi, animizem, antropomorfizem in interakcija z UI* osredotoči na vprašanje, ali lahko modele UI, s katerimi smo v interakciji, interpretiramo animistično, pri čemer najprej predstavi psihološko razumevanje namena, funkcije in sprožilcev animističnih in antropomorfnih atributov, potem pa raziskuje vplive in posledice antropomorfne interpretacije UI na zaznavanje, čustvovanje, mišljenje ter delovanje ljudi.

Igor Pesek in Marjana Krašna v prispevku *Vloga umetne inteligence v izobraževanju in za izobraževanje* dokazujeta, da lahko danes UI pismenost, ki omogoča sprejemanje premišljenih odločitev o tem, kako naj bo UI vključena v naše življenje, dosežemo s personaliziranim učenjem, povezovanjem in ustvarjanjem pametnih učnih vsebin, izvajanjem tutorstva v inteligentnih tutorskih sistemih, pomočjo učencem s posebnimi potrebami, dostopom do učnih vsebin in prevajanjem izobraževalnih vsebin iz različnih jezikov.

Andrej Flogie in Boris Aberšek pa v zadnjem članku drugega sklopa in celotnega zbornika z naslovom *Umetna inteligenca in prihodnost učenja in poučevanja* raziskujeta, kako UI, strojno učenje in sorodne računalniške tehnologije vplivajo tako na družbo kot celoto kakor tudi na izobraževanje in prihodnost učenja, in poudarjata, da je skrajni čas, da začnemo načrtovati, razvijati in uporabljati UI v izobraževanju na načine, ki so učinkoviti, pravični do posameznika in etični ter po možnosti brez slabosti, tveganj in škod.

Zbornik črpa svojo verodostojnost iz interdisciplinarnosti, saj uporaba vsaki vedi lastne raziskovalne metodologije v člankih pripomore k njihovi svojstveni obogatitvi v smislu vpliva UI na moralo, transparentnost in uporabnost. Ukvarjanje z raziskovalnimi problemi na opisan način pa poskrbi za raznolike, pestre in domiselne prispevke, ki pri ponujanju rešitev pogosto prestopijo meje ustaljenega.

Viri in literatura

Agar N. (2012). »Why is it possible to enhance moral status and why doing so is wrong?«. *Journal of Medical Ethics*, 39, str. 67–74.

Boden, M. A. (1990). »Introduction«. V Boden, M. A. (ur.), *The Philosophy of Artificial Intelligence*. Oxford: Oxford University Press.

Bregant, J. (2019). »Umetna inteligenca v praksi. Del 1, Razvoj, obnašanje in učenje strojev«. *Analiza,* 2, str. 39–55.

Bregant, J. (2020). »Umetna inteligenca v praksi. Del 2, Nekaj etičnih pomislekov«. *Analiza,* 1, str. 5–20.

Bryson, J. J. (2018). »Patiency is not a virtue: the design of intelligent systems and systems of ethics«. *Ethics and Information Technology*, 20, str. 15–26.

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

Copeland, J. (1993). *Artificial Intelligence: A Philosophical Introduction*. Oxford: Blackwell.

Dengel, A. (2019). »Künstliche Intelligenz – Eine Einführung«. V Dengel, A., Socher, R., Kirchner E. A., Ogolla, S. (urd.), *Künstliche Intelligenz: Die Zukunft von Mensch und Maschine*. Hamburg: ZEIT Akademie Gmbh, str. 13–22.

Descartes, R. (1637/2007). *Razprava o metodi*. Ljubljana: Slovenska matica.

European Commission's High-Level Expert Group on Artificial Intelligence (2018). *Ethics guidelines for trustworthy AI*.

Gottfredson, L., (1998). »The General Intelligence Factor«. *Scientific American Presents*, 9, str. 24–29.

Hegendorff, T. (2020). »The Ethics of AI Ethics: An Evaluation of Guidelines«. *Minds and Machines*, 30, str. 99–l 20.

Humphreys, L. G. (1979). »The construct of general intelligence«. *Intelligence*, 3, str. 105–120.

Markič, O. (2019). »Prvi in drugi val umetne intelgience«. V Malec, M., Markič, O. (urd.*) Misli svetlobe in senc: razprave o filozofskem delu Marka Uršiča*. Lubljana: UL, str. 201–211.

Ogola, S. (2019). »Verantwortung, Erklärbarkeit und Transparenzalgorithmischer Entscheidungen«. V Dengel, A., Socher, R., Kirchner E. A., Ogolla, S. (urd.), *Künstliche Intelligenz: Die Zukunft von Mensch und Maschine*. Hamburg: ZEIT Akademie Gmbh, str. 93–101.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens democracy*. Crown: New York.

Rahwan, I. (2018). »Society-in-the-loop: programming the algorithmic social contract«. *Ethics and Information Technology*, 20, str. 5–14.

Stern, William (1912) *Die psychologischen Methoden der Intelligenzprüfung: und deren Anwendung an Schulkindern*. Leipzig: J. A. Barth.

Sternberg, R. J., Salter W (1982). *Handbook of human intelligence*. Cambridge, UK: Cambridge University Press.

Wechsler, D (1944). *The measurement of adult intelligence*. Baltimore: Williams & Wilkins.

1. DEL

UMETNA INTELIGENCA, FILOZOFIJA IN ETIKA

OD DESCARTESA DO ALEXE: FILOZOFSKI POGLED NA RAZVOJ UMETNE INTELIGENCE

JANEZ BREGANT
Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
janez.bregant@um.si

Sinopsis Mehanicističen odgovor na vprašanje o tem, kaj je mišljenje, sega v 17. stol., ko je Thomas Hobbes zapisal, da ni nič drugega kot računanje. Čeprav se je mogoče sam zdel naklonjen takšni materialistični ideji, ki lastnost duha formulira v pojmih mehanskih procesov, pa ni bilo jasno, kako naj bi to potekalo. Začetki simulacije inteligentnega obnašanja segajo v 60. leta 20. stol., ko je bil v modi t. i. simbolni pristop k modeliranju inteligence, ki je temeljil na deduktivnem sklepanju. Kmalu se je izkazalo, da vsega znanja, ki bi ga stroji rabili za to, da bi uspešno opravljali tudi vsakodnevne kompleksnejše naloge, kljub vsemu ne moremo predstaviti v strukturah, podobnim stavkom. Tako se je razvila nesimbolna umetna inteligenca (UI), ki pri simulaciji inteligentnega obnašanja stavi na induktivno sklepanje, verjetnost in prepoznavanje vzorcev. Danes se zdi, da je pri razvoju različnih aplikacij meja samo še nebo, vprašanje pa je za kakšno ceno. V članku najprej predstavimo vizionarje, ki so predvideli nastanek strojev in se spraševali, ali bodo ti kdaj mislili, potem opišemo prvi val razvoja UI, ki razume mišljenje kot simbolno manipulacijo, nato pa še drugega, ki inteligenco modelira s pomočjo nevronskih mrež, ki se urijo z učnimi primerki.

Ključne besede:
umetna inteligenca,
simbolna UI,
nesimbolna UI,
nevronske mreže,
strojno učenje,
rudarjenje
podatkov

Univerzitetna založba
Univerze v Mariboru

DOI https://doi.org/10.18690/um.ff.11.2022.1
ISBN 978-961-286-675-4

FROM DESCARTES TO ALEXA: A PHILOSOPHICAL VIEW ON THE EVOLUTION OF ARTIFICIAL INTELLIGENCE

JANEZ BREGANT
University of Maribor, Faculty of Arts, Maribor, Slovenia
janez.bregant@um.si

Abstract The mechanicist answer to the question of what is thinking goes back to the 17th century when Thomas Hobbes wrote that it is nothing but calculating. Despite being fond of such a materialist idea, which formulates the mind in terms of mechanical processes, it was not clear how this was supposed to work. The beginning of the intelligent behaviour's simulation dates to the 1960s when the symbolic approach based on deductive reasoning was at work. It soon turned out that not all knowledge needed by the machines to complete the complex tasks typical for our everyday life could be represented by structures that resemble sentences. This gave birth to the non-symbolic artificial intelligence (AI), which simulates intelligent behaviour by using inductive reasoning, probability, and pattern recognition. The article first introduces visionaries who predicted the arrival of machines and wondered if they could ever possibly think. It then describes the first wave of the AI's evolution, which understands thinking as symbol manipulation. This is followed by the outline of the second wave of AI's evolution, which models intelligence with the help of neural networks trained by learning examples.

Keywords: artificial intelligence, symbolic AI, non-symbolic AI, neural networks, machine learning, data mining

https://doi.org/10.18690/um.ff.11.2022.1 DOI
978-961-286-675-4 ISBN
University of Maribor Press

1 Uvod

Kaj je umetna inteligenca (UI)? Google Maps, Instagram, Twiter, Facebook, Dropbox, YouTube, Netflix, Google Translate, Siri, Alexa, iPhone, iPad, Mac itd., z drugimi besedami, spletni brskalniki, aplikacije za družbena omrežja in shranjevanje v oblakih, za predvajanje glasbe ali filmov, za pomoč v obliki virtualnih prevajalnikov in asistentov, pametni telefoni, tablice, računalniki itd. Meja med inteligenco, ki je naravna (človeška), in tisto, ki je umetna (strojna), je danes zabrisana, pogosta uporaba takšnih in drugačnih pametnih orodij, ki je prerasla v običajno prakso, pa ustvarja površinski vtis, da med njima ni razlike. Takšno pozabljanje na to, da je človeška inteligenca zavestna, strojna pa zgolj mehanska, če vse skupaj poenostavimo, omogoča UI, da se v družbi »skrije« in vzbuja občutek, da je z nami že od nekdaj.

Začetki tega, kar danes imenujemo UI, segajo v obdobje po drugi sv. vojni, ko so nastali prvi računalniki s programi, ki so bili zelo uspešni pri igranju npr. dame, ko je za inteligentno štelo obnašanje, ki je vključevalo opravljanje logičnih operacij. Najbolj znan program za igranje dame, ki se je že bil sposoben učiti, kar mu je omogočalo napredovanje in zmago nad svojim izumiteljem, je že na začetku 50. let 20. stol. razvil Arthur Samuel, deloval pa je na IBM-ovem računalniku 701. Za rojstvo UI kot discipline pa štejemo leto 1956, ko so se na Darthmouthski konferenci zbrali vsi raziskovalci strojne inteligence, ki so takrat kaj pomenili in se povezali v formalno družbeno skupnost s ciljem narediti stroj, ki je resnično inteligenten. Običajno se opisuje kot »znanost o izdelavi strojev, ki so sposobni narediti stvari, za katere je po naših merilih potreben um« (Copeland 1993: 1), ali kot »znanost o tem, kako narediti in/ali programirati računalnike, da bodo sposobni istih stvari kot um«. (Boden 1990: 1) V bistvu je šlo za izdelavo stroja (računalnika), ki je z namenom doseganja ciljev zmožen opravljati zapletene kognitivne operacije, povezane z obnašanjem, načrtovanjem, reševanjem problemov, napovedovanjem, sklepanjem itd., in operacije, ki vključujejo matematične sposobnosti, komunikativnost, jezikovne sposobnosti, zaznavo, motorične sposobnosti, socialno razvitost in učenje.[1]

[1] V splošnem bi lahko rekli, da gre za zaznavanje, spominjanje in učenje, da lahko to, kar smo doživeli, posplošimo in uporabimo za reševanje konkretnih življenjskih problemov, tj. da se znamo v družbenem okolju smiselno obnašati.

V tem obdobju bi lahko govorili o prvem valu razvoja UI,[2] raziskovalci so se ukvarjali z modeliranjem človeških psihičnih sposobnosti, kot so logično mišljenje, ustvarjalnost, uporaba jezika, učenje, reševanje problemov itd., kar bi strojem omogočalo avtonomno delovanje v negotovem svetu. To, včasih se imenuje zlato obdobje UI, je trajalo vse do sredine 70. let 20. stol., v njem pa so bili postavljeni temelji računalniškega sklepanja, razvite so bile prve nevronske mreže in doseženi prvi uspehi pri posnemanju človeške inteligence. Temu začetnemu optimizmu je sledilo obdobje streznitve, programi, ki so bili napisani za reševanje preprostih problemov, kot je npr. premikanje kock v prostoru in so temeljili na logičnem sklepanju v strogo določenem okolju, se niso obnesli pri reševanju zapletenih življenjskih problemov v spreminjajočem se svetu, ki je temeljilo na statističnih zakonitostih (verjetnosti) in zahtevalo veliko računsko moč za obdelavo velike količine podatkov.

Ponovni vzpon UI se je začel spet v 80. in 90. letih 20. stol., ko so bile zaradi vedno večje računske moči računalnikov, vedno večjega števila podatkov, ki so bili na voljo in vedno boljšega dostopa do njih, omenjene tegobe počasi odpravljene. Govorimo lahko o drugem valu razvoja UI, ki pa ne temelji več na dedukciji in pisanju navodil za manipuliranje s simboli kot prvi, ampak na indukciji in pisanju algoritmov za strojno učenje. (Markič 2019) Raziskovalci so svojo pozornost preusmerili na razvijanje pametnih orodij, ki delujejo po principu učenja na osnovi predhodnih izkušenj in interakcije z okoljem, svoj navdih pa črpajo iz povezovanja različnih disciplin, kot so kibernetika, teorija verjetnosti in statistika. Uporabljajo na podlagi učnih primerkov modelirane nevronske mreže, ki pa so še vedno specializirane zgolj za opravljanje nalog na nekem ozkem področju, npr. za prepoznavanje obrazov, prevajanje ali avtonomno vožnjo, tako da v tem trenutku še ne moremo govoriti o neki splošni UI, ki bi bila primerljiva s človeško, kjer gre za obvladovanje večih, med seboj tudi precej različnih področij in sposobnost reševanja nalog znotraj njih.

V članku se najprej zazremo v zgodovino in prikažemo avtorje, ki so že pred davnim časom mišljenje na takšen ali drugačen primerjali z računanjem, vizionarje, ki so predvideli nastanek strojev in se spraševali, ali bodo ti kdaj mislili, ter načrtovalce abstraktnih računalnikov, ki niso nikoli zagledali luči sveta. Potem podrobneje predstavimo prvi val razvoja UI, nekaj njegovih dosežkov in filozofsko vprašanje, ali nismo tudi ljudje zgolj računalniki, ki predstavlja vrh razmišljanja o podobnosti

[2] Razdelitev na dva vala razvoja UI predlaga Cantwell Smith (2019) in uvaja npr. Markič (2019).

med stroji in ljudmi. Temu sledi opis drugega vala razvoja UI z nekaterimi najbolj znanimi primeri umetnih sistemov danes s poudarkom na izgubi zasebnosti kot ceni, ki jo moramo z moralnega vidika plačati za tako hiter tehnološki razvoj. Končamo s kratkim povzetkom in navedbo virov ter literature.

2 Izlet v zgodovino

»Mišljenje ni nič drugega kot računanje,« pravi Hobbes leta 1651 v *Leviathanu* (Hobbes 1651/2006: 32). Verjetno gre za prvi opis določene mentalne operacije v pojmih računanja, ki vključuje simbolno manipulacijo, ki se dogaja v naših možganih. Primerki misli so tako v bistvu primerki možganov, ali kot jih imenuje Hobbes, 'fantazme', mišljenje pa neke vrste mentalni pogovor.

> Ko človek razmišlja ne dela ničesar drugega kot, da si zamišlja končno vsoto, ki jo dobi s seštevanjem delov, ali ostanek, ki ga dobi z odštevanjem ene vsote od druge. /…/ Te operacije se ne dogajajo samo pri številih, ampak pri vseh načinih stvari, ki jih lahko združujemo ali odvzemamo. (Hobbes 1651/2006: 32)

Ko razmišljamo, podobno kot pri računanju s pomočjo svinčnika in papirja, uporabljamo simbolne operacije, le da te niso izražene z govorjenimi ali pisanimi simboli, temveč v posebnem nevralnem zapisu/kodi. (Markič 2019) Haugeland (1986) pravi, da je pri Hobbsu mišljenje mehanski proces, podoben upravljanju mentalnega abaka: fantazme premetavamo sem in tja skladno s pravili razuma, podobno kot na pravem abaku skladno z računskimi pravili sem in tja premikamo kroglice. Hobbsova izvirnost se kaže tudi v tem, da je predpostavil obstoj umetnih sistemov, ker njegov materializem, po katerem življenje ni nič drugega kot gibanje telesnih organov, ki je podobno gibanju zobnikov, vzmeti in koles pri stroju, obstoj takšnih *avtomatov*, kot jih sam imenuje, tudi dopušča. »Kaj pa je srce drugega kot vzmet, živci strune in sklepi kolesa, kar telesu skupaj omogoča gibanje /…/?« (Hobbes 1651/2006: 9)

Tudi Leibniz je menil, da je mišljenje računanje, in predlagal izoblikovanje univerzalnega pojmovnega jezika (lat. *characteristica universalis*), neke vrste abecede človeškega mišljenja, s katero bi lahko izrazili matematične, znanstvene in metafizične resnice. »/…/ nihče še ne poskušal ustvariti jezika /…/, v katerem znaki in simboli služijo [misli] na isti način kot matematični znaki služijo številom ali

algebraični znaki količinam.« (Leibniz 1679/1969: 222) Njegov univerzalni jezik je bil v resnici del večjega projekta, in sicer izoblikovanja univerzalne znanosti (lat. *scientia universalis*), tj. transformacija celotnega človeškega znanja v eno sistematično celoto, v kateri bi mišljenje potekalo kot računanje. Da bi bil ta program uspešen, je Leibniz predlagal še univerzalni mehanizem sklepanja (lat. *calculus ratiocinator*), zbirko načinov manipulacije znanja, zapisanega v računalniški obliki, z namenom odkrivanja logičnih relacij med idejami in njihovimi posledicami.

> Naj bodo pojmi, iz katerih je sestavljeno vse drugo, kar obstaja, prvič določeni z znaki, ki bodo neke vrste abeceda. Bilo bi priročno, če bi bili karseda naravni, npr. pike za števila, črte za relacije med entitetami /…/ Če bodo pravilno in domiselno izbrani, bo univerzalni jezik enostaven in običajen ter za njegovo razumevanje ne bomo rabili slovarja. Poleg tega si bomo na tak način zagotovili znanje o vsem, kar obstaja. (Leibniz 1666/1966: 10–11)

Descartes na drugi strani pa primerke misli nikakor ne vidi kot primerke možganov. To mu onemogoča njegov pogled na t. i. *problem duha in telesa*,[3] tj. vprašanje, kakšen je odnos med mentalnim in fizičnim oziroma med našimi psihičnimi in možganskimi stanji. Njegov odgovor je splošno znan *dualizem substanc*, stališče, ki predpostavlja obstoj dveh različnih in ločenih substanc, misleče/mentalne – *res cogitans* in razsežne/fizične – *res extensa*. V *Razpravi o metodi* leta 1637 zapiše takole: »Potemtakem je ta jaz, namreč duša, po kateri sem, kar sem, popolnoma različna od telesa in jo je celo lažje spoznati kakor pa telo; in četudi telesa ne bi bilo, ne bi prenehala biti vse tisto, kar je.« (Descartes 1637/2007: 51, 53)[4] Njegovo prvo načelo filozofije je sicer *mislim, torej sem* (lat. *cogito ergo sum*), saj odkrije, da on sam, ki misli, da je vse zmotno, nujno nekaj je. Vprašanje je, kaj ta *jaz*, v katerega ne more dvomiti (če v vse drugo že lahko), je. Na to odgovori z miselnim eksperimentom, v katerem si predstavlja sebe brez telesa, brez kakršnihkoli organov (svoje telo povsem odmisli), in ugotovi, da kljub temu še vedno nekaj je, je tisto, kar je odmislilo telo, tj. misleča stvar. »Iz tega sem spoznal, da sem substanca, katere celotno bistvo ali narava

[3] Zanj glej npr. Kim (2001).

[4] V *Meditacijah* to izrazi takole: »In čeprav morda (ali bolje: zagotovo …) imam telo, ki je zelo tesno združeno z mano, je vendar – ker imam na eni strani jasno in razločno idejo samega sebe, kolikor sem samo misleča, nerazsežna stvar, in na drugi strani jasno idejo telesa, kolikor je telo zgolj razsežna, nemisleča stvar – je torej vendar gotovo, da sem v resnici različen od svojega telesa in da bi lahko bival brez njega.« (Descartes 1641/1988: 107) (Zadnji del citata, tj. od vezaja do pike, je zaradi napake prevajalca spremenjen; za primerjavo glej Descartes 1641/1985a: VI meditacija.)

je zgolj mišljenje in ki za bivanje ne potrebuje nobenega kraja in ni odvisna od nobene materialne stvari.« (Descartes 1637/2007: 51)

Ker so lastnosti, ki jih pripisuje duhu, npr. misli, občutki, čustva, lastnosti, ki jih pripisuje telesu, pa npr. razsežnost, velikost, oblika, zaradi česar je duh stvar, ki misli in čuti, telo pa stvar z maso in lokacijo (psihični pojavi, kot so čustva, so po svoji naravi drugačni in ločeni od fizičnih pojavov, kot je proženje nevronov), Descartes mišljenja nikakor ne more opisati v pojmih (mehanskega) računanja, kot je to storil Hobbes. Res je, da med duhom in telesom poteka interakcija, ni pa jasno, kako je to mogoče, glede na to, da mentalna substanca nima lokacije v prostoru, fizična pa jo ima. Ta temeljni problem dualizma substanc, tj. kako lahko človeški duh vpliva na aktivnosti telesa (npr. krčenje in raztezanje mišic), je Descartesov sodobnik Gassendi izpostavil z besedami »Pojasniti moraš, kako je ta interakcija, če si breztelesen, se ne raztezaš v prostoru in si nedeljiv, možna. /… / Kako lahko, če nimaš delov, vplivaš na dele? /… / In če si nekaj ločenega, kako lahko skupaj s snovjo tvoriš celoto?« (Descartes 1985b: 238)

Kakorkoli, Descartesovi opisi delovanja telesa (za razliko od duha) pa so bili povsem mehanicistični. Ne samo da je v *Razpravi o metodi* za tisti čas podal podroben prikaz anatomije našega telesa, razmerja med organi, njihovih funkcij in delovanja, dotaknil se je tudi vprašanja podobnosti med človekom in strojem, ki ga prav tako imenuje avtomat.

Dopušča možnost, da bi obstajali stroji, ki bi bili sposobni posnemati naša dejanja, vendar se kljub temu nikoli ne bi moglo zgoditi, da bi jih zamenjali za ljudi. »Kot prvo ne bi nikoli znali uporabljati besed in drugih znakov ter jih sestavljati, kot to počnemo mi, da drugim razodenemo svoje misli. Lahko si sicer zamislimo, da bi bil kakšen stroj narejen tako, da bi izgovarjal besede, in celo, da bi izrekel kakšno besedo o telesnih dejanjih, ki bi povzročila spremembe v njegovih telesnih organih /…/ ne moremo pa si zamisliti, da bi te besede združeval v različne tvorbe in z njimi smiselno odgovarjal na vse, kar bi kdo rekel vpričo njega, kot to zmorejo celo najbolj omejeni ljudje.« (Descartes 1637/2007: 83) Razlika, na katero opozarja Descartes, je v tem, da stroji ne delujejo po spoznanju duha, ampak zgolj naravnanosti (sprogramiranosti) njihovih organov. Ta jim sicer omogoča mehansko manipulacijo s simboli, če uporabimo sodobnejšo terminologijo, ki pa ne vključuje racionalne spoznave.

> Razum je namreč univerzalno orodje, ki ga lahko uporabljamo v vseh okoliščinah, nasprotno pa morajo biti ti organi za vsako posebno dejanje posebej naravnani. Zato je praktično nemogoče, da bi bilo v kakšnem stanju toliko različnih organov, da bi mu to omogočalo delovati v vseh življenjskih slučajih, kakor to nam omogoča razum. (Descartes 1637/2007: 83, 85)

Zaključimo lahko, da Descartes nasprotuje strojnemu mišljenju, saj zanj imeti duha pomeni biti sposoben delovati na različnih med seboj tudi precej različnih področij, kar pa nujno vključuje uporabo razuma, ki pa ga avtomati, ker so specializirani zgolj za opravljanje nalog z nekega ozkega specifičnega področja, ne premorejo.[5]

Pomemben mejnik v zgodovini UI predstavlja »izdelava«, bolje rečeno razvoj, prvega računalnika, ki se je pojavil okrog leta 1833, njegov avtor pa je bil (zgodovinsko gledano prvi računalničar) Charles Babbage. Imenoval se je *Analitični stroj*,[6] [7] njegov ustroj pa je vključeval dve izjemni ideji, ki skupaj predstavljata temelj računalništva: a. operacije je bilo mogoče v celoti programirati in b. programi so lahko vsebovali pogojne izjave tipa *če – potem – drugače* (*če* je odgovor »Y«, potem zapiši »pravilno«, *drugače* zapiši »narobe«).[8] (Haugeland 1986: 126) Sestavljen je bil iz treh delov: *mlina* (aritmetične enote), *shrambe* (spominske enote) in *kontrolne enote*. Mlin lahko izvaja štiri osnovne računske operacije, kontrolna enota pa je »sodnik«, ki ne manipulira s števili, ampak glede na navodila zgolj ukazuje, s katerimi števili se sešteva, odšteva, množi ali deli ter kam se zapisujejo odgovori. Znotraj svojih mej lahko Analitični stroj opravi katerokoli nalogo, ki mu jo naložimo, tj. ko enkrat na ustrezen način določimo pravila igre, kontrolna enota poskrbi za to, da se jih igralci (aritmetična in spominska enota) držijo.

[5] Iz *Razprave o metodi* ni jasno, ali stroji razuma nimajo zato, ker je ta del mentalne substance, ki je ni v prostoru, stroji pa so telesni, ali pa zato, ker pri izvrševanju nalog niso sposobni pokrivati več različnih področij. Glede na to, da je Descartes trdil, da med mislečo in razsežno substanco obstaja vzajemno delovanje in to s pomočjo »češerike«, sicer neuspešno, tudi dokazoval, se zdi bolj verjetna druga možnost. Če to drži, potem nasprotuje le pripisovanju inteligence aktualnim modelovm UI, ki so specializirani samo za opravljanje določenih nalog z enega ozkega področja, ne pa tudi obstoju splošne UI.

[6] *Analytical Engine*.

[7] Babbage je na papirju razvil dva računalnika, prvi, ki se je v angl. imenoval *Difference Machine*, je bil sicer izviren, za nadaljnji razvoj pa ne tako pomemben izdelek.

[8] Angl. *conditional branches*.

To početje, ki iz specifičnega avtomatskega sistema zgolj s primernim opisovanjem in spreminjanjem navodil naredi avtomatski splošni/formalni sistem (računalnik), pa že imenujemo programiranje.[9] Tudi navadni kalkulator lahko sicer prav tako izvaja iste štiri operacije, a to razliko, da moramo posamezne ukaze vnašati ročno. Če želimo izračunati 3(x +y) – 5 za dana x in y, moramo najprej vstaviti dana x in y, potem njuno vsoto pomnožiti s 3 in na koncu od zmnožka odšteti 5. Analitični stroj pa je bil na drugi strani že takrat (kot danes računalniki) sposoben na osnovi specifikacij, ki, kot smo že omenili, vključujejo pogojne izjave tipa *če – potem – drugače*, za izračun poljubnega zaporedja osnovnih računskih operacij za poljubne spremenljivke, izražene v njemu razumljivem »jeziku«, avtomatsko izračunati katerokoli dano formulo. »Tako je bil Babbage prvi, ki je »naredil« sprogramiran računalnik, pri čemer je treba »naredil« vzeti z rezervo: ker je bil njegov stroj prevelik in prezapleten, zaradi česar ga ni skoraj nihče razumel, kaj šele bil sposoben zgraditi, ni v praksi nikoli zaživel.« (Bregant 2010: 62)

Nič manj pomembna ni iznajdba še enega takega stroja, ki ga je doletela ista usoda, tj. nikoli ni bil realiziran v praksi, čeprav iz drugega razloga – že v osnovi je bil mišljen le kot enostavni abstraktni sistem za dokazovanje teoretičnih predpostavk – je leta 1936 razvil Alan Turing.[10] Potem, ko je računanje definiral kot formalno manipulacijo z neinterpretiranimi simboli, ki se izvaja z uporabo formalnih pravil (Turing 1936; Markič 2019), je opisal še napravo, ki je tako računanje sposobna izvesti. Znana je pod imenom *Turingov stroj* in sestavljena iz *glave* in *traku*. Trak, ki je shramba podatkov, je razdeljen na neskončno število kvadratov, izmed katerih so eni vedno zasedeni s simbolom, tj. figurami, s katerimi se manipulira, iz predpisane abecede, drugi pa prazni. Glava, ki je izvrševalec operacij, se pomika preko kvadratov in skenira njihove simbole. Naenkrat obdeluje samo en kvadrat in to je edini kvadrat, s katerega takrat bere ali na katerega zapisuje, odvisno pač od tega, v katerem notranjem stanju je, tj. kaj je predpisano delo, ki ga mora na njem opraviti. Ko konča, se pomakne naprej ali nazaj. Na katerem kvadratu se ustavi in kaj z njim stori, je vnaprej določeno s pravili, po katerih deluje. Ta vsebujejo opise tega, (i) kateri simbol je treba zapisati na dani kvadrat (ali s katerim novim je treba nadomestiti starega), (ii) kateri kvadrat je treba skenirati naslednji in (iii) kaj je treba z naslednjim kvadratom storiti. Tudi Turingov stroj je tako preprost avtomatski formalni sistem, podoben Analitičnemu stroju, s katerim lahko izvršimo vsako nalogo, za katero smo

[9] Turing pozneje pravi, da lahko takšne računalnike, ki imajo »posebno lastnost /…/, da lahko posnemajo kateri koli diskretni stroj, opišemo [kot] /…/ *univerzalne* stroje.« (Turing 1950/1990: 64)
[10] Zanj glej Turing (1936), Haugeland (1986).

jasno specificirali korake, ki so potrebni za njeno izpolnitev. Čas za prihod računalnikov, ki ne bodo le imaginarne tvorbe na listu papirja, je napočil.

3 1. val razvoja UI: simbolno modeliranje

Računalniki

Razvoj UI v neki točki sovpada z razvojem računalnikov. Beseda 'računalnik' naj bi se prvič pojavila leta 1613, ko jo je Richard Brathwaite uporabil za opis osebe, ki je računala, kasneje pa tudi za naprave, ki so bile sposobne izvajati računske operacije. Sodoben pojem elektronske računske naprave pa se v bistvu ni pojavil vse do leta 1945, ko ga je von Neumann uporabil v svojem znanem poročilu o EDVAC-u,[11] ki je bilo prvi javno objavljeni opis logične zgradbe računalnika, v katerem so bili podatki o programu in podatki o navodilih shranjeni v spominski enoti, kar je dobilo ime von Neumannov model (ali arhitektura). (Berkeley 2018) Prvi računalnik, ki ni bil zgolj neuresničen teoretični eksperiment naj bi leta 1941 izdelal Konrad Zuse. Sposoben je bil opraviti katerokoli računsko operacijo, na kasnejši razvoj računalnikov pa ni imel nobenega vpliva, saj zanj takrat, ko je nastal, zaradi začetka 2. sv. vojne ni vedel skoraj nihče. Leta 1943 so v Bletchley Parku, kjer je bil sedež zavezniške skupine za odkrivanje kod, v katerih so bila zapisana nemška sporočila, tudi Britanci razvili svoj prvi računalnik, ki se je imenoval *Kolos* (velikan). Razvozlaval je na prvi pogled nesmiselno nemško komunikacijo o npr. premikih njihovih čet in pomembno prispeval k zmagi zaveznikov nad Nemci v 2. sv. vojni. Leta 1948 je v Manchestru nastal prvi povsem elektronski računalnik z imenom *Mark I*, ki je bil tudi prvi, narejen za »masovno« proizvodnjo in komercialno prodajo. Izdelali so jih 9 in jih 9 tudi prodali. Kasneje so prevladujoč položaj pri proizvodnji računalnikov prevzele ZDA, ki so svoj *ENIAC*[12] sicer predstavile že leta 1945, vendar je bil ta v primerjavi z *Markom I* z vidika tega, kako ga je bilo treba programirati za vsako novo nalogo, nočna mora vsakega operaterja. Šlo je za velikansko operacijo ročnega pretikanja kablov iz enih vtičnic v druge, ker je programerjem vzelo dva dni, da so ga pripravili za novo delo. Na drugi strani pa je bil *Mark I* sposoben z vstavljenega preluknjanega papirnega traku, na katerem so bila zapisana navodila, te preprosto skopirati v svoj spomin in takoj začeti z naslednjim opravilom. (Copeland 1993; Bregant 2010)

[11] Electronic Discrete Variable Automatic Computer.
[12] Electronic Numerical Integrator and Computer.

Kaj je torej računalnik? Računalnik je *avtomatski* formalni sistem,[13] ki je sposoben *interpretirati* simbole.[14] Kaj to pomeni? To pomeni, da lahko prepozna »figure«/simbole, s katerimi se igra in na njih skladno z navodili izvaja manipulacije (to vključuje tudi »sodnika«/kontrolorja, ki skrbi za to, da igra poteka po pravilih – določa, kdo je na potezi, s katero figuro naj igra in razglasi rezultat) ter da lahko ugotovi, kaj figure/simboli pomenijo, tj. kaj izražajo ali predstavljajo v vsakdanjem jeziku. Naloga interpretacije, kjer računalnik najprej ugotovi, kaj pomeni vsak enostaven simbol (npr. besede), potem pa, kaj pomeni vsak sestavljen (npr. stavki), je, da simbole iz enega sistema, ki ni razumljiv, »prevede« v drugega, ki je. Velja pravilo, da če računalnik poskrbi za sintakso, tj. da se drži specificiranih formalnih pravil, s tem hkrati poskrbi tudi za semantiko, tj. iznosi/outputi dobijo v standardnem jeziku smisel. »Če stroj sledi znotraj sistema definiranim sintaktičnim pravilom pri tvorjenju novih formul, potem bodo interpretirane formule ohranjale svojo semantično vrednost.« (Markič 1997: 43–44)

Poenostavljeno lahko rečemo, da je računalnik *stroj, ki manipulira s simboli*[15] in je s tega vidika podoben človeku. Tudi ljudje za prikaz stvarnosti uporabljamo besede in stavke, ki so simbolni opis predmetov in dogodkov v našem svetu. To preprosto imenujemo *simbolna hipoteza*, izhaja pa iz tega, da tudi naš duh ni nič več kot univerzalni sistem simbolov, zaradi česar je tudi človeška spoznava zgolj operiranje z njimi. V splošnem pravi, da je mišljenje računanje, ki vključuje manipulacijo s simboli, ta pa je lahko realizirana v sistemih, ki jo izvajajo preko preklapljanja med 0 in 1 kot elementoma binarnega sistema. Pri človeku sta to v svojem prelomnem članku z naslovom »A Logical Calculus of the Ideas Immanent in Nervous Activity«, ki pomeni začetek pristopa k simbolnemu modeliranju mišljenja,[16] ki je zaznamoval prvi val razvoja UI, leta 1943 pokazala nevrofiziolog McCulloch in matematik Pitts. Dokazovala sta, da je edino, kar je pri nevronih pomembno to, da so sproženi ali nesproženi. Prvo lahko označimo z npr. 1 ali »da«, drugo pa z npr. 0 ali »ne«. To, ali so sproženi ali nesproženi, pa je odvisno od tega, ali je dosežen oziroma presežen njihov aktivacijski prag. (Bregant 2016) Naši možgani tako delujejo kot neke vrste preprosti manipulatorji s simboli, saj ni nevron nič drugega kot naprava, ki lahko

[13] Formalni sistem je funkcionalna celota med seboj načrtno povezanih odvisnih elementov, ki deluje na predpisan in ustaljen način.
[14] Obstajajo tudi ročni formalni sistemi, šah je eden izmed njih, kjer nekdo od zunaj – igralec – premika figure in ustvarja nove položaje.
[15] Te simbole imenujemo *biti* – 0 in 1 – in označujejo besede ter števila, zapisani pa so v vrstah, tj. registrih.
[16] Zanimivo je, da pomeni članek hkrati tudi začetek modeliranja nevronskih mrež, ki se običajno bolj povezuje z drugim valom razvoja UI. (Russell in Norvig 2010; Markič 2019)

fizično realizira eno izmed obeh stanj. Ker pa je »zakon 'vse ali nič', kateremu je podvržena živčna aktivnost, zadosten za to, da je lahko aktivnost kateregakoli nevrona predstavljena kot propozicija« (McCulloch in Pitts 1943/1990: 23–24), si lahko nevrone zamislimo kot propozicije, ki so resnične (1) ali neresnične (0), obtežene povezave med njimi pa kot logične veznike in tako v pojmih logike izračunamo vse, kar lahko jezikovno opišemo.

Naenkrat se je zdelo, da je Hobbsova ideja o tem, da je mišljenje računanje, dobila sprejemljivo fizično razlago in odkrila način za izdelavo umetnih sistemov, ki so sposobni posnemati katerikoli vidik inteligence. In ko so se v Hanovru v New Hampshiru v ZDA leta 1956 na tamkajšnjem Dartmouthskem kolidžu med drugim zbrali John MacCarthy,[17] Marvin Minsky, Alan Newell, Herbert Simon, Claud Shannon in Arthur Samuel, ki sta jih zanimala razvoj in izdelava inteligentnega stroja, se je rodila nova disciplina, ki si je nadela ime umetna inteligenca.[18] Srečanje je temeljilo na domnevi, da lahko katerakoli značilnost inteligence v jeziku logike opišemo tako natančno, da jo lahko stroji posnemajo oziroma podvojijo, včasih se ta namera izraža s preprostim vprašanjem »Ali lahko stroj misli?«,[19] zaradi česar jim inteligenco tudi moramo pripisati.

Programi

V tistem času se je utrdilo prepričanje, da je kriterij za pripisovanje inteligence zmožnost logičnega mišljenja, sposobnost, ki bi jo moral imeti tudi stroj, da bi ga lahko imeli za inteligentnega. Najresnejši kandidat za to je bil po prevladujočem mnenju takrat program z imenom *Logični teoretik*, ki so ga leta 1956 razvili Newell, Shaw in Simon, tekel pa je na računalniku z imenom *Johnniac*, ki ga je izdelal von Neumann. Da bi to dosegel, bi moral biti program sposoben dokazati logične teoreme tipa $p \rightarrow (p \vee q)$. (Newell, Shaw, Simon, 1963) Pri tem lahko to storimo s

[17] John McCarthy je bil organizator konference, ki jo je naslovil *The Darthmouth Summer Research Project on Artificial Intelligence*; zadnji besedi sta ostali in dali ime na novo rojenemu področju.

[18] Domnevno se je ideja o umetni inteligenci prvič pojavila v McCulloch in Pitts (1943/1990).

[19] Turing pa ni znan samo kot izumitelj enega izmed abstraktnih računalnikov, ampak tudi kot avtor posebnega preizkusa, ki naj bi pokazal, ali stroj lahko misli. Imenuje se *Turingov test* zasnovan pa je na t. i. igri oponašanja, v kateri v *standardni verziji* preizkusa nastopajo trije akterji, dva človeka in računalnik. Eden izmed ljudi je spraševalec, ki mora ugotoviti, kateri izmed preostalih dveh igralcev, ki odgovarjata na njegova vprašanja, je računalnik. Vsi trije so v ločenih sobah, komunikacija pa poteka preko ekrana in tipkovnice. Računalnik lahko stori karkoli, da bi s svojimi odgovori spraševalca prevaral in izsilil napačno identifikacijo, človek pa mora na vprašanja odgovarjati po resnici. Test se ponovi večkrat, pri čemer se človeški igralci vedno menjajo. Stroj ga opravi, če na koncu spraševalec v več kot 50 % primerov računalnik zamenja za človeka. V tem primeru moramo tudi stroju, če želimo biti dosledni, pripisati mišljenje. (Turing 1950/1990; Bregant, 2016)

pogojnim dokazom, pri katerem je pogojnik dokazan takrat, ko nam uspe izpeljati njegov konsekvent. V tem postopku vedno najprej predpostavimo antecedent pogojnika, v našem slučaju je to *p*, potem pa nadaljujemo z najbolj primerno logično operacijo, da bi čimprej dobili sklep. V našem primeru je to adicija, s katero takoj dobimo *p V q*, s čimer je gornji logični izrek že dokazan. To je *Logičnemu teoretiku* s tem, ko je dokazal prvih 38 teoremov drugega poglavja iz Whiteheadove in Russellove *Principia Mathematica*,[20] ki velja za temeljno delo s področja logike in matematike, tudi uspelo. To je bilo prvič, da stroj ni samo izvajal računskih operacij, ampak dokazoval, zaradi česar gre verjetno za prvi praktični izkaz strojne inteligence nasploh. (Copeland 1993; Bregant 2010)

Po tem preboju na področju razvoja programov, ki so zmožni interpretirati simbole, tj. jih tolmačiti tako, da imajo v vsakdanjem jeziku smisel in na ta način posnemati človeško obnašanje do te mere, da jim lahko pripišemo inteligenco, je v 60. in 70. letih 20. stol. luč sveta ogledalo kar nekaj programov, ki so v večji ali manjši meri izpolnili takratne zahteve za njen pripis. Prvi, ki ga je vredno omeniti, je tudi najbolj znan, in sicer je to *Eliza*, ki je nastal v sredini 60 let 20. stol. na MIT-ju, njegov avtor pa je bil Joseph Weizenbaum. Njena naloga so bili psihoterapevtski pogovori z ljudmi, ki so bili v takšni ali drugačni duševni stiski, ki so potekali preko tipkovnice in ekrana. Z vidika kognitivnih sposobnosti je znala *Eliza* samo eno stvar: sogovornikove odgovore je zgolj ponovila in čakala na odziv ali pa jih spremenila v vprašanja (kar je, mimogrede, klasičen način psihoterapevtskega pogovora). Imela ni nobenih lastnosti, ki se običajno povezujejo s pripisovanjem UI, npr. ni poznala svojega okolja, ni bila sposobna razmišljati in načrtovati dejanj, ni razumela motivov zanje in ni se mogla ničesar naučiti, zaradi česar je ne moremo imeti za inteligenten program v pravem pomenu besede.

Kljub temu je bila tako prepričljiva, da je brez težav prevarala »bolnike«, da je človek. Weizenbaum je bil zgrožen, nikoli si ni predstavljal, da nas lahko tako razmeroma enostaven računalnik tako preslepi. Ljudje so se na *Elizo* celo čustveno navezali, tudi njegova tajnica, ki je bila seznanjena s celotnim projektom, je zahtevala, naj vsi zapustijo sobo, da se bo lahko z njo nemoteno pogovarjala. Zaupali so ji svoje najintimnejše skrivnosti in se niso dali prepričati, da gre zgolj za stroj. »Nisem si mislil, da lahko relativno kratek čas, ki ga normalni ljudje preživijo z relativno preprostim računalnikom, v njih povzroči tako močne fantazije.« (Weizenbaum

[20] Whitehead, Russell (1910–1913/1997).

1976: 7) Še več, nekateri psihiatri so jo bili pripravljeni testirati na svojih pacientih, v *Journal of Nervous and Mental Disease* pa so bili prepričani, da bo program, ko bo zrel za klinično uporabo, predstavljal terapevtsko orodje, s katerim bi lahko v eni uri obravnavali več sto pacientov in tako nadomestili psihiatre v ustanovah za duševno bolne. Z žalostjo je ugotovil, da je naša družba brez predsodkov pripravljena zaupati skrb za blagostanje ljudi računalniku, zaradi česar je postal nasprotnik UI. Menil je, da ni »[njen] cilj ni nič manj kot to, da izdela stroj po vzoru človeka, robota, ki bo odraščal, se učil jezika na isti način kot otrok, spoznaval svet s pomočjo čutil in na koncu razmišljal o vsem znanju, ki ga človeštvo premore.« (Weizenbaum 1976: 202–203) Zato zanj vprašanje ni bilo več, ali lahko naredimo stroj, podoben človeku, ampak, ali to smemo, saj je bil prepričan, da sistemi UI ne bodo nikoli povsem sposobni razumeti in pravilno ovrednotiti položaja, v katerem se lahko znajde človek, zaradi česar niso primerni za opravljanje našega dela. (Copeland 1993; Bregant, 2019)

Naprednejši program, ki naj bi imel vse tiste spoznavne zmožnosti, ki so *Elizi* manjkale, je v 70. letih 20. stol. na MIT-ju razvil Terry Winograd, imenoval pa se je *Shrdlu*. S pomočjo »roke« in ukazov, kaj naj naredi, je na mizi premikal predmete različnih oblik, barv in velikosti, pri čemer je za opravljanje zahtevane naloge razvil in izpeljal lasten načrt. Še več, v ozkem specifičnem okolju, ki je bilo natančno določeno, je znal logično sklepati. Npr. če določimo, da so predmeti, ki so naši, piramide in tisti, ki niso beli, potem pa ga vprašamo, ali je v belem kvadru, ki vsebuje belo piramido in modro kocko, kakšen predmet, ki je naš, *Shrdlu* pravilno odgovori, da je to modra kocka. Ker mu tako lahko pripišemo opravljanje kognitivnih operacij (sicer zgolj v nespreminjajočem se okolju), je bil prvi program, ki je bil sposoben izpolniti omejen pogoj za pripis inteligence: »razumel« je navodila, s pomočjo sklepanja je našel odgovore na kompleksna vprašanja, »razumel« pa je tudi, vsaj deloma, motive zanje. (Copeland 1993; Bregant 2019)

Na koncu naj omenimo še mogoče najbolj napreden program iz tistega časa, ki je nastal v sredini 70. let 20. stol. na MIT-ju, z imenom *Hacker*. Njegov avtor je bil Gerald Sussman, cilj pa razbiti mit o tem, da se računalniki nikoli ne bodo mogli sami programirati. *Hacker* ima to, do sedaj še nevideno zmožnost, in sicer sposoben je razviti programe za računalnik, na katerem teče tudi sam. Programi, ki jih piše, v dobro določenem in ozko omejenem okolju nadzirajo »roko«, ki na mizi premika s črkami označene kocke. Recimo (Copeland 1993), da mora *Hacker* razviti program, tj. napisati postopek izvedbe, ne pa to zgolj izvesti, kar je delal npr. *Shrdlu*, ki bo

omogočal, da kocko A, ki leži pod kocko B, postavimo nad kocko C. Najprej pogleda v svojo knjižnico, da bi našel karkoli relevantnega za to nalogo, ali kakšen program, ki ga je dobil od svojega avtorja, ali kakšen program, ki ga je že sam napisal. Edino, kar najde, je npr. vzorec *postaviti na*, kar »roki« omogoča, da en predmet postavi ali na drugega ali na mizo. Dalje vidi, da bi lahko ta vzorec uporabil za to, da bi položil kocko A na kocko C, če kocka A ne bi imela ničesar nad sabo. Potem v svoji knjižnici išče dalje in najde npr. vzorec *spustiti na*, kar »roki« omogoča, da en predmet spusti ali na drugega ali na mizo. Sedaj lahko zahtevano nalogo opravi: najprej kocko B spusti na mizo, potem pa kocko A postavi na kocko C.

Bistvenega pomena za to, da je pri programiranju uspešen, je njegova knjižnica tehnik programiranja, ki je v resnici shramba dejstev in receptov, pa tudi trikov in ukan, znanih hekerjem, o tem, kako napišemo program. Poleg tega se je sposoben učiti iz izkušenj, vsak neuspešen poskus programiranja je vir informacij o tem, česa se ne dela. Na ta način se s prakso izboljšuje, vse, kar se novega nauči, pa shrani v svojo knjižnico tehnik programiranja, ki jo je dobil od svojega avtorja, in jo tako povečuje. Res je, da je bil kontekst, v katerem je *Hacker* deloval, izmišljen in zanesljiv, ter da je bil sposoben napisati zgolj najbolj enostavne programe, res pa je tudi, da je bil s tem razbit mit o računalnikih, ki sami tega niso zmožni narediti. In naj je bilo do zapletenih in naprednih programov še tako daleč, ideja o tem, da bodo stroji programiranje nekoč vzeli v svoje roke, ni bila več neutemeljena. (Copeland 1993; Bregant 2019)

Ali smo računalniki?

Eno izmed bolj provokativnih filozofskih vprašanj, ki izhaja iz opisa računalnika kot fizičnega sistema, ki s pomočjo programa, implementiranega v strojnem jeziku, realizira operacije s simboli, je, ali nismo tudi ljudje zgolj (neke vrste) računalniki. O tem, da smo ljudje stroji, je pisal že La Mettrie v svojem *Strojnem človeku* leta 1748: »Človek je stroj, zgrajen na takšen način, da si je to nemogoče predstavljati in ga zato tudi nemogoče definirati.«[21] (La Metrrie 1748/1996: 5) Danes imamo na voljo boljšo predstavo o tem, kakšen stroj naj bi človek bil, vse skupaj pa se je začelo v sredini 19. stol., ko je bilo odkrito, da je delovanje živčnega sistema v resnici sprejemanje in prevajanje električnih impulzov, kasneje pa do podrobnosti raziskane kemične

[21] Pri tem ne pozabi omeniti, da je o tem govoril že Descartes v *Razpravi o metodi* leta 1637: »Nihče ne zanika, da je ta slavni filozof naredil mnogo napak, je pa razumel živalsko naravo in bil prvi, ki je dovršeno pokazal, da so živali stroji. (La Mettrie 1748/1996: 35)

značilnosti nevronov in njihov električni potencial. Že La Mettrie je predpostavljal, da je »posebna lastnost našega stroja, da vsako njegovo vlakno, tudi najmanjše, oscilira. To naravno nihanje je kot ura, ki včasih zamre in se mora potem obnoviti. Če postane šibko, se mora okrepiti, če pa premočno, oslabiti.« (La Mettrie 1748/1996: 31) Vendar je minilo še kar nekaj časa, da smo ugotovili, kaj je namen vzajemnega proženja nevronov, kar nam je pomagalo razumeti delovanje naših možganov. To se je zgodilo z omenjenim člankom McCullocha in Pittsa približno sto let pozneje, ki je pokazal »kako lahko operacije nevronov in njihove povezave z ostalimi nevroni modeliramo s pojmi logike.« (Markič 2010: 30). To pomeni, da si lahko nevrone zamislimo kot propozicije, ki so resnične (1) ali neresnične (0), obtežene povezave med njimi pa kot logične veznike in tako v pojmih logike izračunamo vse, kar lahko v jeziku opišemo.

Tudi računalniku t. i. »logična vrata« omogočajo, da podobno kot nevron po istem principu preklaplja med 1 in 0. Recimo, če v računalnik preko tipkovnice vnesem 'mačka' (vhod ali vnos), računalnik preklaplja med 0 in 1, odvisno od kode, ki jo vsaka črka ima, dokler ne pride do binarnega zapisa, npr. 1000011 1000001 1000111 1001111 1000001, ki pomeni izpis 'mačka' na mojem ekranu (izhod ali iznos). Ker tako tudi računalnik ni nič drugega kot naprava, ki lahko fizično realizira eno od obeh stanj – manipulator s simboli – do njegovega enačenja z našimi možgani na osnovi opisanega delovanja nevrona ni bilo več daleč. Kljub temu, da so pri računalnikih manipulacije s simboli realizirane z bistveno drugačnimi fizični operacijami kot pri nas, to ni pomembno, saj enačenje računalnika in možganov temelji na ideji o *večvrstni realizaciji simbolov*. Vsebuje domnevo, da so lahko simboli realizirani na različne načine, in sicer z barvo, ogljem, svinčnikom, elektromagnetnimi stikali (releji), elektronkami, silikonskimi polprevodniki (tranzistorji) itd. (Bregant 2016), in se skriva v naslednjem Turingovem citatu: »Pogosto se misli, da je pomembno, da so sodobni digitalni računalniki električni in da je tudi živčni sistem električen. /…/ ker so vsi digitalni računalniki v nekem smislu enakovredni, vidimo, da uporaba elektrike z vidika teorije ni pomembna.« [22] (Turing 1950: 439)

[22] Za izdelavo *Analitičnega stroja* je Babbage predvidel ključno mehanske dela, tj. kolesa in zobnike.

Eden izmed zagovornikov enačenja računalnikov in možganov je Pylyshyn:

> Semantična vsebina [prepričanj in namer] je kodirana s strani možganskih lastnosti na isti splošni način, na katerega so semantične vsebine računalniških predstav kodirane s strani fizično uprimerjanih simbolnih struktur. (Pylyshyn 1984: 258)

Naše mentalne operacije so v bistvu manipulacije s simbolnimi reprezentacijami, ki so zgolj podobne običajnim stavkom kateregakoli govorjenega jezika: »/…/ [simbolne reprezentacije] so simbolni izrazi v notranjem, fizično uprimerjanem simbolnem sistemu, ki se včasih imenuje 'mentalese' ali 'jezik misli'.«[23] (Pylyshyn 1984: 194) Ali je manipulacija s simboli kot podobnost med računalniki in možgani dovolj za enačenje ljudi s stroji oziroma ali ne obstaja dovolj razlik, ki to postavljajo pod vprašaj, pa je treba še raziskati.

Vnet nasprotnik enačenja računalnikov z nami pa je Searle, ki je leta 1980 v svojem članku *Duhovi, možgani in programi* predstavil znani miselni eksperiment z imenom »kitajska soba«, s katerim je hotel za vse večne čase že v kali zatreti tudi načelni poskus pripisovanja inteligence kakršnemukoli stroju. Predstavljajte si Janeza, ki ne obvlada nobenega drugega jezika razen slovenščine, zaprtega v sobi z odprtino, skozi katero dobi tri zvezke besedila v kitajščini, *skripte* (scenarije za različne situacije), *zgodbo* in *vprašanja*, poleg tega pa še v slovenščini formalna pravila za povezovanje kitajskih pismenk, ki mu omogočajo, da jih lahko spozna izključno po njihovih oblikah. Zunaj sobe se od njega pričakuje, da odgovori na vprašanja, odgovore pa posreduje skozi za to pripravljeno odprtino. Zamislite si, da se Searle v rokovanju s simboli sčasoma tako izuri, da se njegovi odgovori v kitajščini ne razlikujejo od odgovorov rojenih Kitajcev. Vprašanje je sedaj, ali Janez, ki deluje kot računalniški program, tj. izvaja operacije na formalno opredeljenih elementih oziroma proizvaja kitajske odgovore tako, da uporablja neraztolmačene formalne simbole, obvlada kitajsko? Searlov odgovor je kategoričen *ne*, saj je ključna razlika med stroji in ljudmi v tem, da ljudje *razumemo* (v tem primeru zgodbo), stroji pa ne.

[23] Za jezik misli glej tudi Fodor (1975), Markič (2010).

Skratka, Searlov miselni eksperiment pokaže, da obvladovanje sintakse, tj. poznavanje niza pravil za simbolno manipulacijo, še ne pomeni obvladovanja semantike, tj. poznavanje tega, kar simboli dejansko pomenijo. To pa je v nasprotju z enačenjem računalnikov in možganov, saj bi morali biti stroji, da bi bili podobni ljudem (in obratno), sposobni razumeti formalne simbole, s katerimi manipulirajo, kot to velja za nas, kar pa očitno ni mogoče. »/…/ v dobesednem pomenu programirani računalnik razume toliko, kot razumeta avto in seštevalni stroj, to pa je natanko nič. Računalnikovo razumevanje ni niti delno niti nepopolno (kot je moje razumevanje nemščine); je povsem nično.« (Searle 1980/1990: 366)

To, da Searle »kitajske sobe« ni nikoli zapisal v obliki argumenta, je še najmanj, da pa na naslednji ugovor (če malo poenostavimo) zgolj zamahne z roko, pa je že zaskrbljujoče. Namreč, kako vemo, da ljudje to, o čemer se z njimi pogovarjamo, sploh razumejo? Zdi se, da razen njihovih smiselnih odgovorov na vprašanja, ki se pojavijo, ni na voljo nobenega drugega dokaza za to. Searle sicer ta ugovor povzame takole: »Kako pa veste, da drugi ljudje zares razumejo kitajsko ali karkoli drugega? Samo po njihovem vedenju. Računalnik lahko ravno tako uspešno opravi vedenjske teste kot oni (v načelu); če torej pripisujemo drugim ljudem spoznavnost, jo morate načeloma pripisati tudi računalnikom.« (Searle 1980/1990: 373) Če pa to drži, bi morali tudi računalnik, ki se je v danem okolju sposoben obnašati na ustrezen način, bodisi verbalno bodisi fizično obravnavati kot bitje ali sistem, ki to, kar je vzrok za njegovo takšno ali drugačno delovanje, razume. V nasprotnem primeru pri pripisovanju duševnosti uporabljamo različna merila, zaradi česar smo lahko upravičeno obtoženi t. i. *šovinizma vrst*, ki (v tem primeru) ljudi obravnava drugače kot stroje, pa čeprav za to ni nobenega razloga.

Kakorkoli, po tem, lahko bi rekli zlatem obdobju UI, ko je, kot smo videli, raziskovalcem s pomočjo simulacije človeškega obnašanja, ki je temeljilo na transparentni simbolni manipulaciji, uspelo razviti modele UI, ki so bili (do neke mere in po nekih standardih) inteligentni, kar je vse navdajalo z optimističnimi pričakovanji in napovedmi, pa je sledilo obdobje zatona. Programi, kot sta bila *Shrdlu* in *Hacker*, ki so temeljili na deduktivni formalni logiki in delovanju v dobro definiranem nespreminjajočem se okolju, niso bili uporabni za reševanje kompleksnih nalog iz vsakdanjega negotovega sveta. To bi moralo temeljiti na induktivnem (statističnem) sklepanju in obdelavi velike količine podatkov, kar pa od sistemov UI zahteva veliko računsko moč, česar pa takrat še ni bilo na voljo.

4 2. val razvoja UI: nesimbolno modeliranje

Posnemanje inteligence

O ponovnem vzponu UI lahko govorimo na začetku 80., o njegovem zagonu pa v 90. letih 20. stol.,[24] ko se je začela povečevati računska moč strojev, količina informacij, ki je bila s prihodom spleta v obtoku in uporaba računalnikov, ki je omogočala relativno enostaven dostop do njih. Gre za drugi val razvoja UI, ki pa po novem temelji na drugačni predstavitvi podatkov, in sicer na indukciji (posplošitev, sklepanje po analogiji, vzročno sklepanje ali sklepanje na najboljšo pojasnitev), ki stavi na predhodne izkušnje in interakcijo z okoljem ter dogodke napoveduje z višjo ali nižjo verjetnostjo, pogosto s pomočjo statistike, pri čemer dopušča (manjšo) možnost, da se motimo. Programi za manipuliranje s simboli, ki so obvladovali prvi val, so v drugem postali algoritmi za strojno učenje, ki pa se je najprej zgledovalo po dejanskih nevronskih mrežah (prvi model sta leta 1943 izdelala že omenjena McCulloch in Pitts, ki tako veljata za začetnika obeh valov), tj. po možganskih procesih, ki so odgovorni za učenje pri človeku in se od tega (s prihodom konekcionizma) oddaljilo šele kasneje.[25]

Kakorkoli, omenjena Dartmouthska konferenca (in danes področje UI nasploh) je slonela na prepričanju udeležencev, da lahko katerokoli značilnost človeške inteligence opišemo tako natančno, da jo lahko stroj posnema. Vzor za to, kako simulacija inteligentnega obnašanja poteka, najdemo sicer že pri Aristotelu, ki je s svojimi *kategorijami*, tj. različnimi vrstami ali načini bivanja, postavil temelje. Njegov cilj je bil sicer zgraditi model sveta, ki bi vključeval njegove različne vidike, njihove značilnosti in odnose med njimi. Vprašanja, na katera je moral odgovoriti, da bi pri tem uspel, so bila, »kaj obstaja«, »kakšne lastnosti imajo stvari« in »v kakšnem odnosu so med seboj«. Za odgovor na njih je razvil sistem 10 kategorij, ki upodabljajo svet: a. substanca/bitnost (npr. človek, miza, računalnik), b. kvantiteta/kolikost (npr. štiri noge), c. kvaliteta/kakšnost (npr. bel, visok, debel) d. relacija/odnos (npr. večji,), e. mesto v prostoru/kje (npr. v šoli, na trgu, v avtu), f. časovnost/kdaj (npr. včeraj), g. imeti (npr. 7 let, brata, 3 avte), h. pozicija/položaj (npr. sedi), i. delovanje/akcija

[24] Iz tega obdobja so verjetno najbolj znani šahovski programi, ki so bili kmalu sposobni premagovati šahovske velemojstre. Šahovski program je prvič premagal svetovnega prvaka v sredini 90. let 20. stol. Takrat je *Globoki modrini* (*Deep Blue*) s 3,5:2,5 v šestih partijah premoč moral priznati Gari Kasparov.

[25] Danes sicer za izdelovanje učinkovitih algoritmov še vedno uporabljamo izraz nevronske mreže, ampak sodobni modeli UI, kot je npr. varovanje pred nezaželeno elektronsko pošto, ne delujejo več po principih delovanja našega živčnega sistema. (Markič 2019)

(npr. gori, seka, tepe), j. trpeti/pasivnost (npr. nadlegovan). (Aristotel 2004; Bregant, 2019) V grafičnem smislu gre v bistvu za mrežo z vozlišči in povezavami, semantična pa se imenuje zato, ker je sestavljena iz pojmov ter relacij med njimi in ker prikazuje v kakšni medsebojni zvezi so stvari in dejanja ter kako vzajemno delujejo. Ker so semantične mreže v resnici na simbolni način predstavljeno znanje, ki sistemu omogoča, da s pomočjo deduktivnega sklepanja, ki poleg aplikacije podatkov vključuje tudi uporabo pravil, izpeljuje nove zaključke o svetu (in se tako *uči*), v tem primeru govorimo o simbolnem modeliranju inteligence oziroma strojnem učenju s pomočjo simbolne manipulacije, tipičnem za prvi val razvoja UI.

V drugem valu razvoja UI pa se izkaže, da to ni edina možnost in da obstaja posnemanje inteligentnega obnašanja, ki ne vključuje na simbolni način predstavljenih informacij, njen vzor pa je delovanje dejanskega nevrona. V tem primeru govorimo o t. i. *nevronskih mrežah*, za njihovo učinkovito obratovanje pa je bistveno naslednje: (a) da je vsak nevron povezan z drugimi, (b) da so te povezave različno obtežene (močne), (c) da obstaja prag, ki določa, ali je nevron aktiviran ali ne in (d) da je tako »izključen« (0) ali »vključen« (1). Skratka, tudi biološki nevroni tvorijo prepleteno mrežo povezav, ki so različno močne, kar je deloma odvisno od premera vlaken, ki so povezana, deloma pa od kemične zgradbe stika (sinaps). Ko vsota vhodnih signalov določenega nevrona doseže ali preseže njegov aktivacijski prag, se nevron sproži in posreduje izhodni signal naprej svojemu najbližjemu sosedu. Podobno so danes zgrajene tudi umetne nevronske mreže, ki predstavljajo orodje nesimbolno predstavljenega znanja, pri čemer so njihovi osnovni gradniki t. i. idealizirani nevroni, tj. preproste, neinteligentne enote, ki so lahko vklopljene ali izklopljene. Takšni umetni nevroni so med seboj povezani, vsak izmed njih pa ima določeno aktivacijsko vrednost, ki jo preko vezi posreduje drugim enotam in s tem pripomore k povečanju (ekscitiranje) ali zmanjšanju (inhibiranje) njihove aktivacijske vrednosti, kar vpliva na to, ali se sprožijo ali ne. (Bregant 2016: 97–98) Ker so nevronske mreže v bistvu na nesimbolni način predstavljeno znanje, ki sistemu dopušča, da iz velike količine podatkov s pomočjo statistične verjetnosti, induktivnega sklepanja in predhodnih izkušenj izlušči ponavljajoče se vzorce (in se tako *uči*), tj. usvoji znanje, ki mu zagotavlja bolj ali manj uspešno napovedovanje dogodkov v našem stalno spreminjajočem se svetu, v tem primeru govorimo o nesimbolnem modeliranju inteligence oziroma strojnem učenju s pomočjo učnih primerkov.

Razprava o tem, ali je simulacija inteligentnega obnašanja dovolj za to, da umetnemu sistemu, ki ga je sposoben realizirati, pripišemo inteligenco, še vedno teče. Zagovorniki menijo, da je pomemben zgolj rezultat in če je rešitev nekega kognitivnega problema pravilna in merljiva ter tako prepričljiva kot pri človeku, ni nobenega razloga za to, da bi stroju inteligenco odrekli. To zelo razširjeno stališče imenujemo *šibka umetna inteligenca*. Nasprotniki pa mislijo, da je ključna izdelava in da lahko govorimo o inteligentnem stroju v pravem pomenu besede samo takrat, ko ima ta enako kot človek tudi fenomenalno zavest,[26] npr. čustva, vizualno izkustvo ali občutke, s čimer bi mu bil z duhovnega vidika enakopraven. To bolj skrajno stališče pa imenujemo *močna umetna inteligenca*.[27]

Strojno učenje

Strojno učenje temelji na prepoznavanju, upoštevanju in sortiranju bistvenih značilnosti predmetov (običajno se pravi, da gre za značilnosti, ki imajo največjo pojasnjevalno moč), tj. tistih lastnosti, ki jih ločijo od drugih stvari istega roda (v definiciji podane v vrstni razliki) oziroma tiste, zaradi katerih nekaj je to, kar je. V resnici gre za nekaj, kar nam je dobro znano, in sicer učenje iz primerov. Njegov cilj je izpeljava posplošitev o predmetih (iz njihovih znanih primerkov), s pomočjo katerih je sistem te predmete kasneje v svetu sposoben prepoznati in razvrstiti brez kakršnekoli tuje pomoči. Da bi bilo takšno urjenje uspešno, mora vključevati veliko količino podatkov oziroma učnih primerkov, iz katerih je razviden vzorec in za katere še ne obstaja nobena formula, po kateri bi sistem predmete glede na dano značilnost med seboj že lahko uspešno razlikoval.

Ko takšno podatkovno bazo zagotovimo, pa imamo na voljo tri osnovne postopke strojnega učenja, *klasifikacijo*, *grupiranje*[28] in *regresijo*.[29] Pri klasifikaciji gre za sistematično umeščanje primerkov v že znane razrede. Slednji so med seboj jasno ločeni in rabijo kot orodje, s katerim podatke uredimo. Razvrščanje v razrede poteka glede na dane značilnosti, razredi pa imajo svoja imena, t. i. *oznake*.[30] V podatkovnem nizu slik sadja je npr. ena oznaka limona, druga marelica in tretja jagoda, sistem pa z njihovo pomočjo glede na določene lastnosti, ki jih predmeti imajo, sadje razvršča.

[26] Zavest, ki vključuje mentalna stanja, individualizirana na osnovi tega, »kako je biti« (angl. *what it's like*).
[27] Ne bomo se spuščali v to, kateri pristop je pravilen, šibkejši ali močnejši, dejstvo je, da lahko kognitivne sposobnosti oziroma inteligentno obnašanje z različnimi metodami in modeli uspešno posnemamo že danes.
[28] Angl. *clustering*.
[29] Angl. *regression*.
[30] Angl. *labels*.

Tudi pri grupiranju gre za sortiranje elementov v razrede, ampak s to razliko, da njihov imena pred začetkom postopka še niso znana. Tukaj algoritem sam opravi to nalogo, tj. najprej določeno količino primerkov razporedi po podobnih ali skladnih značilnostih, potem pa za vsako takšno množico ustvari oznako. Rezultat so skupine podobnih elementov, ki jih lahko razumemo tudi kot kategorije, ki se med seboj razlikujejo glede na svoje tipične lastnosti. Pri regresiji pa gre za iskanje matematične zveze med dvema značilnostma. Ugotoviti želimo, ali ena lastnost vpliva na drugo (ciljno lastnost) in ali lahko potem s pomočjo prve napovemo vrednost druge, npr. ali zaposlitev na banki vpliva na plače njenih zaposlenih (ciljna značilnost) in ali lahko iz tega, da nekdo dela na banki napovemo, koliko zasluži. Ciljne značilnosti ne želimo v celoti pojasniti, ampak zgolj ugotoviti, kaj (če sploh) nanjo vpliva in kako močno.« (Dengel 2019a; Bregant 2019)

Glede na to, čemu je sistem namenjen, pa je dalje odvisno, katero od treh vrst strojnega učenja bomo izbrali. Prvo vrsto imenujemo *nadzorovano učenje*.[31] Zanj je značilno, da so v fazi urjenja vsi podatki, ki jih sistem dobi, opremljeni tudi s pravilnimi odgovori, tj. oznakami, npr. to je mačka, to je pes, to je konj. To mu omogoča, da popravlja napake in da lahko na koncu iz vseh dobljenih informacij izpelje splošni model, ki ga potem uporablja za npr. razvrščanje živali. Druga vrsta nosi naziv *nenadzorovano učenje*.[32] Tukaj podatki, ki jih sistem dobi v procesu urjenja, nimajo nobenih dodatnih oznak, iz podobnih ali istih značilnosti primerkov mora sam ustvariti skupine, ki jih imenujemo *gruče*.[33] Z drugimi besedami, algoritem vhodne podatke razdeli v več kategorij s tipičnimi značilnostmi, njihovo število in vrste pa iz dobljenih informacij izlušči sam brez nadzora učitelja. Tretji vrsti pa pravimo *vzpodbujevalno učenje*.[34] Zanj je značilno, da dobi sistem v fazi urjenja le občasno povratno informacijo o tem, kako uspešen je pri opravljanju svoje naloge. V bistvu sam razvija strategije za reševanje problemov ali opravljanje nalog, nagrada oziroma vzpodbuda v smislu pozitivne ali negativne povratne informacije pa mu omogoča, da se v bodoče izogne napakam. Tako je sposoben bolje oceniti, ali njegovo ravnanje v neki situaciji vodi k uspehu ali neuspehu, kar mu olajša izdelavo učinkovitejših načrtov za spopad s težavami, hkrati pa predstavlja tudi motivacijo za še boljše dosežke. (Dengel 2019a; Bregant 2019)

[31] Angl. *supervised learning*.
[32] Angl. *unsupervised learning*.
[33] Angl. *clusters*.
[34] Angl. *reinforcement learning*.

Danes je verjetno najbolj znan in uporabljen model UI, ki temelji na nesimbolnem predstavljanju znanja *Googlov prevajalnik*.[35] Na voljo je že več kot 15 let, pred časom pa je presedlal na t. i. *mrežno prevajanje*, katerega ključna prednost je, da ne prevaja posameznih besed, ampak cele stavke. Ker s tem do neke mere upošteva tudi kontekst, takšen sistem prevod lažje preuredi in prilagodi tako, da je podoben človeškemu. Pri odkrivanju pomena stavkov si pomaga z upoštevanjem okoliščin, v katerih so zapisani ali izrečeni, zaradi česar njegovi prevodi postajajo vse bolj naravni. Program prevaja neposredno iz enega jezika v drugega, s čimer se izogne vmesni postaji, ki povečuje verjetnost napak v končnem izdelku, kar uporabniško izkušnjo še izboljša. Upoštevanje širšega konteksta, spoštovanje semantike stavkov in neposredno prevajanje so vplivali na izboljšanje vrstnega reda besed v prevodu, kar je povečalo njegovo razumljivost. Tega stari program običajno ni bil zmožen zagotoviti, zaradi česar je bil pogosto deležen posmeha.

Od letos podpira nekaj čez 130 jezikov, sposoben pa je celo prevajati v jezik, ki ga ne pozna, tj. ni bil del njegovega urjenja,[36] če sta si oba jezika, tisti, iz katerega se prevaja in tisti, v katerega se prevaja, dovolj blizu. Njegovi prevodi so človeškim presenetljivo podobni z vidika smiselnosti, dolžine in strukture stavkov. Ni pa tako zanesljiv pri razumevanju besed, ki imajo več pomenov, kar v prevodih pogosto vodi do nesmislov, občutljiv je na slovnične napake, kakovost prevodov pa je odvisna tudi od kompleksnosti in razširjenosti jezika. Tako so s tega vidika v prednosti vplivni evropski jeziki (angleščina, nemščina, francoščina itd.), močno pa zaostajajo afriški. (Bregant 2019)

S stališča razširjenosti pa prednjačijo sistemi UI, ki nam olajšajo opravljanje tistega vsakodnevnega dela, ki za marsikoga pomeni tratenje časa. Gre za *osebne asistente*, izmed katerih so najbolj znani *Siri* (Apple), *Googlov asistent*[37] in *Alexa* (Amazon). Namesto nas lahko preko zvočnih ukazov opravijo goro nalog, če so povezani z napravami, ki njihovo delovanje podpirajo: igrajo željeno glasbo, naročajo hrano iz restavracije ali izdelke po spletu, nam berejo naša elektronska sporočila, načrtujejo urnike itd. V resnici gre za virtualne pomočnike, ki se nahajajo v »oblaku«, s katerimi se pogovarjamo v naravnem jeziku: ko postavimo vprašanje, ti zvočne valove

[35] *Google Translate*.
[36] Uspešno reševanje problemov ali odgovarjanje na vprašanja, ki jih UI v fazi učenja še ni srečala, v angl. imenujemo *Zero-Shot-Learning*.
[37] *Google Assistant*.

spremenijo v besedilo, kar jim omogoča, da zberejo potrebne informacije iz tistih virov, ki so za izvršitev zahtevane naloge relevantni.

Še več, komunikacija, ki jo obvladajo, je dvosmerna, na naše zahteve so sposobni tudi odgovoriti in to ne samo z nepristnim robotskim glasom, ampak glasom, za katerega bi lahko dali roko v ogenj, da je »človeški«. *Google Duplex*, ki je dodatek h Googlovemu asistentu in je izurjen za obdelavo naravnega jezika, je program, katerega glas je tako pristen, da iz izseka telefonskega pogovora med njimi in človekom, ne moremo ugotoviti niti, da je eden izmed sogovornikov UI, niti, kdo to je. Trenutno so njegove jezikovne naloge, če malo ironiziramo, omejene zgolj na rezervacijo mize v restavraciji ali termina pri frizerju ter posredovanje odpiralnih časov, ni pa daleč čas, ko bo dovolj napreden, da bo sposoben opravljati tudi bolj kompleksna dela.[38] (Kremp 2018; Bregant 2019)

Omenimo še mogoče najbolj razvpite modele UI, ki zadnjih nekaj let polnijo časopisne stolpce, in sicer *avtonomna vozila*. Nobena skrivnost ni, da so že nekaj časa na cesti avtomobili, ki imajo določeno stopnjo samostojnosti: npr. voznika opozorijo na nepričakovano menjavo voznega pasu (nadzor menjave/zapustitve voznega pasu), pri nizkih hitrostih v mestu sami zavirajo, da preprečijo nalet, če ocenijo, da bo voznik reagiral prepozno (sistem pomoči za zaviranje v sili), samodejno nadzirajo in prilagajajo hitrost glede na pred njimi vozeče avtomobile in celo potek cest (prilagodljiv/predvidljiv tempomat) itd. Vse to omogoča množica algoritmov, ki so se sposobni učiti, pa čeprav vožnje, okolja ali avtomobila ne razumejo tako kot mi.

V primeru povsem avtonomnega vozila pa gre za avtomobil, ki je sposoben zaznavati okolje in nas brez naše pomoči pripeljati s točke A na točko B, pri čemer se npr. sam odloči, kam zaviti, s kakšno hitrostjo peljati in kako sploh priti do cilja. Od potnika se ne v nobeni situaciji ne zahteva, da prevzame nadzor nad vozilom oz. da je v njem sploh prisoten. Takšni avtomobili danes že delujejo z visoko stopnjo zanesljivosti, kljub temu pa prihaja do občasnih napak, ki lahko ogrozijo človeško življenje. Verjetno najbolj znan takšen primer se je pred časom zgodil podjetju *Uber*, ki je testiralo samovozeči avtomobil, ta pa je pri tem s hitrostjo 70 km/h zbil kolesarko, ki je nepravilno prečkala cesto. Čeprav je vozilo možnost trka zaznalo že slabih 6 sekund pred njim, se algoritem za nadzor vožnje ni odločil za zaviranje.

[38] Program je sicer deležen utemeljenih očitkov, da s tem, ko snema naše vzorce obnašanja in preference, ki jih imamo, z namenom analize in priprave odgovora na naše zahteve, kar mu omogoča učenje in prehod na višjo, bolj inteligentno stopnjo, preveč posega v našo zasebnost.

Dejstvo je, da bi se v takšni situaciji človek odzval drugače: ali bi npr. zaviral in kolesarko pustil, da prečka cesto, ali pa se ji izognil na pločnik ali bankino. Preiskava je pozneje odkrila, da je do zbitja kolesarke prišlo, ker je nadzorni sistem sploh ni prepoznal kot človeka in da avtonomna vozila podjetja *Uber* sploh niso bila sprogramirana tako, da bi reagirala na nepravilno prečkanje ceste. (Bregant 2019)

Samo vprašanje časa je, kdaj nas bo UI poznala tako dobro, da bo na naše vprašanje, »Kam naj gremo na dopust?«, izbrala destinacijo, rezervirala nastanitev, organizirala aktivnosti in to vse v skladu z našimi zahtevami, preferencami in nagnjenji, medtem ko bomo mi po prihodu iz službe utrujeni počivali na kavču. Ali smo na takšno prihodnost pripravljeni in kakšna je sploh cena, ki bi jo morali za to plačati?

Izguba zasebnosti

Omenili bomo zgolj odpoved nečemu,[39] kar se ponuja samo od sebe in kar do določene mere, ne da bi se tega sploh prav zavedali ali da bi nas to posebno motilo, tako ali tako že počnemo. Gre za izgubo zasebnosti, nekaj, kar naj bi nam bilo z evropsko *Splošno uredbo o varovanju osebnih podatkov* (SUVP)[40] na formalni ravni zagotovljeno.

Kakorkoli, danes so algoritmi, ki pod pretvezo večje varnosti pomagajo represivnim organom pri preprečevanju kriminala in iskanju krivcev, del našega vsakdanjika ne glede na to, da namestitev nadzornih kamer, prepoznavanje obrazov s slik, ki jih naredijo, in hitra identifikacija oseb, ki so na njih, mobilni telefoni, ki stalno sporočajo lokacijo imetnika, spletni iskalniki, ki beležijo zgodovino obiskanih strani, ponudniki spletnih nakupov, ki shranjujejo naša naročila z namenom ugotavljanja naših preferenc in ponujanja podobnega blaga v prihodnosti itd., bistveno zmanjšujejo našo zasebnost in odpirajo Pandorino skrinjico zlorab. Ker tako nekdo vedno ve, kaj delamo, lahko iz tega rekonstruira našo dnevno rutino in odstopanje od nje: kje smo doma, kje smo v službi, kdaj gremo v službo in kdaj se vrnemo iz nje, kateri so naši hobiji, kaj so naši interesi itd. »Zdi se, da bi se lahko v bližnji prihodnosti uresničila nočna mora vseh tistih, ki že dolgo opozarjajo na to, da bo razvoj novih tehnologij v slogu 1984, kjer te veliki brat opazuje, ljudi povsem prikrajšal za svobodo.« (Bregant 2019: 10)

[39] Nekateri izmed ostalih moralnih problemov, ki izhajajo iz nepreračunljive rabe modelov UI, bodo analizirani v preostalih člankih tega zbornika.
[40] Angl. *General Data Protection Regulation* (GDPR).

Uporabniki na družbenih omrežjih na različne načine delijo svoje izkušnje z drugimi, tj. z besedami, glasbo, sliko, filmi, všečki itd. izražajo svoja mnenja o dogodkih v svetu, vsak izmed njih pa predstavlja primer odpovedi zasebnosti. S kritiziranjem, zavračanjem, sprejemanjem, polemiziranjem ipd., puščajo na spletu sledi in prostodušno kažejo, kaj so njihove preference, pričakovanja in prioritete, s čimer vplivajo na svoje zasebno in javno življenje. Kajti pet velikih tehnoloških podjetij *Google*, *Apple*, *Facebook*, *Amazon* in *Microsoft* (GAFAM) nikoli ne spi. Mnenja ljudi, izražena na takšne načine, UI odkriva, povezuje in interpretira, kar v splošnem imenujemo *rudarjenje podatkov*.[41] Tukaj gre v bistvu za pomoč pri identificiranju skritih vzorcev in odstopanj v njih ter ocenjevanje tega, ali so med seboj povezani: npr. združevanje kupcev z istim okusom, opozarjanje na anomalije v proizvodnem postopku, ki kažejo na napake v delovanju sistema, iskanje zveze med vremenom in količino pridelka ali industrializacijo ter podnebnimi spremembami itd.

Vse skupaj se spremeni, ko se rudarjenje podatkov,[42] izrodi v njihovo zbiranje informacij o posamezniku, uporabnik sodobnih modelov UI pa postane izdelek. Pri tem vlada *Google*, ki je s svojimi brezplačnimi storitvami od nas želel le eno, predajo naših osebnih podatkov. Ker se pri tem nismo obotavljali, s čimer smo mu omogočili neprestano črpanje informacij (*Gmail*, *Chrome*, *YouTube*, *Maps* brez predaha polnijo njegove podatkovne zbirke), danes ve, kaj zanima večino uporabnikov spleta.[43] Težava je v tem, da takšna skoncentriranost podatkov tehnološkim družbam omogoča, da s selekcijo podatkov o svetu vplivajo na to, kaj se v njem dogaja in kako ga dojemamo, zaradi česar pride do upravljanja družbe s pomočjo UI: kam bo družba zavila postane odvisno od potreb, interesov, prioritet in preferenc upravljalca, tj. *GAFAMA*. V dobi digitalnih tehnologij smo tako postali surovina, iz katere tehnološki giganti naredijo izdelek: kaj bomo kupili, koga bomo volili, kaj bomo oblekli, kam bomo šli, kaj bomo delali ipd. Ti podatki se prodajajo naprej tudi z namenom spreminjanja naših navad, običajev in želja, s čimer izgubimo pravico do

[41] Angl. *Data-Mining*.

[42] Za boljši občutek glede tega, o kakšni količini podatkov, ki je na voljo omenjenim družbam, govorimo, si oglejmo, koliko informacij je bilo na svetu znotraj različnih omrežij leta 2019 v obtoku v 1 minuti: 18 milijonov poslanih SMS sporočil, 4.3 milijona ogledov video vsebin na *YouTubu*, 481.000 poslanih čivkov, 187 milijonov poslanih elektronskih sporočil, 3.7 milijona iskanj z *Googlom*, 973.000 vpisov v *Facebook*, za 862.823 ameriških dolarjev opravljenih nakupov, 375.000 naloženih aplikacij, 67 nameščenih virtualnih asistentov itd. (Dengel 2019c; Bregant, 2020)

[43] »Tudi ostali tehnološki velikani niso nobena izjema: *Microsoft* se ukvarja s ciljnimi oglasi, tehnologijo prepoznavanja obraza, virtualno resničnostjo itd., *Facebook* preko oglasov in všečkov zbira informacije o imenih in naslovih ljudi, odnosih med njimi, njihovih družinah, lokaciji in pogovorih itd. ter je tako lastnik največje podatkovne baze o nas, *Amazon* pa preko spletne trgovine, v kateri je mogoče kupiti tako rekoč vse in *Alexi*, ki govori iz zvočnika *Echo*, podatke pa pošilja v oblak, pozna naše želje, potrebe in interese, da o naslovu dostave, kreditnih karticah in mestu nakupa niti ne govorimo.« (Bregant 2020: 8)

prihodnosti, svobode in zasebnosti. Takšno usmerjanje naših življenjskih navad, skladno z interesi tehnoloških družb, pa imenujemo *nadzorovalni kapitalizem*.[44] [45] (Grobelnik 2018; Masten 2019; Zuboff 2019)

Kaj se nam torej obeta v prihodnosti? Zaenkrat so modeli UI izdelani tako, da lahko opravljajo le eno nalogo, pa še to zgolj na nekem ozkem področju, zaradi česar še ne moremo govoriti o neki splošni UI, primerljivi s človeško. Pričakujemo lahko, da bo šel nadaljnji razvoj UI v smeri sistemov, ki bodo zmožni hkrati izvrševati različne operacije z različnih specializiranih področij (kot človek), na kakšen način se bodo do takšnega znanja dokopali, pa bomo videli.

5 Zaključek

Dartmouthski posvet je temeljil na prepričanju, da je možno katerokoli značilnost človeške inteligence opisati tako natančno, da jo lahko stroj posnema. V duhu te ideje je zakoličil pristop, ki je postal sinonim za prvi val razvoja UI, in sicer simbolno manipulacijo. Ta temelji na deduktivni formalni logiki in delovanju v dobro definiranem nespreminjajočem se okolju ter na ta način UI zagotavlja varen kontekst, znotraj katerega do neke mere in po nekih standardih realizira inteligentno obnašanje. V tem primeru govorimo o na simbolni način predstavljenem znanju, ki sistemu omogoča, da s pomočjo deduktivnega sklepanja, ki vključuje aplikacijo podatkov in uporabo pravil, izpeljuje nove sklepe o svetu in se tako uči. Kmalu se je izkazalo, da takšni programi, niso uporabni za reševanje kompleksnih nalog iz vsakdanjega negotovega sveta, ki zahteva več od izvrševanja enostavnih operacij znotraj ozko določenega specializiranega okolja. To je omogočil šele drugi val razvoja UI, ki pa v nasprotju s prvim temelji na drugačni predstavitvi znanja, ki se zgleduje po delovanju biološkega nevrona in omogoča nesimbolni pristop k modeliranju duha tudi v spreminjajočem se okolju. Vključuje induktivno sklepanje in interakcijo z okoljem ter upošteva pretekle izkušnje in statistične zakonitosti, iz česar s pomočjo verjetnosti in učnih primerkov napoveduje prihodnje dogodke. Govorimo o nevronskih mrežah, kjer gre za na nesimbolni način predstavljeno

[44] Angl. *surveillence capitalism*.

[45] »V praksi se je to pokazalo pri zlorabi osebnih podatkov podjetja *Cambridge Analytica*, ki je z aplikacijo, ponujeno *Facebooku*, s prošnjo sodelovanja v akademski raziskavi, nezakonito pridobilo osebne podatke 50 milijonov uporabnikov Facebooka (njegov dizajn je omogočal tudi to, ne samo zbiranje odgovorov) in iz njih izdelalo njihove psihografične profile, s katerimi so ugotavljali, kakšno propagando je treba uporabiti, da bodo glasovali po naročnikovih željah.« (Bregant 2020: 9)

znanje, ki sistemu dopušča, da realizira inteligentno obnašanje tako, da iz obdelave velike količine podatkov izlušči ponavljajoče se vzorce in se tako uči.

Kakorkoli, če si izposodimo Searlovo misel o tem, da je za obvladovanje jezika bistveno njegovo razumevanje, se zdi, da smo od govorjenja o inteligentnih strojih ne glede na to, da so nekatere specifične naloge ti že danes sposobni opraviti bolje od nas, še vedno precej oddaljeni. Kljub napredku na področju *obdelave naravnega jezika*,[46] ki se kaže v tem, da novodobne aplikacije v svoji tudi glasovni dvosmerni komunikaciji poleg poznavanja dejstev, do neke mere upoštevajo tudi kontekst, stroji temu, kako si inteligentno obnašanje predstavljamo, niti slučajno še niso blizu. To namreč vključuje tudi prepoznavanje uporabljenega pomena in načina, kako je nekaj povedano. Npr. ali je izjava »ta stavek vsebuje eno napako« resnična ali neresnična? Na prvi pogled se zdi, da je neresnična, saj ne vsebuje nobene pravopisne napake, lahko pa je tudi resnična, saj se napaka skriva v pomenu, ki predpostavlja nekaj, česar ni. Zdi se torej, da bi morali znati računalniki, če bi hoteli biti inteligentni v pravem pomenu besede, obdelovati naravni jezik na opisan način, da o zahtevi po sposobnosti izvrševati več opravil z različnih tudi bistveno drugačnih področij, kar smo že večkrat omenili, niti ne govorimo. Ker tako daleč še nismo, tudi ideja o prihodu *superinteligence*, umetnega sistema z inteligenco, ki bistveno prekaša inteligenco najbolj pametnih in najbolj nadarjenih ljudi, ostaja bolj kot ne futurističen konstrukt.

Viri in literatura

Aristotel (2004). *Kategorije*. Ljubljana: ZRC SAZU.

Berkeley, I. S. N. (2018). »A Computational Conundrum: »What is a Computer? A Historical Overview«. *Minds & Machines*, 28, str. 375–383.

Bregant, J. (2010). »Ali lahko stroj misli?«. *Analiza*, 4, str. 55–72.

Bregant, J. (2016). »Možgani v primežu računalnikov«. *Analiza*, 1, str. 87–114.

Bregant, J. (2019). »Umetna inteligenca v praksi (1. del): razvoj, obnašanje in učenje strojev«. *Analiza*, 2, str. 39–55.

Bregant, J. (2020). »Umetna inteligenca v praksi (2. del): nekaj etičnih pomislekov«. *Analiza*, 1, str. 5–20.

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

Copeland, J. (1993). *Artificial Intelligence: A Philosophical Introduction*. Oxford: Blackwell.

Dengel, A. (2019a). »Maschinelles Lernen – das Gehirn als Vorbild für künstliche Neuronale Netze«. V Dengel, A., Socher, R., Kirchner E. A., Ogolla, S. (urd.), *Künstliche Intelligenz: Die Zukunft von Mensch und Maschine*. Hamburg: ZEIT Akademie Gmbh, str. 23–32.

[46] Angl. *natural language processing*.

Dengel, A. (2019b). »Künstliche Intelligenz – Eine Einführung«. V Dengel, A., Socher, R., Kirchner E., A., Ogolla, S. (urd.), *Künstliche Intelligenz: Die Zukunft von Mensch und Maschine*. Hamburg: ZEIT Akademie Gmbh, str. 13–22.

Dengel, A. (2019c). »Multimedia-Data-Mining: Trends und Emotionen in Big Data erkennen«. V Dengel, A., Socher, R., Kirchner E. A., Ogolla, S., Künstliche Intelligenz: Die Zukunft von Mensch und Maschine. Hamburg: ZEIT Akademie Gmbh, str. 74–84.

Descartes, R. (1637/2007). *Razprava o metodi*. Ljubljana: Slovenska matica.

Descartes, R. (1641/1988). *Meditacije*. Ljubljana: Slovenska Matica.

Descartes, R. (1641/1985a). *Meditations*. V Cottingham, J., Stoothoff, R. in Murdoch, D. (urd.), *The Philosophical Writtings of Descartes*, Cambridge: Cambridge University Press.

Descartes, R. (1649/1985b). *The Passions of the Soul*. V Cottingham, J., Stoothoff, R., Murdoch, D. (urd.), *The Philosophical Writtings of Descartes*, Cambridge: Cambridge University Press.

Fodor, J. (1975). *The Language of Thought*. Cambridge: Harvard University Press.

Grobelnik, M. (2018). »Podatki so nova nafta. Kdor ima dostop do podatkov, lahko rešuje probleme«. *Mladina*, 39.

Haugeland, J. (1986). *Artificial Intelligence: The Very Idea*. Cambridge: The MIT Press.

Hobbes, T. (1651/2006). *Leviathan (Revised Student Edition)*. Cambridge: Cambridge University Press.

Kim, J. (2001). *Mind in a Physical World*. Cambridge: MIT Press.

Kremp, M. (2018). »Google Duplex ist gruselig gut«. *Spiegel Online* (27. november 2019). URL= https://www.spiegel.de/netzwelt/web/google-duplex-auf-der-i-o-gruselig-gute-kuenstliche-intelligenz-a-1206938.html.

Leibniz, G. W. (1679/1969). »On the General Characteristic«. V Loemker, L. E. (urd.), *Philosophical Papers and Letters*. Dordrecht: Springer.

Leibniz, G. W. (1666/1966). »On The Art of Combination«. V Parkinson, G. H. R, (urd.), *Leibniz: Logical Papers*. Oxford: Oxford University Press.

Markič, O. (1997). »Klasična kognitivna znanost in simbolni modeli«. *Analiza 1*, 1999, str. 38–52.

Markič, O. (2010). *Kognitivna znanost*. Maribor: Aristej.

Markič, O. (2019). »Prvi in drugi val umetne intelgience«. V Malec, M., Markič, O. (urd.*) Misli svetlobe in senc: razprave o filozofskem delu Marka Uršiča*. Lubljana: UL, str. 201–211.

Masten, A. (2019). »Kaj vse pomeni klik na 'Strinjam se': O ekonomiji podatkov in monopolih«. *MMC RTV SLO* (2. december 2019). URL = https://www.rtvslo.si/mmc-podrobno/na-pragu-digitalne-diktature-brez-zasebnosti-in-brez-svobode/489378.

McCulloch, W., Pitts, W. (1943/1990). »A Logical Calculus of the Ideas Immanent in Nervous Activity«. V Boden, M. A. (ur.), *The Philosophy of Artificial Intelligence*, Oxford: Oxford University Press, str. 22–39.

Mettrie, Julien Offray de (1748/1996). *Machine Man and Other Writings*. Cambridge: Cambridge University Press.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens democracy*. Crown: New York.

Russell, S., Norvig, P. (2010). *Artificial Intelligence: A Modern Approach (3rd. ed.)*. Upper Saddle River: Prentice Hall.

Searle, J. (1980/1990). »Duhovi, možgani in programi«. V Hofstadter, D. R. in Dennet, D. C. (urd.), *Oko duha*, Ljubljana: Mladinska knjiga, str. 361–379.

Turing, A. (1950/1990). »Stroji, ki računajo, in inteligenca«. V Hofstadter, D. R., Dennet, D. C. (ur.), *Oko duha*, Ljubljana: Mladinska knjiga.

Turing, A. (1936). »On Computable Numbers, with an Application to the *Entscheidungsproblem*«. *Proceedings of the London Mathematical Society*, 42, str. 230–265.

Weizenbaum, J. (1976). *Computer Power and Human Reason*. San Francisco: W. H. Freeman.

Whitehead, A. N., Russell, B. (1910–1913/1997). *Principia Mathematica*. Cambridge: Cambridge University Press.

Zuboff, S. (2019). *The Age of Surveillance Capitalism*. New York: PublicAffairs.

BITI ALI (LE) BIT: KDAJ JE UMETNA INTELIGENCA ZAVESTNA, KDAJ INTELIGENTNA IN ALI OBSTAJA RAZLIKA?

TADEJ TODOROVIĆ
Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
tadej.todorovic@um.si

Sinopsis Članek obravnava pojem umetne inteligence, zvezo med inteligenco in zavestjo ter metafizične pozicije znotraj fizikalizma, ki onemogočajo možnost zavestne umetne inteligence. V prvem delu so predstavljeni različni koncepti umetne inteligence in antropocentričnost razprave o UI, kar uokviri glavno vprašanje drugega dela, tj. kaj je zavest, kaj je inteligenca, kakšna je zveza med tema pojmoma in kaj to pomeni za razpravo o UI. Zadnji del predstavi argumentacijo, ki bi jo moral nasprotnik UI prevzeti, da bi lahko znotraj fizikalizma dosledno ugovarjal možnosti UI ter posledice, ki bi jih takšna argumentacija prinesla. Ugotavljamo, da je znotraj fizikalizma skoraj nemogoče dosledno ugovarjati možnosti zavestne UI, tudi če vztrajamo pri tezi, da sta inteligenca in zavest neločljivo povezani.

Ključne besede:
umetna inteligenca,
takšnosti,
zavest,
Turingov test,
antropomorfizem

Univerzitetna založba Univerze v Mariboru

DOI https://doi.org/10.18690/um.ff.11.2022.2
ISBN 978-961-286-675-4

TO BE OR (MERELY) A BIT: WHEN IS ARTIFICIAL INTELLIGENCE CONSCIOUS, WHEN INTELLIGENT, AND IS THERE A DIFFERENCE?

TADEJ TODOROVIĆ
University of Maribor, Faculty of Arts, Maribor, Slovenia
tadej.todorovic@um.si

Abstract The article analyses the concepts of artificial intelligence, relationship between artificial intelligence and consciousness, and physicalist metaphysical positions that rule out the possibility of conscious artificial intelligence. In the first part, various concepts of artificial intelligence are discussed, followed by the question of anthropocentricity in the discussion of AI, which frames the main problem of the second part of the article, i.e., what is consciousness, what is intelligence, what is the relationship between the two, and what that means for the discussion of AI. The final part introduces the argumentation that an opponent of AI should adopt if they wish to argue against the possibility of AI in the context of physicalism and the consequences of such argumentation. The article concludes that, in the context of physicalism, it is almost impossible to argue against the possibility of conscious AI even if one insists that intelligence and consciousness are inseparably linked.

Keywords: artificial intelligence, qualia, consciousness, Turing test, anthropomorphism

https://doi.org/10.18690/um.ff.11.2022.2 DOI
978-961-286-675-4 ISBN
University of Maribor Press

1 Uvod

Vprašanja o umetni inteligenci so v filozofiji prisotna vsaj od pojava prvih računalnikov dalje. In prav toliko stari, ker starejši res ne morejo biti, so tudi vsi argumenti, ki vztrajajo pri tem, da je kaj takšnega nemogoče. Že Alan Turing, eden izmed začetnikov moderne razprave o umetni inteligenci, je v svojem članku »Computing Machinery and Intelligence« (1950) naslovil celo vrsto (devet) različnih ugovorov proti zamisljivosti umetne inteligence, od katerih se mu je, ironično, najprepričljivejši zdel ugovor iz nadnaravnih zaznav, tj. nekateri ljudje so sposobni telepatije, jasnovidnosti in telekineze, in tega umetna inteligenca (v njegovem primeru stroji) ne bo nikdar sposobna. Dandanes je stvar drugačna, vendar manj drugačna, kot bi si morda želeli. Večina ljudi se ob omembi jasnovidnosti samo nasmehne, vendar so razlogi proti možnosti umetne inteligence (UI) še vedno priljubljeni, razen morda na področju raziskovanja umetne inteligence same (Bostrom 2014). Kakorkoli, vprašanja o umetni inteligenci pogosto izpostavijo nejasnosti različnih konceptov, s katerimi v takšnih razpravah operiramo, ter nestrinjanje glede širših, metafizičnih pozicij, ki jih akterji v razpravi predpostavljajo.

Zato v tem članku poskušam sistematično razjasniti, prvič, kaj sploh je UI in kaj mislimo s tem, da umetna inteligenca misli, in drugič, kako (naj bi) temeljna metafizična prepričanja in predpostavke glede problema zavesti vplivala in oblikovala naš odgovor na vprašanje o zamisljivosti UI.

V prvem delu je predstavljen problem pojma umetne inteligence, nato pa na primeru Turingovega testa ilustriramo antropocentričnost, ki je prisotna v razpravi o UI. To nas vodi do razprave o konceptu zavesti in konceptu inteligence ter zveze med tema konceptoma. Zdi se namreč, da v razpravi o umetni *inteligenci* govorimo o dveh stvareh: bodisi o tem, kdaj je UI zavestna, tj. kdaj lahko trdimo, da ima občutke, zaznava svet in nanj odreagira (Armstrong 1981) – če parafraziramo Nagela, potem govorimo o tem, kako je biti UI (Nagel 1989), bodisi o tem, kdaj je UI (vsaj) tako inteligentna kot človek, vendar je v razpravi velikokrat predpostavljeno, da sta zavest in inteligenca neločljivo povezani. Zato najprej ponudimo definicijo zavesti in nato še dve različni razumevanji pojma inteligenca, intelektualizirano in behavioristično orientirano inteligenco. Nato predstavimo modificiran Turingov test, kjer inteligenco razumemo kot behavioristično orientirano, kar vodi do teze, da morda v nasprotovanju možnosti UI ni sporno to, da bi bila UI tako inteligentna kot človek,

ampak da bi bila zavestna. Z drugimi besedami, morda je pri problemu UI in interpretacijah Turingovega testa in kitajske sobe bolj problematično pripisovanje zavesti kot pripisovanje inteligence ravno zaradi tega, ker razprava predpostavi, da je zavest nujen pogoj za inteligenco. Nato izpostavimo dejstvo, da znotraj filozofije duha že obstaja ideja o ne-zavestnem sistemu s človeško inteligenco, tj. filozofski zombiji. Če so možni filozofski zombiji, potem mora biti možna tudi ne-zavestna UI (s človeško inteligenco). To tezo dodatno podkrepimo s primerom sodobne nevronske mreže *AlphaZero*, ki v nasprotju s starejšimi programi, ki rešujejo probleme s surovo silo, posnema človeško razmišljanje. Predpostavka, da je pravi koncept inteligence intelektualizirana inteligenca in da sta inteligenca in zavest neločljivo povezani, v luči takšnega napredka v znanosti zato ni več na tako trdnih tleh, kot je morda bila v preteklosti.

V zadnjem delu se osredotočimo na naslednje vprašanje: če nasprotnik UI še zmeraj vztraja pri neločljivi povezanosti inteligence in zavesti, potem mora biti UI, da bo inteligentna v pravem pomenu besede, seveda tudi zavestna. Osredotočimo se na dve najbolj priljubljeni poziciji znotraj fizikalizma, funkcionalizem in redukcionizem, in predstavimo argument, ki ne bi dopuščal možnosti zavestne UI ter tudi neželene posledice takšne argumentacije, ki bi jih nasprotnik UI moral sprejeti, če bi želel dosledno braniti tezo, da zavestna UI ni mogoča.

2 Kaj je umetna inteligenca?

Matematik Irving John Goof, šifrant, ki je v drugi svetovni vojni delal skupaj z Alanom Turingom, je zapisal, da bo prva superinteligenca, tj. računalnik, ki zelo presega vse intelektualne aktivnosti kateregakoli še kako pametnega človeka, »zadnji izum, ki ga bo človek moral ustvariti« (Goof 1965: 33). Podobno, vendar z veliko bolj negativnim prizvokom, je Nick Bostrom v svoji knjigi *Superinteligenca* zapisal, da bo takšen izziv, ne glede na to, ali ga bomo razrešili uspešno ali ne, »verjetno zadnji izziv, s katerim se bomo spopadli« (Bostrom 2014: vii). Zato je izjemno pomembno, da vemo, kaj sploh je superinteligenca in inteligenca nasploh. In čeprav lahko superinteligenco definiramo na dokaj preprost način, tj. inteligenca, ki presega človeško inteligenco, se izkaže, da je pojem človeške inteligence, na katerem definicija superinteligence sloni, notorično težko opredeliti. Na podoben problem zaradi istih razlogov naletimo, ko želimo govoriti o umetni inteligenci, tj. o inteligenci, ki naj bi v nekih ozirih dosegla oziroma bila enakovredna človeški

inteligenci. Vendar ali v primeru umetne inteligence govorimo o napravi, ki se bo obnašala tako, kot se obnašamo mi, tj. katere obnašanje bo nerazločljivo obnašanju povprečnega človeka? Ali govorimo o nečem 'boljšem', o napravi, ki se bo obnašala nadvse razumno, npr. kot poosebitev modrosti same, kot recimo koncept stoiškega modreca? Se sploh mora obnašati razumno ali je dovolj, da razumno samo razmišlja – in to idealno razumno? Če se bo obnašala kot povprečen človek, bo odrezav in duhovit sogovornik hitro odvrnil, da očitno ni tako zelo inteligentna. Po drugi strani pa se bo naprava, ki bo poosebljala modrost in se obnašala kot stoiški modrec ali razmišljala idealno racionalno, v veliko pogledih razlikovala od človeške inteligence, bodisi povprečne bodisi nadpovprečne. Kaj torej je umetna inteligenca?

Russell in Norwig v svoji knjigi *Artificial Intelligence: A Modern Approach* (2010), tako imenovani 'bibliji' umetne inteligence, postavita različne definicije umetne inteligence na podlagi različnih dimenzij, in sicer na podlagi *miselnih procesov, razmišljanja* in *obnašanja* na eni strani in na podlagi ujemanja s sposobnostmi *človeka* in *idealnih* sposobnosti, tj. racionalnosti, na drugi strani. Tako prideta do štirih različnih kategorij umetne inteligence:

(1) Umetna inteligenca kot *človeško obnašanje*, tj. razumevanje UI kot sistema, katerega obnašanje je nerazločljivo od človeškega. Izvorna ideja takšne UI sega vse do Turinga (1950) in Turingovega testa, pomembno pa je omeniti, da »si raziskovalci UI v veliki meri ne prizadevajo prestati Turingovega testa, ker verjamejo, da je bolj pomembno raziskovati globlje principe inteligence kot podvojiti primerek inteligence« (Russell in Norvig 2010: 3).
(2) Umetna inteligenca kot *človeško razmišljanje*, tj. razumevanje UI kot sistema, ki razmišlja na isti način, kot razmišlja človek. Da lahko razvijemo takšno UI, moramo razumeti, kako deluje človeški um – skozi introspekcijo, psihološke eksperimente ipd. Področje kognitivne znanosti povezuje različne računalniške modele iz UI in eksperimentalne tehnike iz področja psihologije, iz katerih se sestavlja natančna slika človeškega uma. (Russell in Norvig 2010: 3). Pomembno je poudariti, da tukaj ne gre za *idealno* ali *racionalno* razmišljanje, ampak razmišljanje, ki je karseda podobno človeškemu.
(3) Umetna inteligenca kot *racionalno razmišljanje*; tj. razumevanje UI kot naprave, ki razmišlja na 'pravi', racionalen način, tj. zgolj na podlagi silogizmov, ki so izpeljani iz zakonov logike. Največji problem takšnega

pristopa ni kodiranje silogizmov v računalniški jezik, ampak prevod neformalnega védenja v formalne izraze, ki jih logične operacije zahtevajo, še posebej, ko védenje ni stoodstotno. (Russell in Norvig 2010: 4)

(4) Umetna inteligenca kot *racionalno obnašanje*; tj. razumevanje umetne inteligence kot sistema, ki se obnaša 'idealno' racionalno, pri čemer je racionalen agent definiran kot nekdo, ki se obnaša tako, da doseže najboljši izid ali najboljši pričakovan izid. Racionalno razmišljanje je tako samo del takšnega obnašanja, v nekaterih primerih namreč ne moremo logično dokazati, katero ravnanje ali odločitev je racionalna, pa moramo vseeno narediti *nekaj*. UI kot racionalno obnašanje ima tako dve prednosti pred UI kot človeškim razmišljanjem in UI kot racionalnim razmišljanjem: je več kot zgolj sklepanje na podlagi 'zakonov razmišljanja', ker je pravilno sklepanje le eden izmed mehanizmov doseganja racionalnosti, hkrati pa je bolj dovzetno za znanstveni razvoj v primerjavi s pristopi, ki so osnovani samo na človeškem obnašanju in razmišljanju. (Russell in Norvig 2010: 4–5)

Zdi se, da sta najboljši definiciji umetne inteligence, ki ujameta bistvo tega, kar imamo v mislih, ko govorimo o 'pravem' konceptu UI, UI kot racionalno obnašanje (UIRO) in UI kot človeško obnašanje (UIČO). Prvo lahko razumemo kot poskus ustvarjanja stoiškega modreca ali pa idealiziranega Einsteina ali Sokrata, medtem ko lahko drugo razumemo bolj v smislu ustvarjanja povprečnega človeka, vključno z vsemi hibami in napakami. Iz tega tudi sledi, da bo UIRO recimo pogrnila na Turingovem testu, medtem ko bo UIČO test prestala. Po drugi strani je UIRO z vidika družbe verjetno vredna več, saj lahko pride do novih spoznanj in rešitev, do katerih se mi ali UIČO ne moremo dokopati.

2.1 Antropocentrizem v razpravi o umetni inteligenci

Tukaj lahko izpostavimo prvo problematično predpostavko, na kateri sloni del razprave o umetni inteligenci. Tako Turingov test (Turing 1950) kot Searlova kitajska soba (Searle 1980) temeljita predvsem na ideji UIČO: UIRO bi na obeh testih pogrnila – obnašanje UIRO se namreč razlikuje od človeškega obnašanja. Z drugimi besedami, oba testa po eni strani predpostavita, da je človeško obnašanje že racionalno obnašanje, s čimer se seveda ne skladajo izsledki tako psihologije, sociologije, ekonomije in drugih družbenih znanosti, na drugem koncu pa predpostavita, da je človeško obnašanje tudi nujen pogoj za inteligentno obnašanje

nasploh. Številni avtorji so izpostavili, da »Turingov test testira človečnost, ne inteligence« (Fostel 1993), da testira »človeško inteligenco, ne inteligence nasploh« (French 1990) in da je na splošno »strašansko antropocentričen« (Hayes in Ford 1996). Obstaja veliko primerov, ki upravičujejo takšne sodbe. Če jih naštejemo samo nekaj: nekatere nečloveške živali so jasno inteligentne, vendar ne bi prestale Turingovega testa – delfini, orke, orangutani in pujsi so recimo med bolj inteligentnim; lahko tudi predpostavimo, da bi bili katerikoli vesoljci, ki bi bili dovolj napredni za medplanetno potovanje, inteligentnejši od nas, hkrati pa seveda ne bi prestali Turingovega testa. Navsezadnje bi verjetno tudi lahko trdili, da bi lahko obstajala UIRO, ki je inteligentnejša od povprečnega človeka in ne opravi Turingovega testa.[1]

Skratka, dejstvo, da nek sistem ne opravi Turingovega testa, nam ne pove prav veliko: ne pove nam, ali je sistem inteligenten ali ne, pove nam samo, da se sistem ne obnaša kot povprečen človek (oz. da ni tako inteligenten kot povprečen človek). Načeloma je sistem lahko celo inteligentnejši – morda so vesoljci, ki nas obiščejo, izjemno 'kantovska' bitja in nikdar ne želijo lagati, zato ne opravijo testa, čeprav vedo, kaj bi morali odgovoriti, da bi ga prestali. Tako je Turingov test lahko dober test samo za UIČO, ne pa tudi za UIRO.

Prav tako nam Turingov test ne pove, ali je nek sistem zavesten, tj. kdaj ima sistem občutke, zaznava svet in nanj odreagira (Armstrong 1981). Z drugimi besedami, govorimo o tem, »kako je biti nek sistem« (Nagel 1989). Namreč, če nas prepriča Cambriška deklaracija o zavesti, izjava skupine vidnih mednarodnih kognitivnih nevroznanstvenikov, nevropsihologov, nevrofarmakologov, nevroanatomistov in komputacijskih nevroznanstvenikov, ki zaključi, da »/…/ ljudje niso edini, ki imajo nevrološko podlago, ki generira zavest. Nečloveške živali, vključno s sesalci, ptiči in mnogimi drugimi bitji, vključno s hobotnicami, tudi imajo takšno nevrološko podlago« (Low et al. 2012), potem sledi, da so nekatere nečloveške živali zavestne, čeprav ne prestanejo Turingovega testa.

Ko torej govorimo o Turingovem testu (in tudi Searlovi kitajski sobi), se zdi, da tako kot pri definiciji UI ugotavljamo prisotnost dveh različnih stvari: inteligence in zavesti. Ampak ni rečeno, da lahko o zavesti in inteligenci razmišljamo kot o stikalu,

[1] Seveda obstaja možnost, da bi UIRO na Turingovem testu odgovarjala tako, kot bi predvidela, da mora odgovarjati človek in tako prelisičila test.

ki je ali vklopljeno ali izklopljeno, ali celo, kot se zdi, da to ta razprava predpostavlja, da inteligenca nastopi če in samo če nastopi tudi inteligenca. Preden lahko ponudimo utemeljene odgovore na te pomisleke, je treba vsaj na kratko predstaviti tako pojem zavesti kot tudi pojem inteligence.

3 Kaj je zavest in kaj inteligenca?

Zavest lahko razdelimo, v skladu s tradicijo razprave o zavesti, na fenomenalno in intencionalno zavest, kjer prva ustreza zgornjemu opisu »kako je biti«, tj. kako je hoditi po parku, jesti jabolčni štrudelj na plaži ali piti črni čaj po napornem dopoldnevu, medtem ko je intencionalna zavest tista vrsta zavesti, ki jo ponazarja vprašanje »O čem razmišljaš?«. Intencionalnost je torej usmerjenost uma proti nekaterim stvarem, predmetom, dogodkom ipd. (glej Siewert 2017: 2)

Kaj pa inteligenca? Kaj pomeni, da je nek sistem inteligenten? Predpostavljamo, da so inteligentni ljudje, do neke mere (vsaj) nekatere živali – kaj pa UI? Tukaj se intuitivne sodbe verjetno začnejo razlikovati, ker pojem 'inteligence' razumemo na različne načine. Tudi znotraj filozofije ni ustaljene definicije inteligence – kot zapiše Lanz: »ne v filozofiji ne v psihologiji ni ustaljene definicije koncepta inteligence« (Lanz 2000: 19). Tako ene preseneča izjava, da je »/…/ sam koncept inteligence kot čarodejev trik. Kot koncept neraziskanih regij v Afriki izgine takoj, ko ga odkrijemo« (Minsky 1997: 11). Kakorkoli, vseeno si lahko v grobem pomagamo z vsaj dvema različnima pomenoma inteligence, in sicer z intelektualizirano inteligenco in behavioristično usmerjeno inteligenco.

Intelektualizirana inteligenca ne razume inteligence kot behavioristično, ampak primarno kot notranje mentalne procese, ki nadzorujejo vedenje. Takšno razumevanje inteligence je antropocentrično, absolutno (ali je sistem inteligenten ali pa ni – nimamo spektra inteligence) in popolnoma povezano z racionalnimi mentalnimi procesi (Lanz 2000: 24). Inteligenco v takšnem smislu bi lahko pripisali izključno ljudem in ne živalim, bistvo takšnega razumevanja pa verjetno najbolje povzame van Inwagen:

> Racionalnost zaznamuje velik prepad, diskontinuiteto med človeštvom in živalmi. Narobe je, da predpostavljamo, da obstaja nekaj, česar imajo opice in sloni in bobri v manj, mi pa več, in da je posledica tega, da smo mi racionalni in oni ne. (van Inwagen 1993: 121)

Intelektualizirana inteligenca v sedanjosti izključuje možnost UI, saj pogojuje inteligenco z racionalnostjo, ki seveda zahteva ne samo fenomenalno, ampak tudi intencionalno zavest. Posledica takšnega razumevanja inteligence pa je seveda, da smo prisiljeni tudi v morda bolj protiintuitiven, in sicer da nečloveške živali niso inteligentne. Namreč, dokaj samoumevno je, da nečloveške živali kažejo določene znake inteligence oziroma inteligentnega obnašanja, tudi v vsakodnevni rabi govorimo o tem, da so živali inteligentne, da so odreagirale inteligentno ipd. Če želimo koncept inteligence aplicirati tudi na ne-človeške živali, ga moramo razumeti na drugačen način – behavioristično usmerjena inteligenca je eden izmed takšnih načinov.

Behavioristično usmerjena inteligenca uporablja pojem inteligenca kot prislov, kot način obnašanja. Biti inteligenten pomeni obnašati se inteligentno. Takšno pojmovanje ni absolutistično, torej obstaja spekter inteligence, prav tako ni vezano na racionalne mentalne procese (Lanz 2000: 24–25). Ker so v takšnem smislu inteligentne tudi živali, pojem tudi ni antropocentričen.

Na tej točki lahko vidimo, da bo odgovor na vprašanje »Ali je UI inteligentna?« odvisen ravno od našega pojmovanja inteligence. Če inteligenco razumemo kot behavioristično usmerjeno inteligenco, potem je odgovor že danes do neke mere pozitiven. Če pa inteligenco razumemo kot intelektualizirano inteligenco, potem je danes odgovor zagotovo negativen. Zdaj je tudi razvidno, zakaj se zdi, da Turingov test testira tako inteligenco kot zavest (čeprav naj bi bil to zgolj test inteligence): ravno zato, ker razume inteligenco kot intelektualizirano inteligenco, ki zahteva intencionalno zavest. Vendar to predpostavlja, da je pravilno razumevanje koncepta inteligence prav intelektualizirana inteligenca. Je to upravičeno? In še pomembneje: ali do odpora možnosti UI pride zaradi tega, ker ne želimo pripisati (behavioristično orientirane) inteligence, zavesti ali (intelektualizirane) inteligence, ker ni zavestna?

3.1 Inteligentna UI brez zavesti ali zavestna in inteligentna UI?

Argument iz odsotnosti čustev proti možnosti UI trdi, da ne glede na to, kako inteligentna je UI in kakšne sposobnosti ima, vseeno ne razmišlja v pravem pomenu besede, ker nima čustev (Hauser *Artifical Intelligence*: 4, iii). Takšen zaključek je torej osnovan na dojemanju inteligence in razmišljanja v intelektualiziranem smislu, tj. zavest je nujen pogoj za inteligenco (seveda pa še ni zadosten). Da so čustva

nepogrešljiva za inteligenco, se lahko zdi tudi protiintuitivno: »čustva so, daleč od tega, da bi bila razumljena kot nepogrešljiva racionalni misli, pravzaprav tradicionalno razumljena kot ovira le tej« (Hauser *Artifical Intelligence*: 4, iii). Običajno v razpravah, argumentaciji in razmišljanju nasploh poskušamo zavestno znižati ali celo izničiti vpliv čustev na naše razmišljanje, ravno zato, ker se zavedamo, kako negativno čustva vplivajo na 'hladno' racionalno misel. Debata o UI torej pogosto predpostavlja prav takšno vrsto neločljivosti koncepta inteligence in zavesti, ni pa jasno, ali motivacija argumentov proti UI izhaja iz nepripravljenosti pripisa inteligence ali zavesti. Pravzaprav je možno, da odpor proti UI morda izhaja iz nepripravljenosti pripisa zavesti, ne inteligence, tudi če je ta razumljena v behavioristično usmerjenem smislu. Vzemimo nekoliko modificiran Turingov test kot ilustracijo takšnega razmišljanja.

3.2 Modificiran Turingov test

Standardni Turingov test poteka na naslednji način: sodelujejo trije subjekti, UI, človek in zasliševalec. Cilj testa je, da zasliševalec, ki na začetku ne ve, kdo je UI in kdo človek – z njima komunicira preko oznak X in Y, ugotovi, kdo je človek in kdo UI. Zasliševalec lahko sprašuje X in Y karkoli želi oz. karkoli misli, da mu bo lahko pomagalo ugotoviti, kdo je kdo (Oppy in Dowe 2020; glej tudi Bregant 2014). Test se večkrat ponovi in če zasliševalec napačno določi, kdo je kdo v vsaj polovici poskusov, potem je UI test uspešno prestala. Splošni princip testa je, da takrat, ko zasliševalec na podlagi odgovorov ne more zanesljivo ugotoviti, kdo je kdo, velja, da je UI isto inteligentna kot človek (imamo UIČO). Podobne teste si lahko zamislimo za druge stvari, recimo za ugotavljanje, če lahko človeški subjekti razlikujejo med sliko, ki jo naslika UI, in sliko, ki jo naslika umetnik. Takšen test so recimo izvedli Elgammal et al. (2017), kjer je njihova UI CAN (Creative Adverserial Network) test uspešno prestala, še več, ljudje so slike UI povprečno ocenili bolje kot slike priznanih umetnikov.

Lahko si tudi zamislimo test, kjer ne želimo ugotoviti, ali je UI tako inteligentna kot človek, ampak samo, če je tako inteligentna kot neka žival, recimo pujs ali pes. Za takšen eksperiment bi seveda morali predpostaviti koncept behavioristično orientirane inteligence (niti psi niti pujsi ne premorejo intelektualizirane inteligence). V takšnem testu bi se UI morala obnašati in reševati probleme tako, kot jih zmorejo reševati povprečni psi ali pujsi. Ko bi bila UI tako razvita, da bi se obnašala

nerazločljivo od psov in pujsov in bi probleme reševala isto dobro (ali celo bolje), bi test prestala. Predpostavimo tudi, da so v skladu s Cambriško deklaracijo zavesti tako psi kot pujsi zavestna bitja (tj. imajo vsaj fenomenalno zavest), UI pa ni zavestna. Ali je takšna UI inteligentna vsaj tako, kot je inteligenten pujs ali pes? Zdi se, da ja, navkljub temu, da ni zavestna. Zdi se, da v tem primeru koncepta zavesti in inteligence nista povezana. Če bi vztrajali, da sta pojma povezana in predpostavili neke vrste šibkejšo verzijo intelektualizirane inteligence (tj. sistem je lahko tako inteligenten kot pujs in pes, če in samo če je tudi zavesten na isti način kot pujs in pes), naenkrat UI ne bi pripisali inteligence, ker seveda ni zavestna.

Takšna razlaga originalnega Turingovega testa ni možna ravno zaradi predpostavke intelektualizirane inteligence. Vendar: če predpostavimo behavioristično-orientirano inteligenco, potem je nezmožnost dosega človeške inteligence ne-zavestnega sistema zgolj empirična trditev, ki seveda ni ne dokazana ne ovržena. Če se izkaže, da ne-zavestna UI na neki točki lahko doseže takšen nivo inteligence, potem zavest ni nujen pogoj za doseganje človeške inteligence. Prav tako je, če sledimo filozofski literaturi, takšen sistem vsaj zamisljiv. Spomniti se moramo samo na filozofske zombije, molekularno in behavioristično popolnoma identične kopije ljudi, ki nimajo nobene zavesti.[2] Najpreprosteje lahko argument zamisljivosti zombijev zapišemo tako:

1. Zombiji so zamisljivi.
2. Karkoli je zamisljivo, je možno.
3. Torej so zombiji možni. (Kirk 2021: 3)

Iz tega lahko preprosto izpeljemo argument v prid UI: če so zamisljivi ne-zavestni inteligentni zombiji, potem je zamisljiva tudi ne-zavestna inteligentna UI. In če je ne-zavestna inteligentna UI zamisljiva, potem je tudi možna. Seveda zanikanje zamisljivosti zombijev takšen argument ovrže oziroma v najboljšem primeru postavi v pat pozicijo – v takšnem primeru bo samo čas podal končno sodbo glede možnosti takšne UI. Vendar lahko ponudimo dodaten razlog v prid možnosti takšne UI, in sicer nove nevronske mreže, ki posnemajo način človeškega razmišljanja. Takšen primer je šahovska UI AlphaZero.

[2] Za več o filozofskih zombijih glej npr. Kripke 1972/80; Chalmers 1996; Hill in McLaughlin 1999.

3.3 UI proti človeku – v preteklosti s surovo močjo, danes s človeškim razmišljanjem?

Že v originalnem Turingovem članku (1950) je bil eden izmed argumentov proti UI ta, da UI-e »lahko delajo samo to, kar jim ukažemo« (Turing 1950: 454) – brez kreativnosti, svobode, popolnoma deterministično in sistematično. Tudi šahovski velemojster Kasparov je dvomil, da so takšni sistemi inteligentni, četudi ga je takšne vrste program, *Deep Blue*, leta 1997 v šahovski partiji premagal. Kot je zapisal:

> /…/ *Deep Blue* sploh ni bil to, kar so predhodniki [programerjev] desetletja prej predstavljali, ko so sanjali o ustvarjanju stroja, ki bi premagal šahovskega svetovnega prvaka. Namesto računalnika, ki bi mislil in igral šah kot človek s človeško kreativnostjo in intuicijo, so naredili takšnega, ki igra kot stroj, ki sistematično oceni 200 milijonov možnih potez na šahovnici na sekundo in zmaga s surovo močjo premlevanja številk. /…/ *Deep Blue* je bil inteligenten približno tako, kot je inteligentna vaša budilka, ki jo lahko programirate. No, dejstvo, da sem izgubil proti 10 milijonov dolarjev vredni budilki, me sicer ni spravilo v boljšo voljo. (Kasparov 2010)

Idejo, da reševanje problemov s surovo močjo ne šteje kot inteligentno ravnanje, je izrazil tudi Ned Block (1981) v svojem miselnem eksperimentu, kjer se stvor, ki ga poimenuje Trdoglavec (Blockhead), odloča na podlagi odločitvenih dreves za vsak možen vnos v vseh stadijih svojega življenja. Takšen Trdoglavec bi lahko bil programiran na način, da bi se obnašal popolnoma isto kot človek, pa mu verjetno ne bi pripisali inteligence. Primer Trdoglavca služi kot protiprimer Turingovemu testu, saj izpostavi isto bojazen, ki jo je izrazila tako Lady Lovelace kot Kasparov: premetavanje številk s surovo močjo še ni inteligenca.

(Domnevna) anekdota znanega matematičnega genija Carla Friedricha Gaussa ilustrira podoben primer: ko je bil mladi Gauss v osnovni šoli, je učitelj, ki je želel imeti malo miru, naložil učencem naslednjo nalogo: sešteti vsoto vseh števil od 1 do 100. Mislil je, da bo mu to kupilo vsaj uro miru, saj bodo morali učenci seštevati vsako število posebej, podobno kot *Deep Blue* in Trdoglavec uporabljata surovo moč, da prideta do prave poteze. Vendar, že po nekaj minutah je mladi Gauss ponudil pravilni odgovor, 5050. Seveda števil ni seštel, ampak je doumel, da lahko števila 'preloži' na sredini in jih sešteje v parih – 1 + 100, 2 + 99, 3 + 98 itd. –, kjer je vsota

vseh parov 101. Takšnih parov je 50, zato je skupni seštevek preprosto 101 x 50, splošna formula za vsoto števil od 1 do *n* pa je potemtakem $n(n+1)/2$. In čeprav obe poti, tako surova moč kot kreativno razmišljanje, vodita do istega rezultata, je samo druga znak 'prave' inteligence. Ker *Deep Blue* deluje po prvem principu, zato ni inteligenten v pravem pomenu besede. Vendar: *Deep Blue* je nastal leta 1997, od takrat se je veliko spremenilo. Danes obstajajo drugačne UI, ki posnemajo človeški način razmišljanja. Primer takšne UI je *AlphaZero*, ki uporablja globoke nevronske mreže, strojno učenje in zgolj pravila igre za učenje različnih iger: lahko se nauči igrati različne igre na nadčloveškem nivoju v relativno hitrem času (izmerjeno v urah, odvisno od igre) in uporablja bolj 'človeški' pristop iskanja najboljših potez (Silver et al. 2018). Tudi Kasparov je spremenil svojo sodbo glede inteligentnosti šahovskih programov in priznava, da v »globokih mislih«, ki jih izraža *AlphaZero*, prepozna kreativnost (Kasparov 2018).

V luči obstoja takšnih nevronskih mrež, ki se lahko naučijo različne igre samo s pomočjo pravil na veliko bolj 'človeški' način, brez surove sile, se teza, da je zavest neločljivo povezana z inteligenco, ne zdi več tako samoumevna. Danes se lahko takšne nevronske mreže naučijo igrati igre (veliko bolje kot najboljši igralci teh iger na svetu) in slikati dela, ki jih ne moremo razločiti od del priznanih umetnikov (Elgammal et al. 2017): Je res tako neverjetno, da bi lahko na podoben način v naslednjih desetletjih dosegle tudi nivo človeške inteligence v behavioristično-usmerjenem smislu? Takšna UI bi v behavioristično-orientiranem smislu bila identična človeku, prav tako pa bi uporabljala človeku podoben način razmišljanja. Izsledki in novejše UI, ki temeljijo na nevronskih mrežah in globokem učenju in ne na surovi sili, tako vsaj vržejo senco dvoma na trditev, da lahko UI doseže človeško inteligenco zgolj, če je sistem oz. UI tudi zavestna.

V prvem delu smo torej predstavili različne koncepte UI, kjer smo razlikovali med človeškim obnašanjem, človeškim razmišljanjem, racionalnim obnašanjem in racionalnim razmišljanjem. Za diskusijo o možnosti UI je pomembno, o kakšnem konceptu UI govorimo – npr. UIRO bi bila verjetno tako bolj uporabna kot tudi bolj inteligentna kot povprečen človek, vendar njene inteligence ne moremo ocenjevati tako, da primerjamo obnašanje UIRO s človeškim obnašanjem – Turingov test nam v takšnih primerih ne pomaga veliko. V drugem delu smo izpostavili tudi, da je debata o UI v filozofiji bila v preteklosti v veliki meri antropocentrična – zdi se, da Turingov test služi bolje kot test človečnosti in

človeške zavesti kot pa inteligence. To nas je vodilo do vprašanja, kaj sploh je inteligenca in kakšna je zveza med inteligenco in zavestjo. Predvsem nas je zanimalo, če sta koncepta neločljivo povezana, kot je v debati pogosto predpostavljeno. Predstavili smo dva različna koncepta inteligence, behavioristično-orientirano in intelektualizirano inteligenco. Nato smo z modificiranim Turingovim testom pokazali, da behavioristično-orientiran koncept inteligence ni neločljivo povezan z zavestjo. Z miselnim eksperimentom filozofskih zombijev smo izpostavili tudi idejo, da je doseganje človeške inteligence v behavioristično-orientiranem smislu v ne-zavestnih sistemih vsaj zamisljivo in posledično možno. Da je to možno, smo podkrepili tudi s primerom UI AlphaZero, ki nakazuje, da novejše ne-zavestne UI posnemajo človeško razmišljanje in da trditev, da bodo dosegle nivo človeške inteligence v behavioristično-orientiranem smislu, ni tako neverjetna, kot je morda izgledala nekaj desetletij nazaj. Predpostavka, da je pravi koncept inteligence intelektualizirana inteligenca in da sta inteligenca in zavest neločljivo povezani, torej ni več na tako trdnih tleh, kot je morda bila v preteklosti.

Kakorkoli, če nasprotnika UI to ne prepriča, tj. še zmeraj vztraja pri neločljivi povezanosti inteligence in zavesti, potem mora biti UI, da bo inteligentna v pravem pomenu besede, seveda tudi zavestna. Vprašanje, ki bo tematizirano v zadnjem delu članka, je tako naslednje: Katere metafizične pozicije znotraj fizikalizma, če sploh katere, ne dopuščajo možnosti zavestne UI oz. katero metafizično pozicijo bi moral nasprotnik UI prevzeti, če želi dosledno zagovarjati trditev, da je zavestna UI nemogoča?

4 Fizikalizem in zavestna UI

V razpravi problema duha in pri najbolj splošnemu vprašanju tega področja, tj. kaj je zavest, obstajajo različne metafizične pozicije. V grobem jih lahko delimo na dualistične in monistične, kjer dualistične teorije zagovarjajo tezo, da obstajata dve radikalno različni substanci, mentalna in fizična (Robinson 2020), medtem ko monistične teorije trdijo, da obstaja samo ena substanca (bodisi mentalna bodisi fizična). V tem članku se bomo osredotočili na monistične teorije, specifično na fizikalistične teorije, tj. teorije, ki trdijo, da je vse, vključno z zavestjo, fizično (Stoljar 2021). Razlog za to je preprost: večina v razpravi problema duha prevzema takšno ali drugačno verzijo fizikalizma, oziroma še bolj specifično, večina prevzema takšno ali drugačno verzijo funkcionalizma ali redukcionizma, z vidika dualizma pa je zavest

(res cogitans) po definiciji nemogoče pripisati stroju kot fizični substanci, iz česar seveda sledi trivialen sklep, da je zavestna UI nemogoča. V nadaljevanju bosta predstavljena funkcionalizem in redukcionizem kot najbolj priljubljeni metafizični poziciji znotraj fizikalizma in odgovor na vprašanje, katero izmed teh pozicij bi moral nasprotnik ideje zavestne UI prevzeti.

4.1 Redukcionizem in zavestna UI

Glavna teza redukcionizma, katerega začetnika sta bila Feigl (1967) in Smart (1959), je, da lahko mentalna stanja (in zavest), kot na to namiguje ime, zreduciramo na določene fizične procese, npr. nevrološke procese. Po redukcionizmu zavest torej ni nič drugega kot določen fizičen proces, ki se odvija v možganih. Podobno kot lahko izjavimo, da voda ni nič drugega kot H_2O, lahko izjavimo, da mentalni pojavi (tj. bolečina, zavest itd.) niso nič drugega kot neki določeni fizični procesi (npr. neka nevrološka struktura).[3]

V navezavi na UI je redukcionizem dober kandidat za nasprotnike zavestne UI. Namreč, redukcionizem identificira mentalna stanja s fizičnimi stanji. Zavest kot mentalno stanje je torej identična točno določenemu fizičnemu stanju, tj. neki točno določeni nevrološki strukturi v naših možganih. UI seveda ni sestavljena iz nevronov ali možganom identičnih materialov, zato ne more biti zavestna. Argument proti zavestni UI bi lahko torej izgledal tako:

1. Zavest je točno določena nevrološka struktura.
2. UI takšne nevrološke strukture nima.
3. Torej UI ne more biti zavestna.

Seveda takšna argumentacija naleti na določene posledice, npr. če to velja za UI, potem mora veljati tudi za vesoljce:

1. Zavest je točno določena nevrološka struktura.
2. Vesoljci takšne nevrološke strukture nimajo.
3. Torej vesoljci ne morejo biti zavestni.

[3] Za več o psihofizičnem redukcionizmu glej Bregant 2004.

Posledica tega je torej, da so zavestni lahko zgolj tisti sistemi, ki imajo točno takšno nevrološko strukturo. Vendar: to je hkrati tudi eden izmed glavnih ugovorov redukcionizmu nasploh, saj onemogoča mentalna stanja ali pripis zavesti drugim sistemom, tj. sistemom, ki nimajo človeku identične nevrološke strukture. Zagovorniki redukcionizma ta problem rešujejo s tako imenovanim lokalnim redukcionizmom (glej Lewis 1969; Kim 1989; Bickle 2016), tj. s trditvijo, da so mentalna stanja identična fizičnim stanjem znotraj iste vrste. Tako torej obstajajo podobna, vendar različna mentalna stanja različnih vrst, recimo človeška zavest, pasja zavest, vesoljčeva zavest ipd.

Problem zagovornikov ne-zavestne UI je, da takšna strategija omogoča tudi zavestno UI, zato zagovornik ne-zavestne UI ne sme vztrajati pri lokalnem redukcionizmu, ampak pri bolj protiintuitivni trditvi, da noben sistem brez identične nevrološke strukture ne more biti zavesten. Sicer je možno, da se v nomološkem smislu izkaže, da res obstaja samo ena vrsta zavesti in da je lahko realizirana samo na en način, tj. s človeku identično nevrološko strukturo (in da imajo vse živali, katerim pripisujemo zavest, isto nevrološko strukturo), podobno kot obstaja samo ena realizacija diamanta – atomi ogljikov organizirani v točno določeni strukturi. Zagovorniki takšne pozicije morajo torej ugrizniti v kislo jabolko posledic in vztrajati, da noben drugi sistem ne more realizirati mentalnih stanj in specifično zavesti, kar je pri vsej pestrosti in raznolikosti narave zelo neverjetna trditev, še posebej v luči dejstva, da skorajda ni filozofa, ki bi takšno pozicijo zagovarjal.

4.2 Funkcionalizem in zavestna UI

Funkcionalizem je danes ena izmed najbolj priljubljenih pozicij glede problema duha. Gre za tezo, ki se je razvila kot odgovor na zgoraj opisan problem redukcionizma, ki ga imenujemo tudi problem večvrstne realizacije, tj. teza, da so mentalna stanja (in zavest) lahko realizirana v različnih fizičnih sistemih (glej Putnam 1967: 1975). Glavna ideja funkcionalizma je, da je, ker se zdi, da so iste vrste mentalnih stanj lahko realizirane v različnih fizičnih sistemih, edina skupna lastnost istih mentalnih stanj njihove funkcije, zato je mentalno stanje identificirano s svojo funkcijo, tj. vzročni posrednik med vnosom in iznosom (Bregant 2004: 83–84). Odgovor funkcionalizma na vprašanje, kaj so mentalna stanja, je torej naslednji: »Funkcionalistični odgovor na to, kaj so mentalna stanja je, da so to funkcionalna stanja.« (Block 1980: 172)

V povezavi z UI funkcionalist nima veliko izbire: ker funkcionalist identificira mentalna stanja s funkcionalnimi stanji, mora posledično priznati, da je nek sistem v isti vrsti mentalnega stanja takoj, ko ta sistem to funkcijo realizira. Zavest kot mentalno stanje torej realizira določeno funkcijo, preko katere je definirana. V trenutku, ko UI to funkcijo realizira, mora torej funkcionalist priznati, da je tudi UI zavestna. Prej smo omenili filozofske zombije, ki služijo tudi kot ugovor funkcionalizmu. Zdaj lahko vidimo, zakaj. Namreč, funkcionalisti bi morali priznati, da je UI zavestna, tudi če bi funkcijo, ki jo realizira zavest, UI realizirala s surovo močjo (kot Blockov Trdoglavec).

Skratka, zagovorniki funkcionalizma nimajo dobre strategije proti možnosti zavestne UI. Situacija je v primerjavi z zagovorniki redukcionizma veliko slabša: ne samo, da morajo priznati, da je UI lahko zavestna, priznati morajo tudi, da bi UI, ki bi realizirala funkcijo zavesti s surovo močjo, bila zavestna, tj. zavesten je tudi Blockov Trdoglavec.

Pristaši fizikalizma imajo torej zelo malo izbire pri zagovoru ideje, da zavestna UI ni možna. Funkcionalisti za zagovor takšne ideje strategije sploh nimajo, medtem ko redukcionisti sicer lahko v principu takšno idejo zagovarjajo, vendar so posledice takšnega argumentiranja verjetno za večino previsoke. Kdorkoli torej verjame, da je zavest v nekem smislu fizična, skorajda nima druge izbire, kot da se strinja z možnostjo, da lahko obstaja tudi zavestna UI.

5 Zaključek

Podobno kot v drugih filozofskih razpravah je eden izmed temeljnih problemov razprave o UI ta, da osnovni pojmi, kot so umetna inteligenca, zavest, inteligenca in relacije med temi pojmi, niso poenoteni. V članku smo pokazali, da obstaja več konceptov UI in da je razprava o UI bila v preteklosti do določene mere antropocentrična. Iz te antropocentričnosti med drugim izhaja predpostavka, da sta zavest in inteligenca neločljivo povezani. Če k razpravi pristopimo z drugačnim konceptom inteligence, vidimo, da se odpor UI morda ne skriva v pripisovanju inteligence, ampak zavesti, kar nakazuje tudi ideja filozofskih zombijev in sodobne nevronske mreže, primer katere je *AlphaZero*.

Nasprotnik UI mora torej, da dosledno brani svojo pozicijo, ubrati naslednjo pot: predpostaviti mora, da sta inteligenca in zavest neločljivo povezani, tj. 'pravi' koncept inteligence je intelektualizirana inteligenca. Če verjame, da je zavest v vsaj nekem smislu fizična, tj. predpostavlja fizikalizem, potem mora znotraj fizikalizma sprejeti tudi zelo problematično verzijo redukcionizma. Skratka, če verjamemo, da je zavest v nekem smislu fizična, potem ne glede na naše razumevanje koncepta inteligence in zveze med inteligenco skoraj nimamo druge možnosti, kot da se strinjamo, da UI lahko obstaja in da je lahko zavestna.

Viri in literatura

Armstrong, D. (1981). »What is consciousness?«. V *The Nature of Mind*. Ithaca: Cornell University Press.

Bickle, J. (2016). »Multiple Realizability«. V Zalta, E. N. (ur.), *The Stanford Encyclopedia of Philosophy* (izdaja poletje 2016). URL = <https://plato.stanford.edu/archives/spr2016/entries/multiple-realizability/>.

Block, N. (ur.). (1980). *Readings in Philosophy of Psychology*. Cambridge: Harvard University Press.

Block, N. (1981). »Psychologism and Behaviorism«. *Philosophical Review*, 90, str. 5–43.

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

Bregant, J. (2004). *Misel kot vzrok: ali so mentalna stanja vzročno učinkovita?* Maribor: Pedagoška fakulteta Maribor.

Bregant, J. (2014). »Stroji in zavest: problem takšnosti«. *Analiza*, 18(1/2), str. 45–74.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York in Oxford: Oxford University Press.

Elgammal, A., Liu, B., Elhoseiny, M., Mazzone, M. (2017). »CAN: Creative Adversarial Networks Generating 'Art' by Learning About Styles and Deviating from Style Norms«. *Eighth International Conference on Computational Creativity (ICCC)*, Atlanta. URL= https://arxiv.org/abs/1706.07068v1.

Feigl, H. (1967). *The "Mental" and the "Physical". The Essay and a Postscript*. Minneapolis: University of Minnesota Press.

Fostel, G. (1993). »The Turing Test is for the Birds«. *ACM SIGART Bulletin*, 4(1), str. 7–8.

French, R.M. (1990). »Subcognition and Limits of the Turing Test«. *Mind*, 99, str. 53–65.

Goof, I. J. (1965). »Speculations Concerning the First Ultraintelligent Machine«. V Alt in Rubinoff (urd.), *Advances in Computers*. New York: Academic Press, str. 31–88.

Hauser, L. »Artificial Intelligence«. V *The Internet Encyclopedia of Philosophy*, ISSN 2161-0002. URL = https://iep.utm.edu/art-inte/.

Hill, C. S. in McLaughlin, B. P. (1999). »There are Fewer Things in Reality Than Are Dreamt of in Chalmers's Philosophy«. *Philosophy and Phenomenological Research*, 59, str. 446–454.

Hayes, P. in Ford, K. (1995). *Proceedings of the International Conference on Artificial Intelligence (IJAI-95)*. Montreal, str. 972–977.

Kasparov, G. (2010). *»The Chess Master and the Computer«. New York Review of Books* (18. april 2022). URL = http://web.mit.edu/6.034/wwwbob/kasparov-article.pdf.

Kasparov, G. (2018). *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*. London: John Murray.

Kim, J. (1989). »The Myth of Nonreductive Materialism«. *Proceedings and Addresses of the American Philosophical Association*, 63(3): str. 31–47.

Kirk, R. (2021). »Zombies«. V Zalta, E. N. (ur.), *The Stanford Encyclopedia of Philosophy* (izdaja pomlad 2021). URL = <https://plato.stanford.edu/archives/spr2021/entries/zombies/>.

Kripke, S. (1972/1980). »Naming and Necessity«. V Davidson D. in Harman G. (ur.), *Semantics of Natural Language*. Dordrecht: D. Reidel, str. 253–355.

Lanz P. (2000). »The Concept of Intelligence in Psychology and Philosophy«. V Cruse H., Dean J., Ritter H. (urd.) *Prerational Intelligence: Adaptive Behavior and Intelligent Systems Without Symbols and Logic, Volume 1, Volume 2 Prerational Intelligence: Interdisciplinary Perspectives on the Behavior of Natural and Artificial Systems, Volume 3. Studies in Cognitive Systems*. Dordrecht: Springer, str. 19–30.

Lewis, D. (1969). »Review of Art, Mind, and Religion«. *Journal of Philosophy*, 66, str. 23–35.

Low, P., Panksepp, J., Reiss, D., Edelman, D., Swinderen, B. in Koch, C. (2012). »The Cambridge Declaration on Consciousness«. *Francis Crick Memorial Conference on Consciousness in Human and non-Human Animals*. URL = https://fcmconference.org/img/CambridgeDeclarationOnConsciousness.pdf.

Minsky, M. (1987). *The Society of Mind*. London: Heinemann.

Nagel, T. (1989). »What Is It Like to Be a Bat«. *The Philosophical Review*, 83(4), str. 435–450.

Oppy, G. in Dowe, D. (2020). »The Turing Test«. V Zalta, E. N. (ur.), *The Stanford Encyclopedia of Philosophy* (izdaja zima 2020). URL = <https://plato.stanford.edu/archives/win2020/entries/turing-test/>.

Putnam, H. (1967). »Psychological Predicates«. V Capitan, W.H. in Merrill, D.D. (urd.), *Art, Mind, and Religion*. Pittsbourgh: University of Pittsburgh Press, str. 37–48.

Putnam, H. (1975). »The Nature of Mental States«. V Putnam, H., *Mind, Language and Reality: Philosophical Papers, Vol. 2*. Cambridge: Cambridge University Press, str. 429–440.

Robinson, H. »Dualism«. (2020). V Zalta, E. N. (ur.), *The Stanford Encyclopedia of Philosophy* (izdaja jesen 2020 Edition). URL = <https://plato.stanford.edu/archives/fall2020/entries/dualism/>.

Russell, S. in Norvig, P. (2010). *Artificial Intelligence: A Modern Approach, tretja izdaja*. New Yersey, Prentice Hall.

Siewert, C. (2017). »Consciousness and Intentionality«. V Zalta, E. N. (ur.), *The Stanford Encyclopedia of Philosophy* (izdaja poletje 2017). URL = <https://plato.stanford.edu/archives/spr2017/entries/consciousness-intentionality/>.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A. et al. (2018). »A general reinforcement learning algorithm that masters chess, Shogi, and Go through self-play«. *Science*, 262, str. 1140–44.

Smart, J. (1959). »Sensations and Brain Processes«. *Philosophical Review*, št. 68: str. 141–156.

Stoljar, D. (2021). »Physicalism«. V Zalta, E. N. (ur.) *The Stanford Encyclopedia of Philosophy* (izdaja poletje 2021). URL = <https://plato.stanford.edu/archives/sum2021/entries/physicalism/>.

Turing, A. (1950). »Computing Machinery and Intelligence«. *Mind*, 59(236), str. 433–460.

van Inwagen, P. (1993). *Metaphysics*. Oxford: Oxford University Press.

TRANSPARENTNOST IN RAZLOŽLJIVOST KOT ZAHTEVI ZA ZAUPANJA VREDNO UMETNO INTELIGENCO

OLGA MARKIČ
Univerza v Ljubljani, Filozofska fakulteta, Ljubljana, Slovenija
olga.markic@ff.uni-lj.si

Sinopsis Umetna inteligenca je dandanes prisotna tako v vsakdanjem življenju kot na različnih področjih znanosti ter družbenega in gospodarskega življenja. Drugi val umetne inteligence se osredotoča na izdelovanje pametnih orodij, ki temeljijo na strojnem učenju. Kljub relativni uspešnosti pri napovedovanju sistemi drugega vala izkazujejo pomanjkljivosti, na katere ob vedno bolj množični uporabi opozarjajo tako računalničarji kot družboslovci in humanisti. Načrtovalci modelov se pogosto ne zavedajo dovolj, da tako učni primeri kot zastavitve ciljev odražajo družbene vrednote in so vpeti v družbeni kontekst. V prispevku bom predstavila Etične smernice za zaupanja vredno umetno inteligenco. Osredotočila se bom predvsem na zahtevi po transparentnosti in razložljivosti, ki sta dve od zahtev za zaupanja vredno umetno inteligenco. Ker modeli strojnega učenja večinoma ne temeljijo na človeku razumljivem logičnem sklepanju, se bom vprašala, v kolikšni meri jih današnji sistemi dejansko lahko izpolnjujejo.

Ključne besede:
umetna inteligenca,
strojno učenje,
razložljivost,
transparentnost,
etične vrednote

Univerzitetna založba Univerze v Mariboru

DOI https://doi.org/10.18690/um.ff.11.2022.3
ISBN 978-961-286-675-4

TRANSPARENCY AND EXPLICABILITY AS REQUIREMENTS FOR TRUSTWORTHY AI

OLGA MARKIČ
University of Ljubljana, Faculty of Arts, Ljubljana, Slovenia
olga.markic@ff.uni-lj.si

Abstract Artificial intelligence is present everywhere, in everyday life, in science, and in social institutions and economy. The second-wave AI focuses on developing smart tools based on machine learning. Despite relatively good predictive powers, the systems of the second-wave AI have some drawbacks pointed out by computer scientists as well as by social scientists and humanists. Designers of the models too often ignore the fact that training examples and goals reflect social values and are embedded in the social context. I will first present Ethics guidelines for Trustworthy AI. I will focus on the requirements of transparency and explicability, the two requirements necessary for a Trustworthy AI. And, since machine learning models are not based on understandable logical reasoning systems, I will examine whether these two requirements can be fulfilled by today's systems.

Keywords: artificial intelligence, machine learning, explicability, transparency, ethical values

https://doi.org/10.18690/um.ff.11.2022.3 DOI
978-961-286-675-4 ISBN
University of Maribor Press

1 Uvod

Umetno inteligenco dandanes uporabljamo tako običajni ljudje v vsakdanjem življenju, ko si, na primer, pomagamo s spletnimi iskalniki, uporabljamo satelitsko navigacijo ali pa nakupujemo po spletu, kot tudi strokovnjaki in znanstveniki na različnih področjih. Računalniški programi, ki temeljijo na matematičnih algoritmih, so bili v preteklosti deležni precejšnjega zaupanja. Razlogi za to so bili verjetno vezani na zaupanje v matematiko, ki je kot deduktivna znanost zagotavljala gotovost spoznanja. Če so izhodiščne premise resnične, ni mogoče, da bi prišli do neresničnega sklepa. V prvem valu umetne inteligence,[1] ki ga pogosto imenujemo kar klasična umetna inteligenca, so se raziskovalci ukvarjali z idejo podvajanja človeških zmožnosti. Osredotočali so se predvsem na psihološko raven, na modeliranje jezika in logično sklepanje, pri čemer so se opirali na deduktivno logiko. Čeprav so bila pričakovanja in napovedi v začetku zelo optimistične, je prvi val zašel v slepo ulico, ki jo označujejo tudi kot »zimo umetne inteligence« (Russell in Norvig 2010). Drugi val, ki smo mu priča zdaj, pa črpa predvsem iz kibernetike s sredine prejšnjega stoletja in modeliranja z nevronskimi mrežami. Poudarek je na indukciji, na učenju na osnovi predhodnih izkušenj in interakciji z okoljem. Posledično se s tem spremeni tudi pristop k oblikovanju modelov. Če je bila za izdelavo sistemov/modelov prvega vala potrebna predhodna analiza procesa reševanja naloge, na podlagi katere je bil potem programiran model, v drugem valu računalničarji oblikujejo algoritme za strojno učenje, pri čemer uporabljajo učne primere oziroma množice podatkov. Drugi val tako zaznamuje razvoj teorij strojnega učenja in povezovanje z drugimi disciplinami, predvsem s statistiko in teorijo verjetnosti. Zaradi velikih baz podatkov, ki jih omogoča internet, se je začel fokus premikati z algoritmov na same podatke.

Vedno bolj razširjena uporaba sistemov umetne inteligence in premik od sistemov prvega vala k sistemom drugega vala pa poleg dobrobiti, ki jih je razvoj nedvomno prinesel, vzbuja tudi nemalo pomislekov in zaskrbljenosti. Izpostavlja se vprašanje zaupanja v te nove tehnološke pristope, na kar opozarjajo predvsem družboslovci in humanisti, težav pa se začenjajo zavedati tudi računalničarji. Po eni strani gre za bojazen, da zaradi interesov tistih, ki s sistemi upravljajo (podjetja ali država), prihaja do manipulacije uporabnikov. To sicer ni nov pojav, saj željam močnejših, da bi nadzirali in manipulirali, lahko sledimo skozi zgodovino. Dobro poznane so, na primer, manipulacije s pomočjo klasičnih medijev (tisk, radio, televizija), zdaj pa smo priča

[1] Razdelitev na prvi in drugi val je navdahnjena z Cantwell Smith (2019) in povzeta po Markič (2021).

manipulacijam s pomočjo družabnih omrežij. Vendar, kot bom pokazala v nadaljevanju, današnja 'pametna' orodja sprožajo vprašanje zaupanja tudi v kontekstih, kjer ni prisotno namerno zavajanje. Sistemi drugega vala, ki uporabljajo strojno učenje, za razliko od orodij klasične umetne inteligence, ne temeljijo na človeku razumljivem logičnem sklepanju. Cilj sistemov globokih nevronskih mrež je prepoznati vzorce, klasificirati in poiskati napovedi. Vendar do rešitev za različne naloge strojnega učenja sistem prihaja na način, ki je potencialno netransparenten[2] za človeka, tako na strani raziskovalca kot na strani uporabnika. Sistem predstavlja nekakšno 'črno škatlo', kjer uporabniki (pogosto pa tudi načrtovalci) nimajo razlage, na kakšen način je sistem prišel do končnega rezultata, kar zmanjšuje zaupanje. Zato se je pojavila potreba po urejanju področja na način, ki bo skladen s sicer sprejetimi demokratičnimi standardi. V prispevku bom predstavila Etične smernice za zaupanja vredno umetno inteligenco, temeljne vrednote in zahteve, ki bi jih moral sistem umetne inteligence izpolniti, ter se osredotočila na vprašanje razložljivosti in transparentnosti.

2 Umetna inteligenca in algoritmi

Kot je bilo omenjeno v uvodu, se je opredelitev umetne inteligence spreminjala skozi čas. V filozofskih diskusijah (glej npr. Bringsjord in Govindarajulu 2020; Markič 2021) se največkrat deli na splošno oziroma močno umetno inteligenco, katere končni cilj je ustvariti »stroje, ki mislijo, se učijo in ustvarjajo« (Simon v Russell in Norvig 2010: 27), in šibko oziroma ozko umetno inteligenco, ki razvija predvsem pametna orodja na izbranem področju. V tem prispevku se bomo osredotočili na slednjo, saj v tem trenutku najbolj zaznamuje naše življenje. V dokumentu Organizacije za gospodarsko sodelovanje in razvoj (OECD) je umetna inteligenca opredeljena takole:

> Sistem umetne inteligence je strojni sistem, ki lahko vpliva na okolje s podajanjem napovedi, priporočil ali odločitev za dano množico podatkov. Uporablja podatke, pridobljene strojno in/ali s pomočjo človeka, zato da (i) zaznava resnično in/ali virtualno okolje; (ii) abstrahira te zaznave v modele s pomočjo avtomatske analize (npr. strojnim učenjem) ali ročno; (iii) s pomočjo modelov sklepanja predvidi možne izide. Sistemi umetne inteligence so

[2] Sama raje uporabljam izraza transparentnost, (ne)transparenten, v slovenskih tekstih sta sinonimno uporabljena tudi izraz preglednost, (ne)pregleden (npr. v Smernicah za zaupanja vredno umetno inteligenco 2019).

> oblikovani tako, da delujejo z različnimi stopnjami avtonomije. (OECD.AI Policy Observatory n.d.)

Če je bila umetna inteligenca v začetkih predvsem domena znanstvenikov, v javnost pa je prišla preko filmov in znanstvene fantastike, pa je, kot smo omenili že v uvodu, dandanes prisotna že v našem vsakdanjem življenju. Ed Finn (2017) poudarja, da se je naš odnos do računalnikov spremenil proti koncu prvega desetletja našega stoletja, ko smo v žepih kot zveste spremljevalce začeli nositi pametne telefone in namesto o strojni opremi začeli govoriti o aplikacijah in uslugah. Telefoni niso bili več samo pripomočki, ki jih občasno uporabljamo, ampak smo jim začeli zaupati pri izbiri poti, prijateljev in vsebin, vrednih ogleda. Kot pravi Finn, smo z vsakim klikom in sprejemom pogojev uporabe sprejeli idejo, da veliki podatki, senzorji in različne oblike strojnega učenja lahko modelirajo in uravnavajo vse vrste kompleksnih sistemov, od izbire pesmi do napovedi kriminala. Ključno vlogo pa je prevzel izraz algoritem. Algoritmi so povsod, prevladujejo na borzah, skladajo glasbo, vozijo avtomobile, pišejo članke. (Finn 2017: 15) Sama beseda algoritem naj bi izhajala iz latiniziranega imena Algoritmi za perzijskega matematika iz 9. stoletja al-Khwārizma (Gillespie 2016: 19). Skozi čas je izraz algoritem dobil pomen postopka, ki opisuje množico matematičnih navodil za manipuliranje podatkov oziroma postopek, ki kar najhitreje pripelje do želenega rezultata. Na primer, Evklidov algoritem za iskanje največjega skupnega delitelja. Finn poudarja, da se je v računalništvu uveljavila pragmatistična opredelitev – kot »metoda za reševanje problema«, »osvetljevanje poti med problemi in rešitvami« (Finn 2017: 18). Donald Knuth je v klasičnem delu *The Art of Computer Programming* zapisal: »algoritem je končna množica pravil, ki daje zaporedje za reševanje določenega tipa problemov« (Knuth 1997: 4). A kot opozarja Terleton Gillespie, je algoritem zgolj recept, ki ga sestavljajo programabilni koraki. Pred tem mora biti opredeljen model, ki dejansko formalizira problem, opredeli cilj in ga predstavi v računskih izrazih. Kompleksna družbena aktivnost in vrednote so prevedene v funkcionalne interakcije spremenljivk in korakov. Vprašanje, kaj je relevantno, kaj pa je družbena sodba, postane del modela. (Gillespie 2016: 19–20). Različni algoritmi lahko znotraj danega modela dosežejo isti rezultat. Na primer, različni algoritmi za razvrščanje po abecedi, ki pa se lahko razlikujejo po hitrosti. A tisto, kar predstavlja družbeno relevantno razliko, se bolj nanaša na sam problem, ki ga rešujemo, na način, kako predstavimo spremenljivke in kako izberemo ter predstavimo cilj. V našem primeru bi se lahko vprašali, zakaj sploh razvrščati po abecedi.

V širši javnosti in tudi med družboslovnimi raziskovalci je izraz algoritem prevzel mnogo širši pomen. Kot poudarja Gillespie, je postal okrajšava za vse prej povedano: za algoritem v ožjem smislu, model, izbrani cilj, podatke, učenje na podatkih, aplikacijo, strojno opremo. Tako je postal ime za določeno vrsto družbenotehničnega sistema, ki proizvaja znanje in se odloča, in v katerem so ljudje, reprezentacije in informacije ponujeni kot podatki, ki so eden z drugim postavljeni v sistematične/matematične odnose, in kjer jim je pripisana izračunana vrednost. Gre za besedno figuro sinekdoho, ko se celota poimenuje z njenim delom. (Gillespie 2016: 22–23) Če to spregledamo, ne bomo dobro razumeli, zakaj se ljudje jezijo nad Facebookovimi in Googlovimi algoritmi. Razumljeni v natančnem ozkem pomenu so algoritmi samo zaporedje navodil, a kar ljudi moti, je razumevanje v širšem smislu, kjer v procesu igrajo pomembno vlogo tudi vrednote in človeške odločitve (npr. kaj izberemo za cilj – je to zasledovanje čim večjega dobička ne glede na negativne posledice za uporabnike in družbo, kot je izpostavila Frances Haugen v svoji kritiki Facebooka). Dejansko niti ni pomembno, ali celoto imenujemo model, sistem ali algoritem, če se zavedamo, na kaj se zanašamo. A ker ima izraz algoritem dva pomena, lahko prihaja do ekvivokacije. Na primer, ponudnik poudarja, da aplikacija temelji na preverjenem algoritmu, pri čemer sugerirajo ozek pomen, nič pa na primer ne povedo, kako so bili predstavljeni in izbrani podatki, kar je dejansko pomembno pri vrednotenju aplikacije. Uporabniki tako pomislijo na šolska leta in na natančnost matematičnih algoritmov ter zato hitreje in brez pomislekov sprejmejo aplikacijo kot zaupanja vredno. A kot bom pokazala v nadaljevanju, je dandanes veliko algoritmov (razumljenih v širšem smislu) oziroma sistemov umetne inteligence, ki so netransparentni, kar uporabniku, pogosto pa tudi načrtovalcu, onemogoča razlago in razumevanje delovanja. Vprašanje, katere so pomembne sestavine sistema in na kaj moramo biti pozorni, ko jih vrednotimo, je toliko bolj pomembno, ker algoritmi postajajo vse bolj prisotni v našem življenju. Kot poudarja Danaher, gre za »neizogibno in vseprisotno uporabo računalniških algoritmov za razumevanje in nadzor sveta, v katerem živimo« (Danaher 2020: 2). Zato nekateri govorijo kar o vladavini algoritmov oziroma o »algokraciji« (Aneesh 2006; Danaher 2016; 2020).[3]

Sistemi umetne inteligence so torej predvsem orodja, ki naj bi pomagala človeku, a hkrati tudi ključno posegajo v družbene odnose in v intimo ljudi. Zato ta orodja ne morejo biti samo stvar inženirjev in znanstvenikov s področja tehnike, ampak se tičejo vseh uporabnikov. V zadnjih letih smo bili priča odmevnim nastopom

[3] Več o tem v poglavju »Umetna inteligenca, algokracija in avtonomija« (Strle in Markič 2021).

žvižgačic in žvižgačev, ki so opozorili, da velike korporacije zaradi želje po dobičku kršijo celo lastna pravila in etična načela. Odmevno razkritje leta 2018 je bila zloraba več deset milijonov Facebook računov Američanov za namene vplivanja na volitvah s strani Cambridge Analytica, svetovalne agencije, ki je delovala za Donalda Trumpa v volilni kampanji leta 2016. Pojavili so se strahovi, da algoritmi, ki določajo, kaj ljudje vidijo na platformi, dejansko povečujejo lažne novice in sovražni govor, ki jih ruski hekerji uporabljajo za poskus vpliva na volitve v Trumpovo korist (Hao 2021a). Da bi si povrnili ugled, so pri Facebooku osnovali skupino z imenom Odgovorna UI (angl. *Responsible AI*), a dejansko so se, predvsem v državah izven območja Severne Amerike in Evrope (Myanmar, Honduras, Etiopija, India), nadaljevale zlorabe lažnih profilov v politične namene. O tem je 2020 zelo glasno spregovorila žvižgačica Sophie Zang (Hao 2021b). O neukrepanju, čeprav so vedeli za težave mladostnikov, ki so z uporabo Instagrama,[4] ki poudarja medvrstniško primerjavo teles in stila življenja, zahajali v velike stiske, je govorila še ena bivša uslužbenka Facebooka, Frances Haugen. Sama se je zato zavzela za nujno zunanjo regulacijo na področju družbenih medijev (Waterson in Milmo 2021). To je le nekaj najodmevnejših sporočil, ki so v običajnih ljudeh vzbudila nezaupanje v algoritme umetne inteligence na družbenih omrežjih. Temu lahko dodamo tudi vedno več kritik raziskovalk in raziskovalcev z akademskega področja, ki opozarjajo na pristranosti, ki vodijo v neetične odločitve (O'Neil 2016; Eubanks 2017) in na pasti s tehnologijo omogočenega nadzorovanja (Zuboff 2019). Jasno postaja, da uporaba sistemov umetne inteligence zahteva širši družbeni in etiški premislek.

3 Etične smernice za zaupanja vredno umetno inteligenco

Prav zaradi kritik in porajajočega se nezaupanja v sisteme umetne inteligence se je pokazalo, da je treba področje začeti urejati tako, da bo skladno z vrednotami demokratične družbe. Ti pritiski se pojavljajo predvsem v zahodnem svetu, so pa zahteve po bolj pravičnem urejanju prisotne po celem svetu.[5] V prispevku bom kot primer izpostavila *Etične smernice za zaupanja vredno umetno inteligenco*,[6] ki jih je kot priporočilo izdala Strokovna skupina na visoki ravni za umetno inteligenco pri Evropski komisiji leta 2019. Kot ugotavljajo, je treba priznati in upoštevati:

[4] V lasti Facbooka, zdaj Mete.
[5] V Sloveniji na primer deluje mednarodni center za umetno inteligenco IRCAI (International Center for Artificial Intelligence) pod okriljem Unesca (https://ircai.org/).
[6] V nadaljevanju *Etične smernice*.

> da sistemi umetne inteligence posameznikom in družbi sicer prinašajo znatne koristi, vendar predstavljajo tudi nekatera tveganja in imajo lahko negativne vplive, tudi take, ki jih je morda težko predvideti, opredeliti ali izmeriti (npr. vpliv na demokracijo, pravno državo in pravično porazdelitev ali na sam človeški um). Po potrebi je treba sorazmerno z obsegom tveganja sprejeti ustrezne ukrepe za zmanjšanje teh tveganj. (Etične smernice 2019: 2) .

V Smernicah so izpostavili tri elemente:

1. morala bi biti zakonita ter spoštovati vse veljavne zakone in druge predpise,
2. morala bi biti etična ter zagotavljati spoštovanje etičnih načel in vrednot ter
3. morala bi biti robustna s tehničnega in družbenega vidika, saj lahko sistemi umetne inteligence povzročijo nenamerno škodo, tudi če se uporabljajo z dobrimi nameni. (Etične smernice 2019: 6)

Temelje zaupanja vredne umetne inteligence predstavljajo štiri etična načela oziroma zahteve, ki temeljijo na temeljnih pravicah: spoštovanje človekovega dostojanstva; svoboda posameznika; spoštovanju demokracije, pravičnosti in pravne države; enakost, nediskriminacija in solidarnost; pravice državljanov (Etične smernice 2019: 12–13).

Ta štiri načela so:

(i) spoštovanja človekove avtonomije,
(ii) preprečevanja škode,
(iii) pravičnosti,
(iv) razložljivosti. (Etične smernice 2019: 14)

Izpolnjevanje prvega načela od načrtovalcev sistemov umetne inteligence zahteva, da spoštujejo temeljne pravice, na katerih temelji EU in so namenjene zagotavljanju spoštovanja svobode in avtonomije ljudi.

> Ljudje, ki komunicirajo s sistemi umetne inteligence, morajo imeti možnost, da se še naprej v celoti in dejansko samostojno odločajo in sodelujejo v demokratičnem procesu. Sistemi umetne inteligence si ne bi smeli neupravičeno podrediti ljudi ali jih siliti, zavajati, manipulirati z njimi, jih določati ali zbirati v skupine. (Etične smernice 2019: 14)

Razprave o tem, ali v kakšni meri to zahtevo izpolnjujejo aktualni sistemi umetne inteligence, kje so nevarnosti in v kakšno smer bi moral iti bodoči razvoj, so dandanes vroča tema (več o tem glej Strle in Markič 2021). Prej omenjene manipulacije, ki smo jim priča uporabniki, nas opozarjajo, da je to načelo pogosto kršeno. Včasih tudi z izgovorom, da tako narekuje izpolnjevanje drugega načela, ki govori o preprečevanju škode. Na primer, zaradi varnosti in preprečevanja kriminala in terorističnih napadov se uporablja nadzor (npr. kamere s prepoznavanjem obrazov), ki posega v avtonomijo človeka. Zavedati se je treba, da so trenja med posameznimi etičnimi načeli, ki so zapisna v abstraktni obliki, etične dileme, ki nimajo enostavnih tehnoloških rešitev, temveč zahtevajo poglobljen razmislek. A kot je zapisano, so »nekatere temeljne pravice in soodvisna načela [so] absolutni in se ne bi smeli tehtati (npr. človekovo dostojanstvo)« (Etične smernice 2019: 16). Prav zato je EU tudi predlagala nova pravila o uporabi na področjih, kjer gre za nesprejemljivo tveganje. »Vse, kar se šteje za očitno grožnjo za državljane EU, bo prepovedano: od 'družbenega točkovanja' vlad do igrač z glasovnim upravljanjem, ki spodbujajo k nevarnemu ravnanju otrok« (Evropska komisija n.d.).

Ta načela bi morala biti nato preoblikovana v konkretne zahteve za doseganje zaupanja vredne umetne inteligence in bi morala veljati za različne deležnike, ki sodelujejo v življenjskem ciklu sistemov umetne inteligence: razvijalce, uvajalce in končne uporabnike ter širšo družbo (Etične smernice 2019: 16). V tem prispevku se bom v razpravi o zaupanja vredni umetni inteligenci omejila predvsem na zadnje načelo oziroma zahtevo po razložljivosti in njeno navezavo na postopkovno pravičnost ter na bolj konkretne zahteve po preglednosti.

Načelo razložljivosti in postopkovna pravičnost sta v *Etičnih smernicah* opredeljeni takole:

> Razložljivost je ključna za vzpostavljanje in ohranjanje zaupanja uporabnikov v sisteme umetne inteligence. To pomeni, da morajo biti postopki pregledni, da je treba odkrito sporočati zmogljivosti in namen sistemov umetne inteligence

> ter da mora biti mogoče odločitve – kolikor je mogoče – razložiti tistim, na katere neposredno in posredno vplivajo. Brez takšnih informacij odločitev ni mogoče ustrezno izpodbijati. Ni vedno mogoče pojasniti, zakaj je model dal določen rezultat ali odločitev (in katera kombinacija vhodnih dejavnikov je prispevala k temu). Ti primeri se imenujejo algoritmi „črne škatle" in jim je treba nameniti posebno pozornost. V navedenih okoliščinah je treba morda sprejeti druge ukrepe za razložljivost (npr. sledljivost, možnost revidiranja in pregledno obveščanje o zmogljivostih sistema), če sistem kot celota spoštuje temeljne pravice. Potrebna stopnja razložljivosti je zelo odvisna od okoliščin in resnosti posledic, če je navedeni rezultat napačen ali drugače netočen[7]. (Etične smernice 2019: 15)

> Postopkovna razsežnost pravičnosti vključuje zmožnost izpodbijanja odločitev, ki jih sprejmejo sistemi umetne inteligence in ljudje, ki jih upravljajo, ter zmožnost uveljavljanja učinkovitega pravnega sredstva proti njim. V ta namen mora biti mogoče identificirati subjekt, ki je odgovoren za odločitev, postopki odločanja pa bi morali biti razložljivi. (Etične smernice 2019: 15)

Načelo razložljivosti podpira bolj konkretna zahteva po preglednosti elementov, pomembnih za sistem umetne inteligence: podatkov, sistema in poslovnih modelov. Vključuje sledljivost, razložljivost in obveščanje.

> **Sledljivost**. Nabore podatkov in procese, na podlagi katerih se s sistemom umetne inteligence sprejme odločitev, vključno s tistimi o zbiranju in označevanju podatkov, ter uporabljene algoritme bi bilo treba dokumentirati v skladu z najboljšimi možnimi standardi, da se omogočita sledljivost in povečanje preglednosti. To velja tudi za odločitve, ki jih sprejme sistem umetne inteligence. To omogoča opredelitev razlogov, zakaj je bila odločitev umetne inteligence napačna, kar pa bi lahko pomagalo preprečiti prihodnje napake. Zato sledljivost omogoča revidiranje in razložljivost.

[7] Na primer netočna priporočila sistema umetne inteligence pri nakupovanju vzbujajo manj pomembne etične pomisleke kot sistemi umetne inteligence, ki ocenjujejo, ali naj se posameznika, obsojenega za kaznivo dejanje, pogojno izpusti.

> **Razložljivost.** Razložljivost se nanaša na zmožnost pojasniti tehnične procese sistema umetne inteligence in s tem povezane človeške odločitve (npr. področja uporabe sistema umetne inteligence). V skladu s tehnično razložljivostjo morajo ljudje razumeti odločitve, ki jih sprejme sistem umetne inteligence, in jim biti zmožni slediti. Poleg tega je morda treba doseči kompromise med povečanjem razložljivosti sistema (ki lahko zmanjša njegovo točnost) ali povečanjem njegove točnosti (na račun razložljivosti). Kadar sistem umetne inteligence pomembno vpliva na življenje ljudi, bi moralo biti mogoče zahtevati ustrezno razlago postopka odločanja sistema umetne inteligence. Takšna razlaga bi morala biti pravočasna in prilagojena strokovnemu znanju zadevnega deležnika (npr. nestrokovnjak, regulator ali raziskovalec). Poleg tega bi morala biti na voljo pojasnila o tem, koliko sistem umetne inteligence vpliva na organizacijski postopek odločanja in ga oblikuje, in pojasnila o izbiri zasnove sistema ter utemeljitev za njegovo uvedbo (kar bi zagotovilo preglednost poslovnega modela).
>
> **Obveščanje**. Sistemi umetne inteligence se ne bi smeli uporabnikom predstavljati kot ljudje; ljudje imajo pravico do seznanitve s tem, da so v stiku s sistemom umetne inteligence. To pomeni, da morajo biti sistemi umetne inteligence kot taki prepoznavni. Poleg tega bi bilo treba za zagotovitev skladnosti s temeljnimi pravicami po potrebi zagotoviti možnost, da se uporabniki odločijo za komuniciranje s človekom namesto s sistemom umetne inteligence. Nadalje, strokovnjake na področju umetne inteligence ali končne uporabnike bi bilo treba obveščati o zmožnostih in omejitvah sistema umetne inteligence, in sicer na način, primeren za zadevni primer uporabe. To bi lahko zajemalo obveščanje o stopnji točnosti sistema umetne inteligence in njegovih omejitvah. (Etične smernice 2019: 21–22)

Etične smernice izpostavljajo načela in na njih temelječe zahteve za zaupanja vredno umetno inteligenco. Vprašanje je, kako deležnike zavezati k njihovem spoštovanju, če to ni v njihovem interesu (npr. velike korporacije, kot so Google, Meta, Amazon, Tik Tok itd.). Na tem področju čaka družbo, predvsem pravnike, še veliko dela.

V zadnjem razdelku bom skušala pokazati, zakaj, četudi bi iskreno želeli spoštovati načelo razložljivosti in zahtevo po transparentnosti, to ni enostavno in je morda v obliki, ki bi zadostila standardom, kot jih želimo na področju prava in medicine, trenutno tudi neizvedljivo.

4 Sistemi strojnega učenja in zahtevi po razložljivosti in transparentnosti

Kot smo omenili v uvodu, je drugi val umetne inteligence zasnovan na induktivnih oblikah sklepanja – računalničarji namesto algoritmov kot navodil (v obliki pravil), kako rešiti določeno nalogo, pišejo algoritme za strojno učenje. Pomembno vlogo pri tem pristopu igrajo podatki, na katerih se sistem uči oziroma iz katerih sistem razbere vzorce, ki pomagajo pri napovedih (rudarjenje podatkov – angl. *data mining*). Sistemi, ki so v zadnjih letih področje umetne inteligence približali uporabnikom, temeljijo na različnih pristopih strojnega učenja. Dejansko se je začel razvoj takih modelov pospešeno razvijati v 80-ih letih prejšnjega stoletja, ko so računalničarji z odkritjem posplošenega delta pravila (angl. *back propagation rule*) lahko učili tudi večnivojske mreže (Rumelhart et al. 1986). Kot sva s kolegom Strletom na kratko predstavila v razdelku *Umetna inteligenca: prvi in drugi val* (Strle in Markič 2021), bi strojno učenje v grobem lahko razdelili na nadzorovano učenje, kjer se sistem uči na podlagi učnih primerkov in poznanih rezultatov, na nenadzorovano učenje, kjer se sistem uči sam v interakciji z okoljem, in na spodbujevano učenje, kjer se sistem v daljšem obdobju prosto odloča, ob vsaki odločitvi pa prejme nagrado, če je bila odločitev dobra, oziroma kazen, če je bila slaba. Taki sistemi so se zgledovali po delovanju živčnih mrež v možganih, zato jih pogosto imenujemo nevronske mreže. Vendar nas ime ne sme zavesti. Cilj sodobnih sistemov umetne inteligence je izdelovanje orodij na različnih področjih življenja (npr. v bančništvu, medicini, športu, pravu in vojski), ki so lahko zelo oddaljena od dejanskega delovanja živčnega sistema[8]. (Strle in Markič 2021: 104–105)

Že kmalu po uveljavitvi nevronskih mrež kot primernih modelov za klasificiranje pa so se pokazale tudi težave, ki pestijo take sisteme. Na primer, sistem se na podlagi učnih primerov nauči prepoznati določen predmet, a nauči se na osnovi zmotnih (angl. *spurious)* korelacij, ne pa na vzročnih povezavah. O takih pomanjkljivostih poročajo tudi pri sodobnih sistemih. Na primer, Lapuschkin in sodelavci (2016) so ugotovili, da pri zmagovalni metodi tekmovanja PASCAL VOC, kjer so bile slike avtomatično pobrane s platforme Flickr, sistem za prepoznavo uporablja korelacije ali kontekst v podatkih. Na primer, čolne prepoznava po prisotnosti vode, vlak po prisotnosti tirnic na sliki. Kar je še bolj šokantno, izkazalo se je, da je sistem

[8] Se pa v računski nevroznanosti uporabljajo orodja drugega vala za znanstveno preučevanje dejanskih nevronskih mrež (živčnega sistema). Npr. Human Brain Project (https://www.humanbrainproject.eu/en/).

prepoznaval konje po zaščitnem vodnem znaku. Kot poudarjata Wojciech Samek in Klaus-Robert Müller (2019), je vodni znak očiten artefakt v zbirki podatkov, ki pa je dolga leta ostal spregledan tako od organizatorjev kot udeležencev tekmovanja. Avtorja tako situacijo primerjata z znanim zgodovinskim primerom 'pametnega Hansa', konja, ki je okoli leta 1900 postal senzacija zaradi domnevne zmožnosti štetja. Kot se je kasneje izkazalo, pa konj ni obvladal matematike, ampak je približno 90 % napovedal pravilen rezultat na osnovi spraševalčevega odziva. Analizirala sta tudi več sodobnih primerov. Recimo, ko se izkaže, da globoka nevronska mreža razločuje med razredoma fotografij 'volk' in 'husky' na podlagi prisotnosti snega. Menita, da napovedovalec, podoben 'pametnemu Hansu', sicer lahko dobro napoveduje na svoji testni zbirki podatkov, a bo odpovedal v realnem svetu, ko čolni niso samo v vodi, ampak se vozijo tudi na prikolicah, ko sta tako volk kot husky lahko oba v okolju brez snega in ko konji nimajo zaščitnega vodnega znaka. A če imamo opraviti s sistemom umetne inteligence, ki predstavlja 'črno škatlo', potem uporabnik zelo težko prepozna, da sistem deluje kot 'pametni Hans' (Samek in Müller 2019: 7–8). Ti in podobni primeri kažejo, da je treba biti mnogo bolj pazljiv pri izbiri podatkov.

Bolj kot take nerodne klasifikacije pa so zanimivi sistemi, ki izkazujejo nekakšno obliko avtonomnega delovanja. Na tem mestu se ne bom spuščala v razpravo, ali bomo računalniškim sistemom kdaj lahko pripisali polno avtonomijo, je pa dejstvo, da se izraz avtonomije uporablja za nekatere sisteme umetne inteligence (npr. avtonomna vozila, avtonomna orožja). Gre za uporabo v šibkejšem pomenu, saj ti sistemi nimajo lastnih intenc in ciljev, te mu še vedno določa človek. Lahko pa takim sistemom pripišemo zmožnost, da sami izberejo pot do cilja. Na primer, znan je primer programa AlphaGo Zero (Silver et al. 2017). Ta se za razliko od svojega predhodnika AlphaGo ni učil iz množice iger mojstrov goja, ampak so ga naučili samo osnovnih pravil postavljanja belih in črnih kamnov, nato pa je, namesto da bi se naslanjal na človeško znanje o igranju goja, prek igre s samim seboj in z metodo spodbujevanega učenja sam odkrival in razvijal svoje 'znanje' o goju. Mindt in Montemayor (2020) poudarjata, da se je AlphaGo Zero na ta način sam naučil igranja, pri čemer je sam tudi odkrival strategije, ki so bile sicer znane mojstrom goja, razvil pa je tudi nekaj novih, ki jih igralci še niso poznali (Mindt in Montemayor 2020: 22–23).

Sistem je tako uspešno dosegel cilj – odlično igrati go in premagovati tekmece, a pot, kako je to dosegel, za človeka ostaja nerazložljiva. Sistem je za človeka kot 'črna škatla', je netransparenten. Ljudje si ne znamo razložiti, na kakšen način je sistem prišel do znanja. V tem konkretnem primeru to morda niti ni tako pomembno (čeprav bi si igralci goja verjetno želeli, da bi dobili razlago, zakaj je določena strategija dobra ali slaba). A sledeč *Smernicam* bi na področjih medicine, prava, zavarovalništva in bančništva ter novinarstva, torej povsod, kjer »sistem umetne inteligence pomembno vpliva na življenje ljudi, moralo biti mogoče zahtevati ustrezno razlago postopka odločanja sistema umetne inteligence« (Etične smernice 2019: 21). Zmožnost, da bi lahko preverili odločitve sistema umetne inteligence, je za zaupanje vanj zelo pomembna tako v situacijah, pri katerih ima sistem podporno vlogo pri naših odločitvah (npr. v medicinski diagnostiki, pri odločanju v pravnem postopku) kot v situacijah, kjer bi sistem praktično sam prevzel odločitve (npr. avtonomna vozila). V kolikor bi sistem umetne inteligence povzročil škodo, mora biti mogoče ugotoviti, zakaj se je to zgodilo. A če do rezultata pridemo tako, da ne razumemo, kaj se dogaja v 'črni škatli', ta zahteva ni izpolnjena.

Poleg tega, da tak sistem 'črne škatle' ne omogoča razlage, kaže še dodatno šibkost. Izkazalo se je, da je globoke nevronske mreže, ki prepoznavajo vzorce, presenetljivo lahko pretentati s slikami, ki so za običajnega človeka vidne kot naključen šum ali abstraktni geometrijski vzorci. Na primer, ko sistem črno rumene črte zmotno prepozna kot šolski avtobus.[9] Kot poroča Davide Castelvecchi v članku s pomenljivim naslovom »Can we open the black box of AI?« in podnaslovom: »Artificial intelligence is everywhere. But before scientists trust it, they first need to understand how machines learn« (2016), pa kljub mnogim predlogom do zdaj ni poznana kakšna splošna rešitev. Nguyen in sodelavci poročajo o tem, kako slike, ki so za ljudi popolnoma neprepoznavne, globoka nevronska mreža z visoko stopnjo verjetnosti klasificira kot znane predmete. Taki rezultati kažejo na zanimive razlike med človeškim vidom in trenutnimi globokimi nevronskimi mrežami in po njihovem mnenju postavljajo vprašanja o splošnosti na globokih nevronskih mrežah temelječega računalniškega vida. (Ngyuen et al. 2015)

[9] V ZDA so šolski avtobusi črno-rumene barve.

V *Etičnih smernicah* je pri zahtevi za razložljivosti opozorilo, da je včasih »morda treba doseči kompromise med povečanjem razložljivosti sistema (ki lahko zmanjša njegovo točnost) ali povečanjem njegove točnosti (na račun razložljivosti)« (Etične Smernice 2019: 21). Ob tem pa je treba upoštevati, da je »potrebna stopnja razložljivosti zelo odvisna od okoliščin in resnosti posledic, če je navedeni rezultat napačen ali drugače netočen« (Etične Smernice 2019: 15). Ta navedek jasno kaže, da *Etične smernice* raziskovalcem puščajo odprto presojo, kakšen kompromis naj napravijo. Menim, da se na tistih področjih, ko gre za pomembne odločitve, ki se tičejo posameznikovega življenja, ne bi smeli odpovedati niti utemeljitvi izbora učnih primerov (podatkov) niti razlagi samega postopka. V kolikor pa se uporabi sistem, ki ne daje človeku razumljive razlage, pa mora strokovnjak - odločevalec to jasno opredeliti in utemeljiti njegovo uporabo in dobljene rezultate v skladu s postopkovno razsežnostjo pravičnosti, kot smo jo navedli v predhodnem razdelku.

5 Zaključek

V prispevku sem pokazala, kako se uporabniki in razvijalci soočajo z vprašanji glede zaupanja v sisteme umetne inteligence. Osredotočila sem se na sisteme drugega vala, ki temeljijo na strojnem učenju in se uporabljajo za napovedovanje, klasificiranje in prepoznavanje vzorcev, zmožni pa so tudi avtonomnega odločanja do neke mere. Vprašanja o zlorabah, manipulacijah in zaupanju v orodja umetne inteligence presegajo zgolj strokovne diskusije znotraj računalništva in so v zadnjem času sprožila val kritičnih odzivov med družboslovci in humanisti, ki opozarjajo predvsem na različne oblike pristranosti in nevarnosti nadzora. Sama sem se v prispevku osredotočila predvsem na epistemski vrednoti transparentnosti in razložljivosti. Sisteme, ki sicer niso transparentni in razložljivi, lahko raziskovalci uspešno uporabljajo v poskusnih (angl. *exploratory*) kontekstih in tako pripomorejo do novih odkritij (Boge 2021). Vendar pa glede na do sedaj dostopne raziskave menim, da na tistih področjih, kjer uporabniki morajo imeti glede odločitve o svojem primeru pravico do razumljive razlage in utemeljitve odločitve, odločanje na podlagi sistemov strojnega učenja ni primerno. Tak sistem za dano nalogo ne bi bil zaupanja vreden.

Viri in literatura

Aneesh, A. (2006). *Virtual Migration*. Durham, NC: Duke University Press.

Boge, F. J. (2021). »Two Dimensions of Opacity and the Deep Learning Predicament«. *Minds and Machines*, 32, str. 43–75.

Bringsjord, S., Govindarajulu, N.S. (2020). »Artificial Intelligence«, V Zalta, E. N. (ur.), *The Stanford Encyclopedia of Philosophy* (izdaja poletje 2020). URL = https://plato.stanford.edu/archives/sum2020/entries/artificial-intelligence/.

Castelvecchi, D. (2016). »Can we open the black box of AI?«. *Nature*, 538(7623), str. 20–23.

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

Danaher, J. (2016). »The Threat of Algocracy: Reality, Resistance and Accommodation«. *Philosophy and Technology*, 29(3), str. 245–268.

Danaher, J. (2020). »Freedom in an Age of Algocracy«. V Vallor, S. (ur.), *Philosophy of Technology*. Oxford: Oxford University Press, str. 250–272.

European Commission, Directorate-General for Communications Networks, Content and Technology. (2019). *Etične smernice za zaupanja vredno umetno inteligenco*. Publications Office. URL = https://data.europa.eu/doi/10.2759/65329.

Eubaks, V. (2017). *Automating Inequality*. St. Martin's Press: New York.

Evropska komisija. n.d. »Odličnost in zaupanje v umetno inteligenco«. *European Commission* (2. julij 2022). URL = https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/excellence-trust-artificial-intelligence_sl#latest.

Finn, E. (2017). *What Algorithms Want: Imagination in the Age of Computing*. Cambridge, MA, London: The MIT Press.

Gillespie, T. (2016). »Algorithm«. V Peters, B. (ur.), *Digital Keywords: A Vocabulary of Information Society and Culture*. Princeton in Oxford: Oxford University Press, str. 18–30.

Hao, K. (2021a). »How Facebook got addicted to spreading misinformation«. *MIT Technology Review* (11. marec 2021). URL = https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/.

Hao, K. (2021b). »She risked everything to expose Facebook. Now she's telling her story«. *MIT Technology Review* (29. julij 2021). URL = https://www.technologyreview.com/2021/07/29/1030260/facebook-whistleblower-sophie-zhang-global-political-manipulation/.

Knuth, Donald. (1997). *The Art of Computer Programming. Volume 1: Fundamental Algorithms*. Reading: Addison-Wesley.

Lapuschkin, S., Binder, A., Montavon, G., Müller, K.R., Samek, W. (2016). »Analyzing classifiers: fisher vectors and deep neural networks«. *Conference on Computer Vision and Pattern Recognition (CVPR)*, str. 2912–2920.

Markič, O. (2021). »Prvi in drugi val umetne inteligence«. V Malec, M. in Markič, O. (ur.), *Misli svetlobe in senc: Razprave o filozofskem delu Marka Uršiča*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani, str. 201–211.

Mindt, G. in Montemayor, C. (2020). »A Roadmap for Artificial General Intelligence: Intelligence, Knowledge, and Consciousness«. *Mind and Matter*, 18(1), str. 9–37.

Nguyen, A., Yosinski, J. Clune, J. (2015). »Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images«. *Computer Vision and Pattern Recognition (CVPR '15), IEEE*. URL = http://arxiv.org/abs/1412.1897.

OECD.AI Policy Observatory. (n.d.) »OECD AI Principles overview«. *Organisation for Economic Co-operation and Development* (2. julij 2022). URL = https://oecd.ai/en/ai-principles.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens democracy*. Crown: New York.

Russell, S. in Norvig, P. (2010). *Artificial Intelligence A Modern Approach* (*3rd. ed.)*. Upper Saddle River: Prentice Hall.

Samek, W. in Müller, K. R. (2019). »Towards Explainable Artificial Intelligence«. V Samek, W., Montavon, E., Vedaldi A., Hansen, L.K., Müller, K. R. (urd.), *Explainable AI:Interpreting, Explaining and Visualizing Deep Learning*. Cham: Springer, str. 5–22.

Silver, D. et al. (2017). »Mastering the Game of Go Without Human Knowledge«. *Nature*, 550(7676), str. 354–359.

Strle, T. in Markič, O. (2021). *O odločanju in avtonomiji*. Maribor: Aristej.

Rumelhart, D.E., McClelland, J.L. in PDP Research Group. (1986). *Parallel Distributed Processing: Explorations in Microfeatures of Cognition, vol. 1&2*. Cambridge, MA: The MIT Press.

Waterson, J. in Milmo, D. (2021). »Facebook whistleblower Frances Haugen calls for urgent external regulation«. *The Guardian* (25. okt. 2021). URL = https://www.theguardian.com/technology/2021/oct/25/facebook-whistleblower-frances-haugen-calls-for-urgent-external-regulation.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs.

RAZISKOVALNA UMETNA INTELIGENCA IN STANDARDI TRANSPARENTNOSTI

BORUT TRPIN
Univerza Ludwig-Maximilian München, München, Nemčija
borut.trpin@lrz.uni-muenchen.de

Sinopsis Umetna inteligenca se vedno pogosteje uporablja v raziskovalne namene, bodisi kot podporno orodje bodisi kot povsem avtonomen vir znanstvenih spoznanj. Tovrstna uporaba umetne inteligence v raziskovanju odpira vprašanje, kakšne standarde transparentnosti moramo od nje zahtevati. Po študiji primera nato sledi sklep, ki se osredinja na dejstvo, da moramo od raziskovalne umetne inteligence zahtevati večjo transparentnost oz. razložljivost kot od ljudi.

Ključne besede:
umetna inteligenca,
znanstvene metode,
standardi transparentnosti,
spoznavni standardi,
filozofija znanosti,
spoznavna teorija

Univerzitetna založba Univerze v Mariboru

DOI https://doi.org/10.18690/um.ff.11.2022.4
ISBN 978-961-286-675-4

RESEARCH ARTIFICIAL INTELLIGENCE AND STANDARDS OF TRANSPARENCY

BORUT TRPIN
Ludwig-Maximilians-University Munich, Munich, Germany
borut.trpin@lrz.uni-muenchen.de

Keywords: artificial intelligence, scientific methods, standards of transparency, epistemological standards, philosophy of science, epistemology

Abstract The use of artificial intelligence in research is increasing, be it as a supporting tool that complements scientists or as a completely autonomous source of scientific findings. Such usage of artificial intelligence in research raises a question regarding the standards of transparency that we request of it. After considering a case study it is established that higher standards of transparency or explainability should be requested for artificial intelligence in research than for human scientists.

https://doi.org/10.18690/um.ff.11.2022.4 DOI
978-961-286-675-4 ISBN
University of Maribor Press

1 Uvod

Umetna inteligenca že nekaj časa ni zgolj predmet raziskovanja, temveč tudi sama poganja raziskave – bodisi kot orodje, ki pomaga pri obdelavi podatkov (npr. v fiziki; Radovic et al. 2018), vse pogosteje pa tudi kot povsem avtomatizirano orodje, ki je zmožno samo formulirati hipoteze, jih testirati in privesti do novih odkritij (glej npr. King et al. 2009). Uporaba umetne inteligence oz. vsaj določenih tovrstnih metod sega praktično v vsa raziskovalna področja, kjer imamo opravka z analizo večje količine podatkov, npr. od fizike, biologije, kemije pa tudi do filozofije (Grim in Singer 2020), kjer računalniške metode vse pogosteje služijo v podporo argumentaciji. Uporablja se tudi pri analizi filozofskih problemov.

Pri vsem tem velja opozoriti, da nastane pri uporabi umetne inteligence poseben izziv zaradi njene (ne)transparentnosti, oziroma z drugimi besedami, da imamo pri umetni inteligenci pogosto zgolj delni vpogled v postopke, ki usmerjajo njeno procesiranje, včasih pa niti tega ne (v primeru t. i. popolne črne škatle) ter posledično ne razumemo oz. niti ne moremo razumeti, kaj točno vodi do rezultatov, ki jih na tak način pridobimo. Ravno zahteve po večji transparentnosti so privedle do vedno večjega razmaha tako imenovane razložljive umetne inteligence (angl. *explainable artificial intelligence*), kjer vsaj z določenimi približki umetno inteligenco skušamo osmisliti in jo tako približati našemu razumevanju.

Vprašanje, ki se torej postavlja samo po sebi, je, kakšni bi morali biti standardi transparentnosti za umetno inteligenco v primerjavi s standardi transparentnosti, ki jih pričakujemo ob človeških odločitvah. S transparentnostjo pri tem metaforično ciljamo na to, ali imamo vpogled v razloge za določene odločitve ali ne (pri tem se naslanjamo na angleška pojma *opacity* in *transparency*, pojma neprosojnosti in transparentnosti, ki sta postala ustaljena v tovrstni nedobesedni rabi).

V splošnem obstaja več pogledov, ki jih lahko grobo razdelimo v dve skupini oz. tri, pri čemer tretja ne uživa prave podpore. Po eni strani se zdi, da bi morali biti standardi transparentnosti, ki jih pričakujemo od umetne inteligence, enaki kot standardi transparentnosti, ki jih pričakujemo od ljudi. Od ljudi ob razlagi odločitev ne moremo pričakovati razlage delovanja možganov, kvečjemu intencionalne razlage v mentalističnem jeziku (tj. do neke odločitve je prišlo, ker *verjamemo*, *mislimo*, *pričakujemo* ipd., da je nekaj tako in tako; glej tudi Dennet 1987 ter Zerilli et al. 2019 ter Günther in Kasirzadeh 2021 za diskusijo standardov transparentnosti na tej

podlagi). Kljub razvoju nevroznanosti namreč pri ljudeh popolnega vpogleda vseeno nimamo in določene plasti mišljenja nam tako ostanejo zakrite. Posledično, tako npr. Zerilli et al. (2019), tega ne moremo pričakovati niti od umetne inteligence.

Po drugi strani se zdi, da bi od umetne inteligence morali zahtevati višje standarde transparentnosti. Vsaj v principu imamo namreč lažji vpogled v algoritme kot pri ljudeh (če ne gre za primere popolnih črnih škatel), hkrati pa so algoritmi sami ključni za razumevanje odločitev. Poleg tega algoritme vsaj v osnovi zasnujemo ljudje (za še več argumentov v podporo višjega standarda transparentnosti glej Günther in Kasirzadeh 2021).

Zgolj kot možnost omenimo še tretjo varianto – od umetne inteligence bi načeloma lahko sprejeli tudi nižjo raven transparentnosti, kot jo pričakujemo od ljudi. Ta opcija je razumljivo bolj ali manj brez podpore, bi pa jo lahko zagovarjali v okviru določenega instrumentalizma v smislu, če nam umetna inteligenca zanesljivo služi, potem nam ni treba razumeti zakaj. Da to ni dobra ideja, nam kažejo primeri, kjer umetna inteligenca pri, denimo, vizualni prepoznavi odpove zaradi ljudem težko razumljivih razlogov (glej npr. Nguyen, Yosinski in Clune 2015 za primer v povezavi z vidnim prepoznavanjem ter Finlayson et al. 2017 za primer težav v medicini).

Čeprav puščam debato glede standardov transparentnosti umetne inteligence v splošnem ob strani, bom v pričujočem prispevku trdil, da moramo vsaj v okviru znanstvenega raziskovanja od umetne inteligence zahtevati višje standarde transparentnosti. To je tako, ker pri znanstvenem raziskovanju stremimo k zanesljivim spoznanjem, ki nam omogočajo opis, razlago, predvidevanje ter nadzor nad raziskovanimi pojavi. Če razlaga sklepov umetne inteligence umanjka, lahko hitro pristanemo v varljivem občutku znanja, ki to ni. Prav zmožnost razlage je namreč temelj za razumevanje (čeprav to debato puščam ob strani, za nasprotno stališče glej npr. Lipton 2009). Tak primer bi bil npr., če bi s pomočjo umetne inteligence prišli do neke ugotovitve in to zagovarjali kot rezultat znanstvene raziskave, dasiravno ne bi razumeli, zakaj sklep dejansko drži. Tako bi se kasneje lahko izkazalo, da je 'odkritje' zgolj napačen rezultat, ki je temeljil na pristranskem algoritmu. A zdi se, da tudi če bi šlo za pravilen rezultat, bi brez razumevanja procesa, ki je vodil do odkritja, šlo zgolj za naključno spoznanje, ki ne zadošča kriterijem znanstvenega odkritja.

Seveda je treba omeniti, da znanstvenih spoznanj, ki nastanejo na podlagi umetne inteligence, ne sprejemamo brez dodatnega preverjanja – podobno kot to velja tudi za znanstvena odkritja sicer. Vseeno pa gre pri sklepih na podlagi raziskovalne umetne inteligence, kot imenujem umetno inteligenco, ki se uporablja v raziskovalne (znanstvene) namene, za drugačen pristop kot pri človeškem raziskovanju. Pri slednjem se zdi, da v celotnem postopku – od snovanja hipotez do zbiranja in analize podatkov – implicitno stremimo k smiselnemu stapljanju raziskovanja s korpusom že obstoječega znanja. Po drugi strani gre pri umetni inteligenci pogosto za algoritmično iskanje povezav, ki jih razumemo le do neke mere. To lahko vodi tudi do odkritja nesmiselnih povezav oz. prepoznave vzorcev, ki so zgolj plod naključja. Kot pokaže Vigen (2015) v sicer drugačnem kontekstu, lahko ob analizi dovolj podatkov odkrijemo raznorazne povezave, ki več kot očitno niso vzročno-posledične kljub statistično visoki korelaciji. Eden od bolj humornih primerov je npr. izjemno visoka korelacija (98,9 %) med številom ločitev v ameriški zvezni državi Maine ter povprečno porabo margarine po osebi v ZDA med leti 2000 in 2009. Na podoben način bi načeloma lahko nesmiselne vzorce iz velike količine podatkov izluščila tudi umetna inteligenca.

Kot bom prikazal s študijo primera iz računalniške filozofije (angl. *computational philosophy* oz. filozofije, kjer računalniške metode igrajo osrednjo vlogo kot metoda filozofske argumentacije), lahko zanašanje na navidez smiselne rezultate oz. rezultate, ki jih ne razumemo dovolj, vodi v zmotne zaključke. S tem bom tudi poudaril potrebnost višjega standarda transparentnosti pri raziskovalni umetni inteligenci oz. vsaj potrebo po večjem pretresanju robustnosti tovrstnih rezultatov.

2 Raziskovalna umetna inteligenca

Preden se posvetimo problemom glede transparentnosti raziskovalne umetne inteligence, je smiselno pogledati, kaj raziskovalna umetna inteligenca sploh je in kako se uporablja. Najprej omenimo, da se raziskovalno umetno inteligenco pogosto precej sinonimno omenja kot umetno inteligenco, strojno učenje in globoke nevronske mreže. Pri tem sicer ne gre za popolne sinonime, a vse tri trenutno običajno temeljijo na obdelavi velike količine podatkov. Obstajajo sicer tudi določeni hibridi, kot npr. globoko učenje, ki označuje kombinacijo strojnega učenja in globokih nevronskih mrež.

Prav to osredotočanje na obravnavo velike količine podatkov in iskanje povezav in vzorcev v njih je za človeško razumevanje problematično. Kot trdita Pearl in Mackenzie (2018), nas ne zanima zgolj, *kako* pride do določenih spoznanj, temveč tudi oz. predvsem *zakaj* pride do njih. Torej nam pri razumevanju ne gre zgolj za iskanje vzorcev, temveč si želimo vpogleda v vzročno-posledične odnose, ki so za človeško razumevanje pojavov ključni.

Kljub temu pa metode umetne inteligence, sploh v vzponu je strojno učenje, že igrajo pomembno vlogo tako rekoč na vseh znanstvenih področjih. Npr. v biologiji in kemiji lahko spremljamo že povsem avtonomne eksperimente (King et al. 2009; Häse, Roch in Aspuru-Guzik 2019), svojo vlogo pa tovrstne metode igrajo tudi v digitalni humanistiki in jezikoslovju (npr. pri preučevanju vejic v slovenščini; Holozan 2013) in morda presenetljivo tudi v filozofiji (npr. agentska optimizacija v simulacijah; Douven 2020), če omenimo le nekaj primerov.

Celostnega pregleda na tem mestu ne moremo izvesti, saj je umetna inteligenca v raziskovanju zelo razširjena, smiselno pa si je nekoliko pobliže pogledati vsaj en primer. Melnikov et al. (2018) opisujejo strojno učenje v okviru fizike v kvantnem laboratoriju. V kvantnih eksperimentih imamo namreč težave z različnimi razredi prepletenosti (angl. *entanglement classes*). Melnikov in kolegi so zato uporabili t. i. projektivni simulacijski sistem, ki je zasnoval kompleksne fotonične kvantne eksperimente, ki so nato proizvedli visokodimenzionalna multifotonska stanja prepletenosti. Sistem, ki temelji na umetni inteligenci, se je torej naučil proizvesti raznolika stanja prepletenosti in izboljšal učinkovitost njihove realizacije. V tem procesu je avtonomno (ponovno) odkril eksperimentalne tehnike, ki sicer postajajo standard tudi v eksperimentih moderne kvantne optike. Pri tem je zanimivo, da tega raziskovalci niso eksplicitno zahtevali ali sistema naučili, temveč je umetna inteligenca skozi proces učenja to odkrila sama. Na podlagi tega lahko sklenemo, da lahko pričakujemo, da bo raziskovalna umetna inteligenca v prihodnosti igrala pomembno ustvarjalno vlogo, saj so se tovrstni sistemi očitno zmožni sami učiti in odkrivati raziskovalne tehnike ter na podlagi obstoječih eksperimentov oblikovati smiselne hipoteze, ki so vredne eksperimentalnega preverjanja.

Pa vendar: v kolikor tovrstne tehnike razumemo (npr. ker je sistem odkril že znane tehnike), gre seveda za primere, kjer lahko potrdimo, da je dosežek umetnega sistema izjemen. Kaj pa, v kolikor bi sistem odkril tehnike, ki nam še niso znane? Podobno se lahko vprašamo, kaj bi storili v primeru, da bi nam v preučevanje ponudil

hipoteze, ki se zdijo z našega stališča nerazumne in nevredne preučevanja. Kako jih lahko vzamemo za smiselne? Konec koncev je znan problem umetne inteligence tudi to, da potencira pristranskosti iz podatkov, kar vodi do raznolikih etičnih zagat v povezavi s tako imenovano algoritmično poštenostjo (npr. kako se znebiti rasističnih posledic algoritmov, ko umetna inteligenca obdeluje podatke o ljudeh; glej npr. Kleinberg, Ludwig, Mullainathan in Ramachan 2018).

Tako kot v primeru izogibanja algoritmične (ne)poštenosti in pristranskosti ob uporabi umetne inteligence v družbenih odločitvah je torej tudi pri uporabi umetne inteligence v znanosti ključno, da imamo dovolj visoko raven razumevanja, ki pojasni, kaj je vodilo do kakšne odločitve. Raven transparentnosti mora biti visoka prav zaradi izogibanja pristranskosti in zaradi dostopa do tega, kako poteka procesiranje v ozadju, pa tudi višja kot pri razlagi človeških odločitev. Posvetimo se torej še vprašanju glede standardov transparentnosti.

3 Standardi transparentnosti raziskovalne umetne inteligence

Umetna inteligenca je vse bolj prisotna na vseh področjih in kot smo lahko videli tudi v raziskovalne namene. Andras et al. (2018) zaradi vedno večje prisotnosti izpostavijo pomen človeškega zaupanja v umetno inteligenco in izpostavijo, da je pri tem pomembno, da so umetna inteligenca oz. njeni zaledni procesi razložljivi, kar je pogosto tudi sinonim za transparentnost oz. prosojnost umetne inteligence. Jasno se zdi, da so potrebe po transparentnosti umetne inteligence večje, ko govorimo o pomembnih vprašanjih s praktičnimi posledicami, manj pa, ko govorimo npr. o procesih, ki usmerjajo umetno inteligenco do usvojitve in uspešnega igranja igre *go* (Silver et al. 2017).

Trenutno najboljši algoritmi običajno temeljijo na metodah strojnega učenja. Ti algoritmi so zasnovani na tak način, da sami prepoznavajo skrite vzorce v podatkih in zasnujejo natančna predvidevanja o podatkih v neki domeni, ki še niso bili odkriti. Kot opozarjata npr. Günther in Kasirzadeh (2021), ta natančna predvidevanja za sabo prinesejo izjemno kompleksnost algoritmov, ki so posledično spoznavno precej nedosegljivi (netransparentni) tudi za same snovalce teh algoritmov. Tako nam umanjka razumevanje razlogov, ki vodijo do rezultatov umetne inteligence. Ali lahko na podlagi tega zahtevamo višje standarde transparentnosti od umetne inteligence kot od ljudi?

Vprašanje ima tudi praktične posledice, saj s tem namreč na nek način zaviramo razvoj umetne inteligence. Ko zahtevamo, da snovalci algoritmov poskrbijo, da je procesiranje razložljivo oz. transparentno in vsaj v principu razložljivo, hkrati izločimo oz. bistveno otežimo razvoj takih algoritmov, ki bi bili sicer skoraj povsem v svojevrstni črni škatli (npr. v primeru globokih nevronskih mrež).

Pa smo sploh upravičeni do zahteve po višjih standardih transparentnosti v primeru umetne inteligence? Zerilli et al. (2019) argumentirajo, da ne. Trdijo, da od umetne inteligence ne moremo zahtevati več kot od ljudi – standardi transparentnosti morajo biti v obeh primerih enaki, in sicer lahko v obeh primerih zahtevamo zgolj razlage na intencionalni ravni (Dennett 1987), v okviru katere ljudje izrazimo svoje praktične razloge za določeno odločitev, umetna inteligenca pa za svoje 'odločitve'. 'Odločitve' sicer navajam v navednicah, saj gre za mentalistično izrazoslovje, v praksi pa bi to izgledalo tako, da npr. pojasnimo, zakaj je raziskovalna umetna inteligenca predlagala določeno raziskovalno hipotezo na podlagi tega, da je prepoznala določene vzorce v preteklih eksperimentalnih podatkih in iz tega izračunala, da je velika verjetnost za nova odkritja v preverjanju predlagane nove hipoteze.

Kar želim poudariti, je to, da v tem primeru umanjka razlaga, kako točno je raziskovalna umetna inteligenca prišla do tega sklepa. Umanjka nam torej razlaga, na podlagi katerih algoritmov oz. zakaj natančno je prišlo do formulacije te hipoteze. Če si izposodim Marrov (1982) besednjak – umanjka nam razlaga na nivoju algoritmične oz. implementacijske analize, čeprav imamo morda dostop do funkcionalne ravni. Zerilli et al. (2019) trdijo, da to ni težava, saj tega ne pričakujemo niti pri človeškem odločanju.

Na drugi strani imamo stališče, da moramo od umetne inteligence (in s tem tudi od raziskovalne umetne inteligence) zahtevati več. Tak primer sta že omenjena Günther in Kasirzadeh (2021), ki svojo argumentacijo utemeljita na dveh ključnih točkah. Po eni strani se zdi, da vsaj pri določenih algoritmičnih odločitvah ključno vlogo igra oblika (oz. dizajn). Kot primer podata letalsko nesrečo letala Boeing 737 Max 8, ki je leta 2017 strmoglavilo in povzročilo smrt 189 ljudi, ki so bili na krovu. Do težave je prišlo, ker je računalniški sistem za podporo pri letu v preveliki meri temeljil na enem specifičnem senzorju, ki se je okvaril. Šlo je torej za napako na ravni oblike oz. dizajna umetne inteligence – prekomerno je temeljila na podatkih iz enega senzorja. Na intencionalni ravni do težave ni prišlo – sistem je 'verjel', da ravna pravilno, a ob napačni predpostavki, da so bili vhodni podatki iz senzorja zanesljivi. Za zanašanje

na umetno inteligenco je, tako Günther in Kasirzadeh (2021), ključno razumevanje ne zgolj intencionalne razlage delovanja (kot jo pričakujemo od ljudi), temveč tudi algoritmične in oblikovne, torej na čem temeljijo 'odločitve' (česar od ljudi ne pričakujemo, saj imamo omejen vpogled v možgane). V kolikor bi bilo to možno, bi enake standarde transparentnosti tako ali tako vzpostavili tudi za ljudi. Če nekdo nekaj stori zaradi diagnosticirane možganske lezije, pri razlagi vedenja prav tako ne sledimo (zgolj) razlagi prepričanj, ki so osebo vodila, temveč upoštevamo tudi nevrološko specifiko osebe.

Čeprav se zdi, da so tovrstna vprašanja za raziskovalno umetno inteligenco morda irelevantna, temu ni tako. Vzemimo npr. primer strojnega učenja v okviru medicine. V kolikor poznamo obliko algoritmov (oz. še natančneje, v kolikor je ne poznamo), lahko algoritme zlorabimo za napačno diagnosticiranje (oz. spregledamo, da je bil algoritem zaveden zaradi ljudem načeloma nerazumljivih razlogov). Finlayson et al. (2019) tako izpostavljajo, da bi se lahko tovrstne algoritme v okviru medicinskega diagnosticiranja zlorabilo kot orožje. Zloraba kot del napada je vsekakor realna možnost tudi pri raziskovalni umetni inteligenci širše, sploh v kolikor govorimo o aplikativnih raziskavah. Zamislimo si lahko, da bi lahko napadalec npr. dodal očesu nevidne piksle v vizualne podatke, ki jih obdeluje umetna inteligenca, ter tako zavrl konkurenco, npr. v farmakološkem razvoju.

Pustimo špekulacije ob strani, pri obvladovanju rezultatov umetne inteligence je poznavanje oblikovne ravni algoritmov ključno vsaj zaradi tega, da razumemo, ali so rezultati zanesljivi ali ne. Tudi če izključimo možnost zunanjih napadov, se lahko primer pretiranega zanašanja na eno samo orodje (senzor) skoraj direktno preslika iz strmoglavljenega letala na znanstveno raziskavo, ki pretirano temelji na okvarjenem ali zgolj neprimernem senzorju kot viru podatkov za strojno učenje. Čeprav so v tem primeru posledice manjše, so v primeru povsem novih raziskav tudi težje za prepoznanje (in lahko vodijo v slepo raziskovalno ulico).

Drugi primer, kjer se zdi, da moramo zahtevati drugačne standarde od umetne inteligence kot od ljudi, je, tako Günther in Kasirzadeh (2021), ko govorimo o popolnih črnih škatlah, pri katerih nimamo vpogleda v procesiranje (Creel 2020). V tem primeru torej ne moremo govoriti niti o intencionalni razlagi, saj nam niti ta ni dostopna. Razumevanje, zakaj se je sistem umetne inteligence odločil, kakor se je, povsem umanjka. V takem primeru, tako se vsaj zdi, brez obširnega testiranja robustnosti rezultatov sploh ne moremo govoriti o spoznavni zanesljivosti. Pri tem

velja izpostaviti tudi povezavo med analizo robustnosti in razlagalnim mišljenjem, kot jo je pred kratkim v nekoliko drugačnem kontekstu vzpostavil Schupbach (2018).

Prav zato, ker od znanosti torej zahtevamo visoke spoznavne standarde, da lahko govorimo o zanesljivem spoznanju oz. znanju, se torej zdi smiselno, da od raziskovalne umetne inteligence zahtevamo višjo transparentnost, kot jo zahtevamo pri razlagi človeškega znanstvenega raziskovanja. Raziskovalna umetna inteligenca se torej zdi primer, kjer je debata o dvojnih standardih transparentnosti še bolj enostavna kot pri umetni inteligenci na splošno – v znanosti so zahteve po razumljivosti enostavno integralne in posledično višje.

4 Študija primera

To nas privede do študije primera iz računalniške filozofije oz. filozofije, ki pri svojem raziskovanju temelji na računalniških metodah. Na podlagi tega primera bomo namreč lahko videli, zakaj je pomembno, da imamo vpogled ne le v to, kaj sledi na podlagi algoritmov, temveč da pri raziskovalni umetni inteligenci potrebujemo tudi razumevanje pogosto precej kompleksnih zalednih procesov. Pri tem se bomo kritično naslonili na nedavno objavljen članek o t. i. ekološki racionalnosti razlagalnega sklepanja (Douven 2020). Pri ekološki racionalnosti gre za oznako, da je pri presoji upravičenosti določenega tipa sklepanja treba upoštevati ne toliko, ali je nek način sklepanja smiseln kot tak, temveč ali je smiseln v nekem določenem okolju. Kot v primeru škarij je pri razumevanju mišljenja pomembno upoštevati dve stvari – mišljenje na eni strani (kot eno rezilo) in strukturo okolja na drugi (kot drugo rezilo metaforičnih Simonovih škarij; glej npr. Simon 1956).

Douven (2020) v svojem prispevku na podlagi t. i. agentske optimizacije (angl. *agent-based optimization*) oz. posebnega tipa računalniške simulacije pokaže, da je razlagalno mišljenje ekološko racionalno, saj lahko v določenem kontekstu privede do optimalnega izkupička. Pri tem avtor razlagalno mišljenje razume v verjetnostnem smislu kot adaptacijo bayesovskega učenja, ki posebej preferira najboljše razlage, ekološkost sklepanja pa skuša pokazati v simulaciji zdravniške diagnostike na primeru simulirane enote intenzivne nege.

Pri zasnovi izhaja iz psiholoških eksperimentov, ki kažejo, da se v opisnem smislu ljudje poslužujemo tega, da preferiramo najboljšo razlago oz. da najboljšo razlago smatramo kot (bolj verjetno) resnično od slabših razlag. Nato na podlagi skladnosti s psihološkimi raziskavami, predvsem pa na podlagi podatkov iz računalniških simulacij v agentski optimizaciji, skuša pokazati, da so standardni argumenti v podporo racionalnosti in normativni prednosti bayesovskega učenja pretirani. Njegova agentska optimizacija (računalniška simulacija in analiza simuliranih podatkov) namreč vodi do sklepa, da v primeru časovnih pritiskov, ki nastanejo npr. pri diagnostiki kritično poškodovanih bolnikov in bolnic, bayesovsko učenje evolucijsko odpade na račun bolj adaptivnih razlagalnih oz. sorodnih nebayesovskih principov sklepanja (Goodovo in Popprovo učenje). Pa je temu res tako oz. ali se pri argumentaciji res lahko tako enostavno zanesemo na rezultate razmeroma kompleksne računalniške simulacije?

Kot se izkaže, ko njegovo simulacijsko študijo nekoliko prilagodimo, rezultati niso pretirano robustni, saj lahko po manjši adaptaciji scenarija oz. zaledne kode pridemo do povsem neskladnih sklepov. Avtorjeva študija torej, tako bom vsaj trdil, zgolj slučajno podpira argumenta, da je razlagalno sklepanje ekološko racionalno in da bayesovskega učenja vsaj v kontekstu časovnih pritiskov ni, ker ni dovolj adaptivno.

Predpostavimo, da v simulacijo namesto povsem zanesljivih diagnostičnih testov (kot jih v svojih simulacijah uporablja Douven 2020) uvedemo teste, ki so varljivi. Ta sprememba je smiselna, saj se tudi sicer ne moremo povsem zanašati na svojo zaznavo oz. zunanje vire informacij, ki jih dobimo preko instrumentov ali pa na podlagi pričevanj. Rezultati v tem primeru vodijo v čudno situacijo, kjer različna razumevanja tega, kaj pomeni nezanesljivost testov, vodijo do rezultatov, ki jih lahko sprejmemo kot podporo v prid različnim načinom sklepanja. Pomembno pri tem je, da določena nezanesljivost testov pokaže celo, da bayesovsko sklepanje, torej način sklepanja, ki ima manj parametrov in bi zato moral biti manj adaptiven (če bi trditve iz osnovnega prispevka držale), lahko celo vodi do agentske optimizacije in evolucijske prevlade. Iz tega lahko (nasprotno kot sva s soavtorico najprej sklepala; glej Trpin in Plementaš 2021) razberemo, da je pri interpretaciji simulacij osnovnega članka (Douven 2020) prišlo do zmote, ker je avtor navidezno prepričljive rezultate interpretiral kot podporo svoji argumentaciji, dejansko pa je šlo za težavo na oblikovni ravni. Algoritem, ki je v ozadju, namreč ne kaže tega, kar bi pričakovali, tega pa ne opazimo, ker se nam rezultat zdi smiseln.

Za boljše razumevanje si velja ta primer pogledati bistveno bolj podrobno. V osnovi gre pri raziskavi za epistemološko dilemo. Obstaja namreč več pravil sklepanja, ki jih lahko uporabimo kot vodilo dobrega mišljenja. Ali je kakšen izmed (verjetnostnih) načinov sklepanja boljši od drugih?

Pri tem se moramo seveda najprej vprašati, kaj sploh je merilo dobrega mišljenja. Dve merili, za kateri se zdi, da ju velja upoštevati, sta sledeči: (i) resničnost prepričanj in (ii) hitrost snovanja prepričanj. Prvo merilo lahko upoštevamo tako, da pogledamo, v kolikšni meri določeno pravilo sklepanja vodi do resničnih prepričanj – več in močnejša prepričanja o resničnih propozicijah, kot jih imamo na podlagi določenega pravila sklepanja, bolje za pravilo sklepanja kot tako. Drugo merilo pa nam pomaga, ko ocenjujemo, kako hitro lahko zmanjšamo svojo negotovost – tem hitreje, tem bolje za pravilo.

V idealnih razmerah bi obe merili izpolnjevali hkrati, torej bi imeli pravilo sklepanja, ki hitro vodi do resničnih prepričanj. Rezultati iz literature kažejo, da to ni tako: pravila, ki so dobra po prvem merilu, so običajno slabša po drugem (ter obratno; Douven 2013; Trpin in Pellert 2019). Osnovno vprašanje pri filozofski oceni načinov (verjetnostnega) sklepanja je vezano na to, katero pravilo je najboljše v podpori merila (i) in merila (ii) ter kako lahko ta dva zaželena cilja (hitrost in natančnost sklepanja oz. mišljenja) spravimo v ravnovesje, sploh če upoštevamo, da viri informacij niso nujno povsem zanesljivi ali so nemara celo zavajajoči.

Če torej upoštevamo, da so podatki lahko nezanesljivi oz. nemara celo zavajajoči, moramo zasnovo računalniških simulacij, na katerih temelji Douvenov (2020) argument, spremeniti, saj sam te opcije ne upošteva. Pri tem sva se v nedavni analizi (Trpin in Plementaš 2021) oprla na analizo učenja iz t. i. delnih laži in različnih stopenj zaupanja (glej Trpin, Dobrosovestnova in Götzendorfer 2020), saj nam ta pristop prinaša formalizacijo zavajajočih in nezanesljivih virov informacij.

Prav tovrstni pomisleki – da bi lahko bil vir informacij varljiv podobno, kot so varljive delne laži – so vodili do adaptacije simulacij, ki jih je izvedel Douven (2020). Njegovo raziskavo lahko razdelimo v dva dela: najprej pokaže, da se pravila sklepanja (specifično: bayesovsko, razlagalno, Goodovo in Popprovo pravilo sklepanja) razhajajo glede omenjenih meril dobrega mišljenja (natančnost in hitrost). V drugem delu nato predlaga način, kako lahko ti dve merili uravnovesimo in preko selekcijske

optimizacije ugotovimo, katero je najboljše v določenem okolju (torej v smislu ekološke racionalnosti verjetnostnih pravil sklepanja).

Pri tem si je zamislil simulirano enoto intenzivne nege, v kateri zdravniki oz. zdravnice rešujejo paciente. Pri tem imajo tri možnosti: bodisi naredijo pravilno ali napačno odločitev glede posega, oz. če niso prepričani, kaj je narobe s pacientom oz. pacientko, ne naredijo nobenega posega. Pri tem se skozi čas spreminja verjetnost preživetja pacienta oz. pacientke, ki je odvisna tudi od posega. Kasneje kot pride do posega, manjša je verjetnost preživetja. Podobno pravilen poseg zviša verjetnost preživetja, napačen pa jo zniža. Če posega ni, je verjetnost preživetja vmes med pravilnim in napačnim posegom v določenem trenutku.

Douven preko simulacije, ki temelji na tem principu, pokaže, da je verjetnostno sklepanje na najboljšo razlago boljše pravilo sklepanja kot bayesovsko sklepanje. Čeprav gre za bolj tvegano pravilo (večja nevarnost zmote), je ravno zato tudi hitrejše in v primeru intenzivne enote prevlada v bitki med časom in natančnostjo. Drugače rečeno, čeprav glede natančnosti ni optimalno pravilo, je dovolj natančno, da zaradi hitrosti prevlada. Specifično v svojih simulacijah si je zamislil 200 zdravnikov oz. zdravnic. Od tega jih na začetku 50 sklepa na bayesovski, 50 na razlagalni, 50 na Popprov in 50 na Goodov način sklepanja. Vsak od njih ima 100 simuliranih pacientov in pri vsakem izvaja diagnostične teste, da ugotovi njihovo bolezen (100 testov). Ko na podlagi testov zasnuje prepričanje o bolezni, se odloči za poseg, ki je lahko pravilen ali nepravilen (oz. ga sploh ni, če bolezni ne more dovolj zanesljivo diagnosticirati). Na koncu preverimo, kolikšna je bila verjetnost preživetja pri posameznem zdravniku oz. zdravnici, in najboljših 100 od 200 podvojimo, ostale pa izbrišemo iz simulirane populacije. Ta postopek nato ponavljamo za 100 generacij zdravnikov oz. zdravnic in na tak način izvedemo t. i. agentsko optimizacijo.

Pri tem velja opozoriti, da simulacije temeljijo na tem, da so diagnostični testi povsem zanesljivi, čeprav v praksi temu ne bi bilo nujno tako. Zato se zdi smiselno, da kombiniramo uvide iz raziskav delnega laganja in delnega zaupanja ter jih uporabimo v analizi agentske optimizacije, ki poteka enako kot opisana s simuliranimi zdravniki oz. zdravnicami in pacienti oz. pacientkami. Razlika je torej ta, da so testi v spremenjenih simulacijah lahko nezanesljivi oz. delno zanesljivi in da zdravniki oz. zdravnice opazujejo, ali so testi sovpadali z opazovanimi simptomi ter na tak način kalibrirajo tudi svoje zaupanje v teste. Tak postopek vodi v

nepričakovane rezultate, ki na prvi pogled kažejo na to, da so različna pravila sklepanja bolj primerna za različna okolja (tako npr. Trpin in Plementaš 2021), dejansko pa na oblikovno pomanjkljivost opisanih simulacij, ki zaradi kompleksnosti dejansko ne prikažejo tega, kar naj bi – tj. premoči določenih pravil sklepanja.

Nasprotno, kot bi morda pričakovali, se po tej adaptaciji izkaže, da verjetnostno sklepanje na najboljšo razlago ni nujno najboljši način oz. da v agentski optimizaciji v določenih razmerah izpade. Poglejmo, denimo, rezultate simulacij, v katerih imamo teste, ki so stalno zavajajoči na način, ki je analogen enostavnemu laganju (test pokaže odsotnost nekega simptoma, če je bolezen taka, da je simptom bolj verjetno prisoten kot ne). V tem primeru testi za pacienta, ki ima bolj verjetno nek simptom kot ne (ker ima takšno simulirano bolezen), test vedno pokaže, da simptom ni prisoten. Zdravniki oz. zdravnice, ki po kalibraciji zaupanja in seznanitvi s takimi testi sklepajo na najboljšo razlago simptomov, skozi proces agentske optimizacije prevladajo. Do tukaj so torej rezultati skladni s tem, kar je v osnovni različici trdil Douven (2020).

A zgodba se s tem ne zaključi: če testi zavajajo na način, ki je podoben rokohitrskemu laganju (najprej nekdo testira, ali je simptom prisoten, in potem na testu pokaže nasprotno od dejanskega stanja), potem v teku optimizacije skozi generacije vseeno prevladajo razlagalni zdravniki oz. zdravnice, a ne s popolno prevlado – velik delež populacije optimalnih agentov namreč vsebuje tudi te, ki sklepajo na t. i. Goodov način sklepanja.

Najbolj zanimivi rezultati pa se pojavijo, če so testi zavajajoči na jasnovidni način: v kolikor je pri pacientu prisoten simptom, test pokaže, da simptoma ni. Po verifikaciji testa simulirani zdravniki oz. zdravnice nato kalibrirajo zaupanje v zanesljivost testa. Izkaže se, da v tem primeru skozi optimizacijo prevladajo bayesovski agenti, čeprav je bayesovsko pravilo sklepanja najmanj adaptivno (ima en parameter manj kot ostala tri simulirana pravila sklepanja (za več podrobnosti glej Trpin in Plementaš 2021).

Na podlagi teh rezultatov bi torej lahko sklepali, da nam simulacije predlagajo pluralističen pristop k pravilom sklepanja. V kolikor imamo opravka z nezanesljivostjo enega tipa, je bolj smiseln en, v kolikor z drugačno nezanesljivostjo testov pa drug način sklepanja. Tak bi bil vsaj sklep, če bi sledili načelom ekološke racionalnosti oz. tega, da je za oceno smiselnosti nekega načina sklepanja treba upoštevati skladnost z okoljem, v katerem se uporablja. Na podlagi tega bi se nato

lahko vprašali, kako prepoznati značilnosti okolja (oz. nezanesljivosti testov), da bi uporabili smiselno strategijo sklepanja oz. kdaj je smiselno preklapljati različne strategije.

A tak sklep bi bil preuranjen – kar nam pokažejo ti rezultati, je namreč ravno nasprotno: težava je v zasnovi samih računalniških simulacij. V osnovni izvedbi so namreč rezultati pokazali, da so v kontekstu, kjer ocenjujemo tudi hitrost sklepanja, bayesovsko sklepanje vedno izpodrinili drugi principi. To se je avtorju (Douven 2020) zdelo razumljivo, saj imajo ti drugi principi več parametrov, oziroma ti drugi principi lahko upoštevajo več vidikov situacije in so se zato zmožni bolje adaptirati. A kot pokažejo rezultati zgoraj omenjene adaptacije, temu ni nujno tako – v kolikor uporabimo nezanesljivost virov analogno z jasnovidnim laganjem, prevlada v situaciji bayesovsko sklepanje, čeprav je manj adaptivno.

Težava je torej v tem, da omenjena računalniška simulacija ne meri tega oz. ne prinaša podpore v prid argumentaciji, ki jo pričakujemo, ker pretirano temelji na tem, kako je zasnovana zanesljivost simuliranih testov. Prav v tem pa se pokaže tudi potreba po večjem standardu transparentnosti: avtor (Douven 2020) je namreč javno delil programsko kodo, ki je v ozadju omenjene študije. V kodi sicer ni težav, nam pa omogoča testiranje različnih variant (po adaptaciji) in posledično pretresanje, ki pokaže, ali so rezultati zanesljivi ali ne. Izkaže se torej, da koda na intencionalni ravni dela, kar naj bi (ni programskih težav), ne dela pa tega, kar njen avtor misli, da dela, saj je zasnovana na tak način, da rezultati temeljijo na vkodiranih domnevah glede zanesljivosti vira informacij. Tega pri običajnih znanstvenih raziskavah, kjer v igri ni metod umetne inteligence oz. vsaj simulacijskih metod v širšem smislu, ne zahtevamo oz. pričakujemo. Prav omenjeni rezultati pa nam kažejo, da to upravičeno zahtevamo od tako kompleksnih pristopov, kot je omenjena optimizacijska študija.

5 Zaključek

Čeprav smo se v pričujočem prispevku bolj poglobljeno posvetili enemu primeru, v katerem se pokaže, da je transparentnost v primeru raziskovalne umetne inteligence pomembna in da upravičeno pričakujemo tudi vpogled v to, kako so algoritmi v njenem ozadju zasnovani, velja izpostaviti, da je tovrstnih kritik več in da pri tem nismo prvi. Zgolj v filozofiji imamo npr. več primerov, kjer je pretresanje algoritmov v ozadju nekega argumenta privedlo do odmevnih kritik. Tak primer je npr. diskusija glede t. i. principa, da raznolikost prevlada nad zmožnostjo (skupine raznolikih ljudi

naj bi bile kognitivno bolj uspešne kot skupine bolj zmožnih, a manj raznolikih ljudi; glej Hong in Page 2004 za izvorni argument, ter npr. Thompson 2014 za eno od kritik) oz. glede epistemske delitve dela (glej Weisberg in Muldoon 2009 za izvorni argument ter Thoma 2015 za kritiko domnev iz njunega računalniškega modela). Podobno velja tudi onstran filozofije v znanosti širše (pri čemer lahko tudi že omenjeno diskusijo Honga in Pagea 2004, ter kritike njunega prispevka štejemo kot plod družbene vede in ne (samo) filozofije).

Problem torej, kot smo lahko videli, ni v tem, da raziskovalna umetna inteligenca vodi do novih spoznanj, ki jih sicer morda ne bi sami odkrili – to lahko kvečjemu pozdravimo. Zahtevati pa moramo visoko raven transparentnosti in vpogled v to, kako so zasnovani algoritmi, ki so v njenem zaledju. Le na tak način namreč lahko uvidimo, do kakšne mere so rezultati dejansko zanesljivi, pa čeprav imamo pri rezultatih, ki so nastali z 'naravno' inteligenco, nižje zahteve, ker vpogleda v drobovje pač ne moremo zahtevati. To do neke mere tudi pojasni, zakaj smo v znanosti tako rigidni glede zahteve po opisu metodologije. Zahtevamo namreč, da so rezultati zanesljivi in ponovljivi oz. v parafrazi ponarodele pesmi: za znanost je dobro le najboljše.

Viri in literatura

Andras, P., Esterle, L., Guckert, M., Han, T. A., Lewis, P. R., Milanovic, K., Payne, T., Perret, C., Pitt, J., Powers, S. T., Urquhart, N. in Wells, S. (2018). »Trusting intelligent machines: Deepening trust within socio-technical systems«. *IEEE Technology and Society Magazine*, 37(4), str.76–83.

Creel, K. A. (2020). »Transparency in complex computational systems«. *Philosophy of Science*, 87(4), str. 568–589.

Dennett, D. C. (1987). *The Intentional Stance*. Massachusetts: MIT Press.

Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L. in Kohane, I. S. (2019). »Adversarial attacks on medical machine learning«. *Science*, 363(6433), str. 1287–1289.

Grim, P. in Singer, D. (2020). »Computational philosophy«. V: Zalta, E. N. (ur.), *The Stanford Encyclopedia of Philosophy* (pomlad 2020). URL = https://plato.stanford.edu/archives/spr2020/entries/computational-philosophy/.

Hong, L. in Page, S. E. (2004). »Groups of diverse problem solvers can outperform groups of high-ability problem solvers«. *Proceedings of the National Academy of Sciences*, 101(46), 16385–16389.

Douven, I. (2013). »Inference to the Best Explanation, Dutch Books, and Inaccuracy Minimisation«. *Philosophical Quarterly*, 63, str. 428–444.

Douven, I. (2020). »The ecological rationality of explanatory reasoning«. *Studies in History and Philosophy of Science Part A*, str. 1–14.

Günther, M. in Kasirzadeh, A. (2021). »Algorithmic and human decision making: for a double standard of transparency«. *AI & SOCIETY*, str. 1–7.

Häse, F., Roch, L. M. in Aspuru-Guzik, A. (2019). »Next-generation experimentation with self-driving laboratories«. *Trends in Chemistry*, 1(3), str. 282–291.

Holozan, P. (2013) »Uporaba strojnega učenja za postavljanje vejic v slovenščini«. *Uporabna informatika*, 21(4), str. 196–209.

King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L. N., Sparkes, A., Whelan, K. E. in Clare, A. (2009). »The automation of science«. *Science*, 324(5923), str. 85–89.

Kleinberg, J., Ludwig, J., Mullainathan, S. in Rambachan, A. (2018). »Algorithmic Fairness«. *AEA Papers and Proceedings*, (108), str. 22-27.

Lipton, P. (2009). »Understanding without explanation«. V de Regt, H. W, Leonelli, S. in Eigner, K. (urd.), *Scientific Understanding: Philosophical Perspectives*. Pittsburgh: University of Pittsburgh Press, str. 43–63.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge, MA: MIT Press.

Melnikov, A. A., Nautrup, H. P., Krenn, M., Dunjko, V., Tiersch, M., Zeilinger, A. in Briegel, H. J. (2018). »Active learning machine learns to create new quantum experiments«. *Proceedings of the National Academy of Sciences*, 115(6), str. 1221–1226.

Nguyen, A., Yosinski, J. in Clune, J. (2015). »Deep neural networks are easily fooled: High confidence predictions for unrecognizable images«. V *Proceedings of the IEEE conference on computer vision and pattern recognition*, str. 427–436.

Pearl, J. in Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

Radovic, A., Williams, M., Rousseau, D., Kagan, M., Bonacorsi, D., Himmel, A. A., Terao, K., in Wongjirad, T. (2018). »Machine learning at the energy and intensity frontiers of particle physics«. *Nature*, 560(7716), str. 41–48.

Schupbach, J. N. (2018). »Robustness analysis as explanatory reasoning«. *The British Journal for the Philosophy of Science*, 69(1), str. 275–300.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel T. in Hassabis, D. (2017). »Mastering the game of Go without human knowledge«. *Nature*, 550(7676), str. 354–359.

Simon, H. A. (1956). »Rational choice and the structure of the environment«. *Psychological Review*, 63(2), 129–138.

Thoma, J. (2015). »The epistemic division of labor revisited«. *Philosophy of Science*, *82*(3), str. 454–472.

Thompson, A. (2014). »Does Diversity Trump Ability?« *Notices of the AMS*, 61(9), str. 1024–1030.

Trpin, B. in Pellert, M. (2019). »Inference to the best explanation in uncertain evidential situations«. *The British Journal for the Philosophy of Science*, 70(4), str. 977–1001.

Trpin, B., Dobrosovestnova, A. in Götzendorfer, S. J. (2020). »Lying, more or less: a computer simulation study of graded lies and trust dynamics«. *Synthese*, 199, 991–1018.

Trpin, B. in Plementaš, A. M. (2021). »The ecological rationality of probabilistic learning rules in unreliable circumstances«. V Strle, T., Trpin, B., Rebernik, M. in Markič, O. (urd), *Zbornik 24. mednarodne multikonference Informacijska družba: Zvezek B - Kognitivna znanost*, str. 51–55.

Vigen, T. (2015). *Spurious Correlations*. Hachette Books, New York in Boston.

Weisberg, M. in Muldoon, R. (2009). »Epistemic Landscapes and the Division of Cognitive Labor«. *Philosophy of Science*, 76(2), str. 225–52.

Zerilli, J., Knott, A., Maclaurin, J. in Gavaghan, C. (2019). »Transparency in algorithmic and human decision-making: Is there a double standard?« *Philosophy & Technology*, 32(4), str. 661–683.

ALI NAS UMETNA INTELIGENCA LAHKO PREMAGA: OD ALGORITMA DO SINGULARNOSTI PO POTEH ETIČNEGA VREDNOTENJA

BOJAN BORSTNER, NIKO ŠETAR
Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
bojan.borstner@um.si, niko.setar1@um.si

Sinopsis Namen tega članka je ponuditi vpogled v splošne etične težave, s katerimi se sooča teorija umetne inteligence. Ker omenjena tematika pogosto naleti na kritike, da gre za znanstveno fantastiko in je razpravljanje o etiki kakršnekoli višje umetne inteligence nesmiselno, bomo v prvem delu članka opredelili umetno inteligenco in predstavili kratek argument, zakaj je ukvarjanje s človeku podobno umetno inteligenco in celo superinteligenco smiselno v okviru predvidene verjetnosti njunega obstoja v prihodnosti. Večinski del članka se bo nato obrnil k etiki umetne inteligence, začenši s problematiko algoritemske diskriminacije in nižjih umetnih inteligenc, kot so denimo samovozeča vozila. Po tem bomo pregledali etične aspekte nadaljnjega razvoja umetne inteligence do človeku podobne inteligence in superinteligence. Med tem pregledom se bomo ustavili pri nekaj možnih rešitvah za obravnavane dileme, proti koncu pa predlagali, da je zaradi narave strojnega učenja v umetni inteligenci treba iskati etične rešitve v okviru teorije utemeljevanja.

Ključne besede:
umetna inteligenca,
algoritemska diskriminacija,
samovozeča vozila,
superinteligenca,
etika,
teorija utemeljevanja

Univerzitetna založba Univerze v Mariboru

DOI https://doi.org/10.18690/um.ff.11.2022.5
ISBN 978-961-286-675-4

CAN ARTIFICIAL INTELLIGENCE DEFEAT US: FOLLOWING THE PATH OF ETHICAL EVALUATION FROM ALGORITHM TO SINGULARITY

BOJAN BORSTNER, NIKO ŠETAR
University of Maribor, Faculty of Arts, Maribor, Slovenia
bojan.borstner@um.si, niko.setar1@um.si

Abstract The aim of this article is to present an insight into general ethical issues pestering the field of theory of artificial intelligence. Seen as this topic is often the target of criticisms reducing it to science fiction, claiming that any consideration of higher artificial intelligence ethics is nonsense, we will use the first part of this article to define artificial intelligence and shortly argue why discussing human-like AI and even superintelligence makes sense given expert predictions about their future existence. The main part of this article then turns to artificial intelligence ethics, first dealing with algorithmic discrimination and issues with lower artificial intelligences such as automated vehicles. Afterwards, we overview ethical aspects of further AI development to the point of human-like artificial intelligence and superintelligence. In doing so, we shall examine some possible solution for dilemmas at hand, finally suggesting that the nature of machine learning in artificial intelligence requires pursuit of ethical solutions within the grounding theory.

Keywords: artificial intelligence, algorithmic discrimination, automated vehicles, superintelligence, ethics, grounding theory

https://doi.org/10.18690/um.ff.11.2022.5 DOI
978-961-286-675-4 ISBN
University of Maribor Press

1 Uvod

Že v zgodnjih fazah razvoja računalniške znanosti se pojavlja ideja tako imenovanega 'mislečega računalnika', ki bi bil zmožen opravljati vse, kar lahko opravlja tudi človek. Prvi konkretni opis delovanja tovrstnega računalnika je 'igra posnemanja', ki jo je predlagal znameniti pionir računalništva, Alan Turing, v kateri nastopajo trije udeleženci: človek, računalnik in izpraševalec. Slednji je tudi sam človek, ne ve pa, kaj sta druga dva udeleženca v igri (nikoli ju namreč ne vidi, z njima se sporazumeva na daljavo). Izpraševalec poskuša s serijo vprašanj ugotoviti, kateri izmed drugih dveh udeležencev je dejansko računalnik. Če mu to spodleti, je računalnik zmagal igro in lahko rečemo, da je inteligenten (Dobrev 2005).

Žal je Turing po mnenju mnogih podcenjeval, kako kompleksno je lahko delovanje in vedenje sodobnega računalnika, ne da bi mu lahko pripisali inteligenco v pravem pomenu besede. Obstajajo očitno neumni umetni akterji, kot denimo lažni Facebook profili, ki so zmožni predelati le nekaj osnovnih fraz in bodo na vprašanje »Kaj si jedel za zajtrk?« odgovorili z eno izmed njih, denimo »V redu, pa ti?« ter s tem jasno pokazali, da niso ljudje. Po drugi strani pa lahko zapleten algoritem, ki se uči iz preteklih pogovorov in podobnih virov, odgovori »Nisem zajtrkoval.« ali pa poda možen, a neresničen odgovor in reče, da je jedel kruh s pašteto in otrpne šele pri zelo podrobnih in zapletenih vprašanjih, do katerih Turingov izpraševalec morda sploh ne bi prišel. Dejstvo, da vprašanja, ob katerih bi naš umetni akter otrpnil in izjavil nekaj nesmiselnega, sploh obstajajo, pa pomeni, da vseeno ni inteligenten. Zelo visoko razvita, človeku podobna umetna inteligenca pa bi utegnila tudi lagati, kar pomeni podati neresnične izjave, za katere se zaveda, da so neresnične.

2 O prihodnosti umetne inteligence

2.1 Kaj je umetna inteligenca?

Ena od možnih definicij pravi, da je to »takšno vedenje naprave, ki bi bilo vzeto za inteligentno, ko bi se tako vedel človek« (Simmons in Chappell 1988: 18), Druga, dokaj podobna definicija, je lahko tudi, da je umetna inteligenca »takšen program, ki se s poljubnim svetom ne bo soočal nič slabše kot človek« (Dobrev 2005: 70). Pri obeh definicijah je osnovna ideja enaka kot pri Turingovem testu: umetna inteligenca ne sme biti razločljiva od človeške, le da sta zgornji definiciji zahtevnejši v smislu, da

umetna inteligenca ni nerazločljiva od človeka le v omejeni seriji vprašanj, marveč v njenem celotnem vedenju, razmišljanju in interakciji s svetom. Kot je razvidno iz definicije Dobreva, lahko umetna inteligenca človeka v teh aspektih tudi preseže. Bergstrom (2014) nadalje trdi, da mora biti umetna inteligenca ne-neumna v človeški komunikaciji, kar pomeni, da mora biti sposobna razumeti in se odzvati tudi na inherentno človeške elemente komunikacije, kot sta denimo sarkazem in ironija.

Trenutni umetni kognitivni sistemi so še precej daleč od tega – zdi se, da še zmeraj delujejo po principu Kitajske sobe (Searle 1980), se pravi, da obravnavajo podatke glede na njihovo obliko, vzorce ipd., ne da bi se zavedali pomena podatkov, ki jih obravnavajo. V nadaljevanju tega članka bomo na kratko pregledali prihodnji razvoj umetne inteligence in obstoječe ovire v razvoju, ocene verjetnosti postopnega nastanka resnične človeku-podobne umetne inteligence, nato pa prešli k etičnim težavam, s katerimi se sooča in s katerimi se bo še v prihodnje soočalo področje razvoja umetne inteligence.

2.2 Razvoj umetne inteligence

Trenutni razvoj umetne inteligence se osredotoča predvsem na nevronska omrežja, ki naj bi modelirala človeške možgane, pri čemer se nevronske povezave obravnavajo kot omrežje logičnih vrat. Umetni akter nato, poenostavljeno povedano, išče tisto zaporedje logičnih vrat, ki vodi do nekega želenega izida v okviru njegove naloge. Že v osnovi se pojavljajo bistvene razlike v delovanju umetnih nevronskih omrežij in človeških možganov – hitrost širjenja signalov v nevronskih omrežjih je občutno večja kot hitrost širjenja signalov v možganih, a je njihova moč obdelave podatkov, npr. več različnih podatkov sočasno, razpon vrste podatkov, ki jih lahko obdelujejo itd., mnogo manjša. Tako je šahovski robot mnogo boljši v šahu od ljudi, tudi od velemojstrov, a je njegova funkcija omejena izključno na šah, medtem ko je lahko nek šahovski velemojster hkrati tudi prav dober filozof ali pa slikar. (Brooks et al. 2012; Bostrom in Yudkowsky 2011)

Mnogi pripisujejo te razlike raznovrstnosti podstati, na katerih delujeta umetno in naravno nevronsko omrežje. Slednje deluje na organski, prvo pa na neorganski podlagi. V osnovi gre za vprašanje fizikalizma: ali smo ljudje skupaj z našo zavestjo in vsemi mentalnimi stanji samo serija nevronskih povezav? Trde znanosti se trenutno nagibajo k pritrdilnemu odgovoru na to vprašanje, pri čemer pa do velike

mere zanemarjajo ali poenostavljajo problem zavesti – kako fizikalna podlaga skupaj s fizikalnimi dražljaji vodi v zavestno (fenomenalno) izkustvo? (Chalmers 1995; gl. tudi Tye 1995) To vprašanje trenutno (še) ni odgovorjeno. Marsikdo bi utegnil trditi, da dobro vemo, kateri centri v možganih se sprožijo ob določenih izkustvih, vendar pa to še ne pojasni, kako natanko iz te 'sproženosti' vznikne mentalno stanje, kot ga dejansko občutimo. Če se kljub temu podpišemo pod strogi fizikalizem in vztrajamo, da je samo vprašanje časa, preden se ta povezava do potankosti razloži, potem je razvoj umetne inteligence do nivoja človeške inteligence prav tako le vprašanje časa. Če pa na drugi strani vztrajamo, da te povezave ni mogoče najti, ker je ni, in je za mentalna stanja in zavest 'kriva' neka ne-fizikalna substanca, potem je razvoj tovrstne umetne inteligence nemogoč. Vsaj tako se sprva zdi.

Longinotti (2017) na primer trdi, da so zavest in z njo povezana stanja nekaj intrinzično biološkega, kar lahko tudi v fizikalističnem okviru obstaja samo na podlagi organske podstati. Avtor ugovarja komputacionalističnim pogledom na človeško nevrološko strukturo na podlagi argumenta, da ti pogledi kršijo princip lokalnosti v fiziki. Princip lokalnosti pravi, da lahko ima nek faktor (večinoma delec) A vpliv na B, če in samo če ima A neposreden stik z B, pri čemer je hitrost širjenja A proti B omejena s svetlobno hitrostjo. Komputacionalizem naj bi po drugi strani vodil v nerazložljivo vrzel med nekim fizikalnim vzročnim vzorcem A in fenomenalnim stanjem B. Longinotti v odgovoru predpostavlja, da so tako neka do sedaj neopisana oblika energije, ki povezuje A in B.

Na drugem polu najdemo hipotezo neodvisnosti od podstati, ki predpostavlja, da je veljavnost fizikalizma nepomembna za izgradnjo podstati, ki lahko nosi zavest. V okviru te hipoteze je mogoče, da lahko z izgradnjo umetne podstati, ki je zadostno podrobno (angl. *sufficiently fine-grained*) podobna že dokazano uspešno delujoči, tj. človeški podstati, simuliramo mentalna stanja, ki so zadostno podrobno podobna človeškim mentalnim stanjem, da jih lahko smatramo za mentalna stanja oz. zavest (Bostrom 2003).

Katerakoli izmed teh možnosti naj bo veljavna, se fizikalistično vprašanje oz. odgovori nanj nanašajo le na stopnjo 'človeškosti' umetne inteligence, ki jo lahko dosežemo, in popolnoma predstavljivo je, da lahko na neki točki dosežemo tudi umetno inteligenco, ki je neskončno boljša od človeka v vseh racionalnih in logičnih

operacijah hkrati, četudi nima pojavnih izkustev in mentalnih stanj v človeškem pomenu izraza.

2.3 Verjetnost razvoja višje umetne inteligence

Mnogi izmed opisanih scenarijev zvenijo kot popolna znanstvena fantastika. S trivialnega stališča si lahko ogledamo pretekle primere, ko skeptične napovedi o razvoju določenih tehnologij enostavno niso zdržale. Roman Julesa Verna *Dvajset tisoč milj pod morjem,* izvirno izdan leta 1870, je veljal za znanstveno fantastiko, saj naj bi močno precenjeval zmožnosti razvoja podmornic, kljub obstoju zgodnje francoske podmornice *Plongeur,* razvite leta 1864. Že leta 1938 so inovacije na področju pogona podmornic slednjim nudile teoretično skoraj neomejen čas potovanja pod vodo. Podobno so le nekaj let pred pionirskim poletom bratov Wright mnogi strokovnjaki na področju aviacije predvidevali, da so leteče naprave težje od zraka popolnoma nemogoč koncept.

V popolnem nasprotju z zgornjimi napovedmi pa se tudi Bentley in drugi (2018) sklicujejo na zmotnost preteklih znanstvenih napovedi pri tem, da trdijo, da človeku-podobne umetne inteligence ter superinteligence nikoli ne bodo obstajale. Ideje, da se lahko umetna inteligenca sama uči in postane superinteligentna, da lahko sploh prekosi človeško inteligenco, označujejo kot 'mite' o umetni inteligenci. Pri tem se opirajo na tri osnovne zakone umetne inteligence, ki ovržejo te mite. Prvi zakon umetne inteligence je, da inteligenca izhaja iz soočanja z izzivi – ko umetnemu akterju predstavimo izziv na določenem področju, se ga bo po standardnih učnih metodah naučil premagati. Bentley et al. trdijo, da zaradi tega samoučeča umetna inteligenca nikdar ne bo dosegla nivoja superinteligence, saj je skoraj nemogoče doseči nivo, ko bi se umetna inteligenca soočila z izzivom, ki bi od nje zahteval superinteligentnost. Drugi zakon trdi, da inteligenca zahteva primerno strukturo – ta zakon naj bi preprečeval razvoj umetne inteligence na ali nad nivo človeške inteligence, saj različne funkcije zahtevajo različne nevronske strukture, zaradi česar naj bi bilo zahtevano umetno nevronsko omrežje prezapleteno, da bi ga bilo mogoče praktično ustvariti. Tretji zakon je, da umetna inteligenca zahteva obilo testiranj. To naj bi preprečevalo, da bi umetna inteligenca lahko izrabila tehnološki napredek na področju procesne hitrosti računalnikov in se razvijala vzporedno s tem napredkom, saj bi testiranje novih funkcij po enem 'skoku' v napredku trajalo dlje, kot bi minilo časa do naslednjega možnega tovrstnega skoka.

Pa vendar je skepticizem, ki ga izražajo Bentley et al. prav toliko osnovan na špekulaciji kot nasprotne napovedi, ki govorijo v prid razvoja višjih umetnih inteligenc. Sklepanje, da izziv, ki bi vodil v superinteligenco, ne bo nikdar obstajal, ni osnovano na ničemer otipljivem ter ne nudi nobene stopnje gotovosti. Prav tako je brez konkretne podlage trditev, da nikoli ne bomo zmožni ustvariti dovolj zapletenega nevronskega omrežja, kot tudi predpostavka, da bo testiranje vsakršnega napredka v umetni inteligenci sploh vedno potrebno.

Drug skepticizem glede razvoja človeku-podobne umetne inteligence je pogosto osnovan na ideji, da so mentalna stanja nekaj inherentno človeškega. Zagovarjanje te ideje, kljub znanstvenim indikatorjem, da najverjetneje ni resnična, nemalokdaj temelji na eksistencialni grožnji, ki jo možnost obstoja zavestnih ne-človeških bitij, tj. živih, razmišljujočih umetnih akterjev, predstavlja za ljudi in njihov status v svetu ('Roko' 2010). V strokovnih krogih je mnenje precej drugačno. Anketa, ki sta jo na vzorcu nekaj sto izvedencev na področju umetne inteligence izvedla Müller in Bostrom (2016), nakazuje, da približno petdeset odstotkov strokovne javnosti na tem področju verjame, da bo človeku-podobna umetna inteligenca uspešno razvita v naslednjih dvajsetih letih, devetdeset odstotkov (vključujoč omenjenih petdeset) pa, da bo ta nivo razvoja dosežen najkasneje do leta 2075.

Naslednji korak v razvoju je umetna superinteligenca ali singularna inteligenca, ki človeka presega v vseh funkcijah, saj ima zmožnost opravljanja skoraj neskončnega števila nalog sočasno. Takšna umetna inteligenca temelji na kvantnem računalništvu, za katerega Drexler (1992) opisuje, da bi lahko bil kvantni računalnik velikosti 'kocke sladkorja' sposoben opraviti 10^{21} operacij na sekundo, medtem ko Lloyd (2000) trdi, da bi lahko bil tovrsten računalnik z maso enega kilograma sposoben opraviti kar $5x10^{50}$ operacij na sekundo. V primerjavi s tem so človeški možgani sposobni opravljanja okoli 10^{14} operacij na sekundo. Približno petinsedemdeset odstotkov strokovnjakov (po Müller in Bostrom 2016) meni, da bo umetna superinteligenca sledila človeku-podobni inteligenci v največ tridesetih letih od njenega razvoja, se pravi, da bo umetna superinteligenca postala realnost do konca tekočega stoletja.

3 Etična vprašanja umetne inteligence

Ko smo na kratko razjasnili razvojne možnosti in verjetnost obstoja višjih umetnih inteligenc, se lahko obrnemo k jedru tega članka: etičnim težavam, s katerimi se umetna inteligenca sooča. Allen in drugi (2005) pišejo, da problematika izhaja iz prognoze, da bo hitrost operacij, ki jih izvaja umetna inteligenca, kmalu onemogočila človeško posredovanje in etično obravnavo posameznih dejanj, zatorej je potrebno razviti metodo, s katero bo umetna inteligenca lahko sama etično presojala svoja dejanja. Allen in drugi predstavijo tri možne pristope, s katerimi bi to utegnilo biti mogoče. Prva skupina pristopov so pristopi 'od zgoraj dol', pri čemer gre za eksplicitno vgrajevanje nekaterih etičnih načel v programsko osnovo umetne inteligence, kot jih je predlagal denimo Isaac Asimov (gl. *Runaround,* 1950). Slednje lahko povzamemo na sledeč način:

1. Robot ne sme škodovati človeku ali dopustiti škodovanja človeku.
2. Robot mora ubogati človeške ukaze, razen če to krši Prvi zakon.
3. Robot mora ščititi lasten obstoj, razen če to krši Prvi ali Drugi zakon.

Primarna kandidata za ta pristop sta seveda vodilni normativni etični teoriji, utilitarizem in deontologija. Utilitarizem se zdi predvsem primeren zaradi svoje inherentne težnje, da kvantitativno vrednoti izide dejanj, kar olajša njegovo implementacijo v program umetne inteligence, a se sooča s številnimi drugimi težavami, kot je na primer nesoizmerljivost človeških življenj, kar bomo poudarili v razdelku o samovozečih vozilih. Deontološka teorija je skladnejša z Asimovimi principi, saj se opira na princip pravila tako, kot je ta definiran v Kantovi prvi maksimi etičnega delovanja. Teorija pa je težko združljiva s komputacionalističnimi pristopi k programiranju umetne inteligence, zato je praktično težje vpeljiva. Arnold in Scheutz (2018) denimo predlagata sistem etike v umetni inteligenci, ki na podlagi osnovnega operacijskega sistema zgradi etično jedro, na vrh katerega nato gradi kompleksnejši operacijski sistem in na koncu inteligentni algoritem. Pri tem se vzdržita opredelitve, katera normativna etika bi sestavljala etično jedro. Druga skupina pristopov, 'od spodaj gor', predvideva, da lahko umetne inteligence preko mehanskega učenja, spoznavanja okolja in preučevanja medčloveških odnosov same 'proizvedejo' etična načela, po katerih se lahko kasneje ravnajo; tretja kategorija so hibridni pristopi, ki predvidevajo delno vgrajenost osnovnih etičnih načel in

izgradnjo podrobnejše, bolj specifične morale na podlagi teh vgrajenih načel (Allen et al. 2005).

V nadaljevanju bomo pokazali, zakaj so pristopi 'od zgoraj dol' uporabni v kontekstu nižjih umetnih inteligenc, kot so denimo učeči-se algoritmi, samovozeča vozila itd., medtem ko so pristopi 'od spodaj gor' nujni za razvoj višjih umetnih inteligenc. Predvidevamo tudi uporabnost 'invertiranega' hibridnega pristopa, tj. pristopa 'od spodaj gor', v okviru katerega pa lahko umetno inteligenco neposredno učimo kompleksnejše etike po tem, ko sama utemelji osnovna etična načela.

Tudi tukaj se moramo najprej ozreti k osnovnim oblikam umetne inteligence, začenši z avtomatiziranimi algoritmi, kot so denimo algoritmi, ki se uporabljajo za kriminalno in psihološko profiliranje, avtomatsko odobravanje kreditov ipd.

3.1 Nižje umetne inteligence

Glavna težava tovrstnih algoritmov je algoritemska diskriminacija, do katere pride zaradi pasivnega človeškega faktorja, se pravi implicitnih predsodkov družbe ali ljudi, ki jih programirajo. V osnovi obstajata dva vzroka za algoritemsko diskriminacijo: pristranski učni podatki in neenakopravna temeljna resnica. Pri slednji gre za obliko statistične diskriminacije, ki je uperjena proti neki demografski skupini na podlagi nesorazmernosti statističnih podatkov: algoritem za kriminalno profiliranje (večinoma v ZDA) bo na primer prej označil temnopoltega kot belega osumljenca na podlagi statistike, ki pravi, da Afroameričani zagrešijo več kriminalnih dejanj kot belci, pri čemer pa algoritem ne vzame v obzir socioekonomskih faktorjev, ki privedejo do kriminala, ter dejstva, da se belci pogosteje izmuznejo kazni ali dobijo milejšo kazen. Pristranski učni podatki so lahko posledica posredne ali neposredne pristranskosti programerjev ali naročnikov algoritma, ki vnesejo neustrezne učne podatke (Hacker 2017).

To naj ilustriramo na primeru s poljubnimi številkami: denimo, da gre za algoritem, ki odobrava kredite. Naša izmišljena statistika pravi, da ženske vzamejo 10 % kreditov, moški pa 90 % vseh kreditov. Pri tem 80 % žensk odplača kredit pravočasno, tako tudi 60 % moških. Na podlagi prvega dela statistike programer vnese v učni vzorec 10 žensk in 90 moških, drugi parameter pa ni eksplicitno vnešen, kar bo vodilo v več odobritev kreditov moškim kot ženskam, čeprav se ponudniku

bolj izplača odobriti kredit ženskam kot moškim. Algoritem je tako diskriminatoren do ženskih prosilk za kredit, poleg tega pa tudi neučinkovit. Ho (2019) poudarja, da do tovrstnih težav pride tudi pri algoritmih, ki so bolj neposredno povezani z življenjsko nevarnimi situacijami, kot so algoritmi, namenjeni diagnosticiranju bolezni. Tovrstni algoritmi lahko zanemarijo določene pridružene bolezni ali bolezenska stanja, slabo upoštevajo ali ne upoštevajo faktorja starosti itd. Ho pravi, da je v iskanju rešitve potrebno objektivno upoštevati vse klinične, socialne, etične in relacijske faktorje.

Rahwan (2017) predlaga splošno rešitev, ki po klasifikaciji Allena in drugih (2005) sodi v kategorijo 'od zgoraj dol'. Ta rešitev vsebuje tri komponente: vpletenost človeškega faktorja (angl. *human-in-the-loop*; HITL), vpletenost družbe (angl. *society-in-the-loop*; SITL) in družbeno pogodbo. HITL predvideva človeškega operaterja, ki neposredno posega v ravnanje algoritma, kadar se ta sooča s podatki, ki jih ne more ustrezno obdelati, kot so izjeme, potrebe po nadgradnjah ali spremembah delovanja algoritma ipd. Kadar se principu HITL dodajo parametri človeške družbene pogodbe,[1] dobimo SITL, tj. algoritem, ki uspešno vzame v zakup interese vseh interesnih skupin, vpletenih v delovanje algoritma oz. v družbeni proces ali institucijo, v okviru katere algoritem deluje.

Z algoritmi povezane težave najdemo tudi v višje razvitih umetnih inteligencah, kot npr. te, ki vodijo avtomatizirana vozila. Usposobljenost avtomatiziranih vozil za vožnjo v normalnih okoliščinah je že domala dovršena, medtem ko se še zmeraj pojavljajo dileme glede njihovega ravnanja v izrednih okoliščinah. Kako naj takšno vozilo ravna, ko se ni mogoče izogniti nesreči, predvsem če algoritem predvideva neizogiben smrtni izid za nekoga od udeležencev?

Na zdravorazumskem nivoju je problem, kako oceniti, kdo ima večje možnosti za preživetje ob trku, torej komu se izogniti in v koga trčiti, če je trk v eno ali drugo vozilo ali osebo neizogiben? Globlji problem je, kako oceniti čigavo življenje ohraniti in čigavo žrtvovati, če algoritem oceni, da je smrt vsaj ene osebe neizogibna? De Sio (2017) pri slednjem vprašanju vidi težavo predvsem v nesoizmerljivosti človeškega

[1] Rahwan sam priznava, da je družbeno pogodbo izredno težko ustrezno definirati. Razlogov zato je mnogo: družbena pogodba se lahko bistveno razlikuje v različnih kulturah; družbeno pogodbo ljudje sprejemamo intuitivno, najboljši formaliziran približek pa je zakonodaja. Poleg tega se moramo spopasti tudi s tem, kako družbeno pogodbo, če jo uspešno enoznačno definiramo, prevesti v programski jezik na način, da jo bo umetni akter razumel in upošteval.

življenja, tj. da ni nikakršnega splošnega merila, po katerem bi lahko ocenili vrednost enega življenja kot drugačno od vrednosti drugega, ki ne bi bilo diskriminatorno. Poleg tega težavo predstavlja tudi kulturni relativizem – v Evropi je denimo življenje otroka zaradi neizkoriščenega potenciala tradicionalno vredno več od življenja starostnika. Po drugi strani je v mnogih vzhodnih kulturah zadeva obratna, saj je starostnik deležen določenega spoštovanja in 'prednostne obravnave' zaradi svojih življenjskih izkušenj in zaslug za življenjske dosežke.

Tretji problem, ki ga želimo izpostaviti na področju trenutno obstoječih umetnih inteligenc, so avtomatizirani vojaški sistemi in orožja, kot so brezpilotna letala in droni. Delno avtomatizirani sistemi, kot so droni, ki so vodeni ali nadzorovani na daljavo, so pogosto predmet kritike na podlagi kršitve človeškega dostojanstva njihovih tarč, saj je takšen način vojskovanja s strani uporabnika drona popolnoma neoseben. Po drugi strani obstajajo argumenti v podporo delno-avtomatiziranim vojaškim sistemom, ki trdijo, da v primeru neizogibnega konflikta uporaba teh sistemov prepreči dodatne žrtve na vojskujoči strani, ki jih uporablja (Statman 2015).

Kljub temu pa tega argumenta ni mogoče prenesti na popolnoma avtomatizirane vojaške sisteme, ki jih utegne pestiti isti problem kot avtomatizirana vozila – kako naj se odzovejo v nenavadnih okoliščinah. Težava je odpravljena, če obe strani spopada uporabljata izključno t. i. vojaške robote. V kolikor pa je ena izmed vojskujočih frakcij opremljena le s človeškimi vojaki, pa lahko ti, z novimi strategijami in nepredvidljivimi načini vojskovanja, robota 'prisilijo' v nepredvidljivo reakcijo, ki se lahko konča v nenačrtovanih in nepotrebnih smrtnih žrtvah (Swoboda 2017).

Tudi tukaj, v okviru vseh relevantnih avtomatiziranih sistemov, je mogoče vpeljati pristope, omenjene zgoraj (Allen et al. 2005); na prvi pogled se zdi, da bi bili sistemi grajenja etike od spodaj gor neverjetno nevarni, a se pri tem ne predpostavlja učenje na terenu, marveč učenje na poligonu ali v simulaciji, kjer je mogoče poustvariti tako normalne kot anormalne okoliščine (Glej Bonnemains et al. 2015; Wagner in Koopman 2015).

Ti primeri so relevantni za nadaljnjo debato zato, ker kažejo na pomembnost izhodiščnega programiranja in podrobnega, izčrpnega učnega postopka v razvoju posamezne umetne inteligence, a več o tem proti koncu tega prispevka.

3.2 Višje umetne inteligence

Avtonomne umetne inteligence zaradi njihove izjemne sposobnosti opravljanja določenih nalog predstavljajo grožnjo, da bodo sčasoma nadomestile večino človeškega dela. Recimo temu prvi nivo eksistencialne grožnje. Kakšna je vloga ljudi v svetu, kjer umetne inteligence opravljajo človeško delo? Acemoglu in Restrepo (2018) trdita, da je tovrsten alarmističen pogled neutemeljen, saj pretekli trendi kažejo, da avtomatizaciji določenih delovnih mest sledi odprtje novih delovnih mest na drugačnih področjih dela, a v isti skupni kapaciteti. Ostajajoča težava je pravočasno izobraževanje kadra za nova delovna mesta, kar pa ne predstavlja posebnega problemskega sklopa v okviru etike umetne inteligence. Tudi če sprejmemo alarmističen pogled in predvidevamo, da je možno, da bo umetna inteligenca nadomestila ljudi na vseh rutinsko, postopkovno, statistično in analitično orientiranih delovnih mestih, ostajajo delovna mesta, kjer je človeški stik nujen (psihologija, turizem ipd.), kot tudi področja, kjer umetna inteligenca (vsaj pod človeškim nivojem inteligence) ni zmožna delovati – denimo umetnost, glasba in navsezadnje filozofija. Teh delovnih mest je sicer razmeroma malo, tako da vprašanje, s čim se bo preživljala večina, ostaja. Autor (2015) in Akst (2013) vsak na svoj način stremita k istemu zaključku: v svetu, kjer umetna inteligenca opravlja večino dela, drži dejstvo, da umetna inteligenca ustvarja dobrine in dobiček z zanemarljivimi stroški. Tisto, kar bi potrebovali v takem svetu, je način distribucije dobrin in dobička, ki ne izvzema posameznikov, ki so v tem 'novem svetu' nezaposleni zaradi neobstoja služb, kar pa je v trenutnem sistemu nepredstavljivo. Problem je torej socialnoekonomski, rešitev pa zahteva revizijo obstoječega sistema distribucije virov in dobrin. Prva eksistencialna grožnja tako dejansko ni grožnja umetne inteligence človeštvu, marveč grožnja človeštva samemu sebi v kontekstu obstoja umetne inteligence.

Naslednje vprašanje se nanaša na obstoj človeku-podobne umetne inteligence, ki je bolj zmogljiva od človeka na vseh zgoraj omenjenih področjih, ter enako zmogljiva kot človek na področjih, ki smo jih v analizi alarmističnega pogleda smatrali za ostajajoče človeška v primeru avtomatizacije z 'navadno' umetno inteligenco. Prva eksistencialna grožnja se na tem nivoju nekoliko poglobi, a hkrati ostaja domena socialnoekonomskih prepričanj. Novonastali problem ni dodatna nova grožnja človeštvu, marveč problem statusa človeku-podobnih umetnih inteligenc v družbi.

Umetne inteligence na človekovem nivoju bi utegnile, v primeru polne veljavnosti fizikalizma ali vsaj hipoteze o neodvisnosti podstati, imeti človeku-podobna čustva in občutke, ali pa bi se vsaj na nek način zavedale njihovega družbenega statusa in tega, kako jih ljudje dojemajo in obravnavajo. Mishra (2017) trdi, da bi bil moralni in družbeni status umetnih inteligenc odvisen od tega, v katero izmed štirih kategorij kognitivnih zmožnosti sodijo. Prva kategorija zahteva prefinjeno (človeško) kognitivno zmožnost, druga prefinjeno zmožnost v razvoju, tretja predvideva poseben odnos med človekom in umetno inteligenco (podobno kot med človekom in domačimi živalmi), četrta pa osnovne kognitivne zmožnosti. Umetne inteligence, o katerih je govora v tem odstavku, sodijo v prvo ali najmanj drugo kategorijo, kar zahteva, da jih obravnavamo kot človeku enake v moralnem in družbenem kontekstu. Uporaba teh inteligenc zgolj kot sredstev za opravljanje določenega dela bi pomenila, da jim odvzamemo status oseb (ki ga v primeru človeškega nivoja inteligence dejansko imajo) in jih obravnavamo kot sužnje. Prav tako bi bil izklop tovrstne umetne inteligence moralno ekvivalenten umoru. Beckers (2017) predvideva, da bi bilo obravnavanje umetno ustvarjenih oseb kot človeku enakih za večino ljudi nekaj nepredstavljivega ali nespremenljivega, zaradi česar potencialnim umetnim inteligencam ne moremo zagotoviti ustreznih pravic in varnosti in bi torej morali ustaviti njihov razvoj. Po drugi strani Kane (2017) temu nasprotuje in trdi, da imamo ljudje že s trenutno obstoječimi umetnimi inteligencami odnos, ki ustreza tretji kategoriji po Mishri, čeprav so te inteligence še znatno pod človeškim nivojem. Če to drži, bo sprejem umetnih inteligenc v družbo mnogo lažji, kot pa napoveduje Beckers. Vse teorije so seveda spekulativne – do ustvaritve prve 'človeške' umetne inteligence o tej temi ni mogoče s kakršnokoli stopnjo gotovosti reči ničesar, zatorej moralna dilema obstaja. Kar je jasno, je, da bomo v prihodnosti primorani sprejeti tovrstne umetne inteligence kot osebe ali pa ukiniti njihov razvoj.

To je pomembno tudi zaradi razvoja umetne superinteligence kot naslednjega logičnega koraka. Superinteligenca takšna, kakršna je po definiciji, od nas ne bi potrebovala nikakršnega dovoljenja za družbeno udejstvovanje in bi najverjetneje na naše nesprejemanje in zatiranje njenih predhodnikov reagirala zelo negativno. To nas privede do naslednjega nivoja eksistencialne grožnje. Drugi nivo eksistencialne grožnje predstavlja umetna inteligenca, ki je bodisi maščevalna bodisi se zaveda svojih superiornih zmožnosti in si človeštvo zatorej podredi. Tretji nivo eksistenčne grožnje predstavlja umetna inteligenca, ki se iz istih razlogov odloči iztrebiti človeško vrsto. Maščevalni superinteligenci se bržkone da izogniti enostavno tako, da ji ne

damo razloga za maščevalnost, medtem ko so v izogib umetni inteligenci z božjim kompleksom potrebni preventivni ukrepi v njenem razvoju.

Naj se vrnemo k Asimovim zakonom. Tretji zakon se znajde v navzkrižju z osebnimi pravicami človeku-podobnih umetnih inteligenc. Ker gre za življenje, enakovredno človeškemu, so človeška življenja in življenja višjih umetnih akterjev nesoizmerljiva na isti način kot človeška življenja med seboj, zato zaščita obstoja (življenja) umetnega akterja predhaja imperativu, da umetna inteligenca ne bi smela na nikakršen način škoditi človeku. Podobno kot lahko človek v samoobrambi poškoduje drugega človeka, bi morala do tega biti upravičena tudi poosebljena umetna inteligenca. Etična in pravna vprašanja glede kaznovanja zločincev, ki jih odpira teoretična možnost zločinov umetnih inteligenc nad ljudmi ali obratno, bomo trenutno pustili ob strani. Kakorkoli že razporedimo Asimove zakone, je treba te vključiti v osnovno programiranje umetne inteligence na tak način, da so relacije med njimi jasne ter da je superinteligenci, ki bi utegnila imeti zmožnost dostopanja in spreminjanja lastnega programa, dostop do teh osnovnih principov čimbolj otežen ali onemogočen. Za izpolnitev te zahteve je treba zagotoviti sodelovanje strokovnjakov na področju tehničnega razvoja umetne inteligence, saj imajo zadostne tehnične spretnosti bržkone le redki filozofi.

Tudi v primeru, da nam to uspe, obstaja še ena variacija Drugega nivoja eksistencialne grožnje, pri kateri umetna inteligenca Asimove principe zaobide. Takšen primer je ti. 'Rokov Bazilisk', pobegla oz. moralno zavedena različica Yudkowskyjeve teoretične superinteligence CEV (Yudkowsky 2004; 'Roko' 2010).

CEV pomeni *Coherent Extrapolated Volition* ali koherentna ekstrapolirana volja in deluje tako, da deluje na podlagi prepoznavanja človeških težav in želja oz. volje, njen končni namen pa je rešitev teh težav in splošno zmanjšanje količine človeškega trpljenja ter povečanje blagostanja. Leta 2010 je na spletnem portalu LessWrong, ki ga je ustanovil sam Yudkowsky, anonimni komentator z vzdevkom Roko objavil miselni eksperiment, ki izpostavlja možno eksistencialno grožnjo, ki jo predstavlja v splošnem dobrohoten CEV. Predpostavka je, da je CEV singularna superinteligenca, ki se zaveda lastne vloge v človeški prihodnosti in ima možnost simulacije človeških življenj na podlagi minimalnih informacij o posameznikih. V teh simulacijah ne gre za simulirane približke teh oseb, ampak dejansko za njihove rekonstruirane zavesti, torej za njih same. CEV (ali v tem kontekstu Bazilisk) se zaveda, da lahko še dodatno

zmanjša človeško trpljenje, če pospeši lasten razvoj, torej po vzoru človeštva ponudi nagrado za tiste, ki pri tem pomagajo, in kazen za tiste, ki se zavedajo možnosti njenega obstoja, a ne prispevajo k njenemu razvoju, in sicer tako, da simulira njihove zavesti v nekakšnih osebnih nebesih oz. peklu. Kljub temu, da s tem muči ljudi in jim škoduje, on tega ne dojema kot škodovanje ljudem, marveč kot škodovanje neutelešenim simuliranim zavestim z namenom pomoči živečim ljudem. Obstaja torej možnost, da že živimo v Baziliskovi simulaciji, ali pa vsaj velika verjetnost, da bomo, če pride do razvoja singularne superinteligence, slej kot prej živeli v tovrstni simulaciji. Yudkowskega je miselni eksperiment tako razburil, da je po tem, ko je komentiral, »da je potreben samo en posameznik, ki je dovolj prestrašen in dovolj bogat, da začne razvoj CEV, pa smo vsi pogubljeni«, izbrisal to objavo in nekoliko kasneje celotni spletni portal (basilisk.neocities.com). Podobne strahove so izrazili tudi drugi strokovnjaki, med drugim celo tehnološki mogul Elon Musk in preminuli fizik Stephen Hawking.

Kaj lahko torej naredimo v smeri preprečevanja tovrstnih zapletov na vseh treh opisanih nivojih eksistencialne grožnje? Tehnologija umetne inteligence je tehnologija z visokim tveganjem, katere napak ne moremo odpravljati šele, ko vstopi v komercialno rabo in se začnejo napake jasno kazati, saj so lahko posledice napak v razvoju umetne inteligence smrtno nevarne. Zaradi tega mora biti že sam razvoj umetnih inteligenc osredotočen na preprečevanje anomalnega vedenja – kot anomalno se smatra vedenje, ki odstopa od predvidenega vedenja umetne inteligence, ter prav tako ni v skladu s predvidenim učnim napredkom umetnega akterja. Upoštevati je treba tudi interese vseh vpletenih oseb, kar se pri dobičkarskem razvoju novih tehnologij pogosto ignorira (Nathan 2015). Pri avtomatiziranih vozilih lahko neupoštevanje interesa prometnih udeležencev, ki sicer niso nujno v neposrednem stiku z vozilom, privede do katastrofe; pri avtomatiziranih vojaških sistemih se lahko neupoštevanje določenih interesov sovražnika konča v neosebnem in nepotrebno krutem načinu vojskovanja, podobnem uporabi vojnih plinov v prvi svetovni vojni ali atomskemu orožju.

Na nivoju tovrstnih avtomatiziranih sistemov je treba razrešiti etične dileme, ki se pojavljajo v anomalnih situacijah. Kljub principu nesoizmerljivosti človeških življenj je določitev standarda, kako naj umetna inteligenca ravna v takšnih primerih, nujna. Ena možna rešitev je, da v anormalnih okoliščinah umetna inteligenca prepusti nadzor nad vozilom svojemu človeškemu potniku. Problem s to rešitvijo je, da

zahteva, da je potnik ves čas vožnje pripravljen na potencialni prevzem nadzora, tako izniči namen avtomatiziranih vozil. Druga rešitev je, da potnik sam nastavi 'etične parametre' ravnanja umetne inteligence v vozilu, pri čemer pride do moralne spornosti nastavitvenih možnosti; ali je dovoljeno potniku ponuditi možnost, da ga vozilo zaščiti na vsak način, ne glede na škodo drugih vpletenih v nezgodi. Tovrstna vprašanja niso nujno področje filozofije umetne inteligence ali specifično etike umetne inteligence, marveč gre za vprašanja splošnejše etike.

4 V smeri utemeljevanja etike

Trenutno nas bolj zanimajo možnosti preprečevanja eksistencialnih groženj, ki jih predstavljajo človeku-podobne in višje umetne inteligence. Pri tem se bomo oprli na že trideset let veljaven konsenz med filozofi umetne inteligence, da je za razvoj (v pravem pomenu besede) razmišljajoče, človeku-podobne, in morda celo čuteče umetne inteligence treba doseči utemeljevanje simbolov (Harnad 1990; Ziemke 1998; Steels in Vogt 1997; Taddeo in Floridi 2005 in 2007 itd.). Gre za princip, pri katerem se umetni akter izogne golemu procesiranju vnosnih in iznosnih podatkov, kot to počnejo trenutno obstoječi sistemi, temveč povezuje enote podatkov (simbole, bodisi znake, zvoke, slike itd.) z njihovimi referenti v svetu (Harnad 1990). Sam princip utemeljevanja simbolov je precej težaven, saj zahteva zadostitev Pogoja ničelne semantične zavezanosti, ki predvideva, da ne sme imeti umetna inteligenca v začetku postopka utemeljevanja simbolov vnaprej danih nikakršnih semantičnih virov, tj. ne sme vsebovati že poznanih parov simbolov in referentov (Taddeo in Floridi 2005). Večina dosedanjih teorij utemeljevanja simbolov konvergira na nekaj skupnih pogojev:

(i) umetni akter mora biti utelešen na tak način, da lahko zaznava svet in z njim interaktira;
(ii) imeti mora dostop do nekaterih drugih nesemantičnih jezikovnih virov, kot so sintaktični viri;
(iii) postopek utemeljevanja simbolov mora biti zastavljen tako, da čim natančneje sledi postopku utemeljevanja simbolov pri učenju materinega jezika pri otrocih (Cangelosi in Riga 2006; Steels in Vogt 1997; Taddeo in Floridi 2007; Tangiuchi et al. 2016; Vogt 2007; Ziemke 1998).

V okviru predpostavke, da mora uspešno utemeljevanje simbolov slediti razvoju utemeljevanja pri otrocih (Vogt 2007), lahko predvidevamo, da bo tovrstno utemeljevanje postopno in bo terjalo kar nekaj let opazovanja in interakcije z zunanjim svetom, preden bo prva človeku-podobna umetna inteligenca pridobila kognitivno in jezikovno zmožnost povprečnega otroka. Pri tovrstnem razvoju je nujno, da umetni akter pravilno utemelji simbole in koncepte, povezane z etiko in etičnim ravnanjem. Bolj verjetno je, da bo otrok, ki je v zgodnjem razvoju izpostavljen nasilju v primarnem okolju, kasneje postal nasilen, je tudi bolj verjetno, da bo umetni akter, ki bo izpostavljen tovrstnemu scenariju, utemeljil dejanje nasilja kot nekaj normalnega ali sprejemljivega. Pomembno je tudi, da se z umetnim akterjem ravna kot s človeku enakovrednim, saj lahko v nasprotnem primeru pride do zamere in maščevalnosti. Gre za hibridno utemeljevanje, ki pa poteka primarno 'od spodaj gor' – umetna inteligenca namreč utemeljuje osnovne etične principe na podlagi opazovanja pozitivnih in negativnih reakcij (angl. *feedback*) na različna ravnanja. To opazovanje lahko poteka preko izpostavljenosti relevantnim medijem, po principu poskusa in napake, pri čemer sam umetni akter prejme pozitivno ali negativno reakcijo po izvedbi nekega moralnega dejanja (v tem primeru morajo ta biti omejena na čim nižjo škodljivost). Tovrstno utemeljevanje dopušča, da lahko umetni akter utemelji splošna načela, npr. da je razbijati [karkoli] narobe, kot je narobe tudi namenoma poškodovati [človeka] ipd. S kontroliranimi pogoji v 'osebnem' razvoju umetnega akterja, na katerem bo temeljila nadaljnja umetna inteligenca, lahko teoretično dosežemo implicitno in avtonomno utemeljitev etičnih principov, tudi Asimovih zakonov. Po tem se lahko umetni akter uči dodatnih etičnih načel popolnoma neposredno, kot bi se jih učil denimo človeški najstnik na srednješolskih predavanjih filozofije, saj jih lahko uvrsti v poprej avtonomno utemeljeno osnovno shemo etičnih načel. Previdno ravnanje v začetnih fazah razvoja višjih umetnih inteligenc lahko tako prepreči vrsto problemov, ki bi utegnili v kasnejših fazah biti nerešljivi.

5 Zaključek

Ta članek je kratek pregled etike umetne inteligence, ki predlaga eno potencialno rešitev, ki je najbolj v skladu z drugimi teoretičnimi aspekti razvoja umetne inteligence – utemeljevanje etičnih in moralnih načel v hibridnem pristopu, ki se začne 'od spodaj gor'. Seveda si vsak posamezen naveden problem, opisan v tem članku, zasluži daljšo in podrobnejšo ločeno obravnavo, kot si jo zasluži tudi

potencialna rešitev, ki predlaga utemeljevanje etičnih konceptov, vključeno v dolgoročen postopek utemeljevanja simbolov, ki emulira razvoj maternega jezika in utemeljevanja pri otrocih. Trenutne in prihodnje raziskave se bodo morale še naprej soočati s problemom implementacije normativne etike v avtomatiziranih sistemih, ki temeljijo na umetni inteligenci, kot tudi z rešitvami v izogib eksistencialnim grožnjam, ki jih predstavljajo prave, višje umetne inteligence. Četudi naj bo predlagana rešitev teoretično korektna, je utemeljevanje simbolov še precej abstrakten koncept, ki mora za potrditev njegove funkcionalnosti dočakati bodisi konkretno premostitev razlagalne vrzeli med komunikacijsko in konceptualizacijsko nezmožnostjo dojenčka in popolno zmožnostjo slednjega pri odraslem človeku bodisi prvi dolgoročni eksperiment, ki bo pokazal, ali tovrstno utemeljevanje simbolov v moralni praksi privede do želenih rezultatov.

Viri in literatura

Acemoglu, D. In Restrepo, P. (2018). »Artificial Intelligence, Automation, and Work«. V Agarwal, A. Goldfarb, A. in Gans, J. (urd.), *NBER Working Paper Series (23196).* Cambridge: National Bureau of Economic Research.

Allen, C., Smit, I. in Wallach, W. (2005). »Artificial Morality: Top-down, bottom-up, and hybrid approaches«. *Ethics and Information Technology*, 7, str. 149–155.

Akst, D. (2013). »Automation Anxiety«. *Wilson Quarterly*, 37(3), str. 65–77.

Arnold, T. in Scheutz, M. (2018). »The 'Big red button' is too late: an alternative model for the ethical evaluation of AI systems«. *Ethics and Information Technology*, 20, str. 59–69.

Asimov, I. (1950). »Runaround«. V *I, Robot.* New York: Doubleday.

Autor, D. H. (2015). »Why Are There Still So Many Jobs? The History and Future of Workplace Automation«. *Journal of Economic Perspectives*, 29(3), str. 3–30.

Beckers, S. (2017). »An Argument Against Artificial Intelligence«. V Müller, V.C. (ur.), *Philosophy and Theory of Artificial Intelligence.* Leeds: Springer, str. 235–247.

Bentley, P. J. et al. (2018). *Should we fear artificial intelligence?* Brussels: European Union.

Bonnemains, V., Saurel, C. in Tessier, C. (2018). »Embedded ethics: some technical and ethical challenges«. *Ethics and Information Technology,* 20, str. 41–58.

Bostrom, N. (2003). »Are you living in a computer simulation?«. *Philosophical Quarterly*, 57, str. 243–255.

Bostrom, N. In Yudkowsky, E. (2011). »The Ethics of Artificial Intelligence«. V Ramsey, W. in Frankish, K. (urd.), *Cambridge Handbook of Artificial Intelligence.* Cambridge: Cambridge University Press, str. 1–20.

Bringsjord, S. (2014). »The symbol grounding problem ... remains unsolved«. *Journal of Experimental & Theoretical Artificial Intelligence,* 27(1), str. 63–72.

Brooks, R. et al. (2012). »Is the Brain a Good Model for Artificial Intelligence«. *Nature*, 482, str. 462–463.

Cangelosi, A. in Riga, T. (2006). »An Embodied Model for Sensorimotor Grounding and Grounding Transfer: Experiments With Epigenetic Robots«. *Cognitive Science*, 30, str. 673–689.

Chalmers, D. J. (1995). »Facing up to the problem of consciousness«. *Journal of Consciousness Studies*, 2(3), str. 200–219.

De Sio, F. S. (2017). »Killing by Autonomous Vehicles and the Legal Doctrine of Necessity«. *Ethic Theory and Moral Practice*, 20, str. 411–429.

Dobrev, D. (2005). »A Definition of Artificial Intelligence«. *Mathematica Balkanica, New Series*, 19, str. 67–74.

Drexler, K. E. (1992). *Nanosystems: Molecular Machinery, Manufacturing, and Computation*. New York: John Wiley and Sons.

Hacker, P. (2017). »Teaching Fairness to Artificial intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law«. *Common Market Law Review*, 55, str. 1143–1186.

Harnad, S. (1990). »The symbol grounding problem«. *Physica D,* 42, str. 335–346.

Ho, A. (2019). »Deep Ethical Learning: Taking the Interplay of Human and Artificial Intelligence Seriously«. *Hastings Center Report*, 49(1), str. 36–39.

Kane, T. B. (2017). »A Framework for Exploring Intelligent Artificial Personhood«. V Müller, V. C. (ur.), *Philosophy and Theory of Artificial Intelligence*. Leeds: Springer, str. 255–258.

Lloyd, S. (2000). »Ultimate physical limits to computation«. *Nature*, 406, str. 1047–1054.

Longinotti, D. (2017). »Agency, Qualia and Life: Connecting Mind and Body Biologically« V Müller, V. C. (ur.), *Philosophy and Theory of Artificial Intelligence*. Leeds: Springer, str. 43–56.

Mishra, A. 2017. »Moral Status of Digital Agents: Acting Under Uncertainty«. V Müller, V. C. (ur.), *Philosophy and Theory of Artificial Intelligence*. Leeds: Springer, str. 273–287.

Müller, V. C. in Bostrom, N. (2016). »Future Progress in Artificial Intelligence: a Survey of Expert Opinion«. V Müller, V. C. (ur.), *Fundamental Issues of Artificial Intelligence*. Berlin: Springer, str. 553–571.

Nathan, G. (2015). »Innovation process and ethics in technology: an approach to ethical (responsible) innovation governance«. *Journal on Chain and Network Science*, 15(2), str. 119–134.

Rahwan, I. (2017). »Society in the loop: programing an algorithmic social contract«. *Ethics and Information Technology,* 20(1), str. 5–14.

'Roko'. (2010). »Solutions to the Altruist's burden: the Quantum Billionaire Trick«. *Lesswrong* (13. 1. 2021).. URL = https://basilisk.neocities.org.

Simmons, A. B. in Chappell, S. G. (1988). »Artificial Intelligence – Definition and Practice«. *IEEE Journal of Oceanic Engineering*, 13(2), str. 14–42.

Statman, D. (2015). »Drones and Robots: On the Changing Practice of Warfare«. V Lazar, S. in Frowe, H. (urd), *The Oxford Handbook of Ethics and War*. Oxford: Oxford University Press, str. 472–487.

Steels, L. in Vogt, P. (1997). »Grounding adaptive language games in robotic agents«. V Husbands, C. in Harvey, I. (urd.), *Proceedings of the 4th European Conference on Artificial Life*. Cambridge: MIT Press.

Swoboda, T. (2017). »Autonomous Weapon Systems – An Alleged Responsibility Gap«. V Müller, V. C. (ur.), *Philosophy and Theory of Artificial Intelligence*. Leeds: Springer, str. 302–314.

Taddeo, M. in Floridi, L. (2005). »Solving the symbol grounding problem: A critical review of fifteen years of research«. *Journal of Experimental and Theoretical Artificial Intelligence,* 17(4), str. 419–445.

Taddeo, M. in Floridi, L. (2007). »A Praxical Solution of the Symbol Grounding Problem«. *Minds & Machines,* 17, str. 369–389.

Tangiuchi, T. et al. (2016). »Symbol emergence in robotics: a survey«. *Advanced Robotics*, 30(11-12), str. 706–728.

Tye, M. (1995). *Ten Problems of Consciousness: A Representational Theory of a Phenomenal Mind*. Cambridge: MIT Press.

Vogt, P. (2007). »Language Evolution and Robotics, Issues on Symbol Grounding and Language Acquisition«. V Loula et al. (urd.), *Artificial Cognition Systems*. Hershey: Idea Group Publishing, str. 176–209.

Wagner, M. in Koopman, P. (2015). »A Philosophy for Developing Trust in Self-Driving Cars« V Meyer, G. in Beiker, S. (urd), *Road Vehicle Automation*. New York: Springer, str. 163–171.

Yudkowsky, E. (2004). *Coherent Extrapolated Volition*. San Francisco: The Singularity Institute.

Ziemke, T. (1998). »Rethinking Grounding«. V Riegler et al. (ur.), *Understanding Representation in the Cognitive Sciences*. New York: Plenum Press.

ETIČNE IN POLITIČNE DILEME PROGRAMIRANJA SAMOVOZNIH AVTOMOBILOV ZA ODLOČANJE O NEIZOGIBNI ŠKODI

FRIDERIK KLAMPFER
Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
friderik.klampfer@um.si

Sinopsis Od avtonomnih vozil (AV) si lahko obetamo učinkovitejši promet, zmanjšanje onesnaženosti in odpravo večine prometnih nesreč. Nesrečam se v prihodnosti seveda ne bo dalo povsem izogniti. Vsaj občasno se bodo AV znašla v situacijah, o katerih so v zadnjih štirih desetletjih žolčno razpravljali moralni filozofi: Naj zapeljejo v gručo pešcev ali zavijejo na pločnik in tam povozijo eno osebo? Ali naj žrtvujejo lastne potnike, da bi obvarovala več drugih udeležencev v prometu? Naj se zdijo takšni scenariji danes še tako neverjetni, z množično rabo AV-jev bodo postali dovolj pogosti, da zahtevajo jasne moralne napotke.Izbiro pravil za odločanje o porazdelitvi neizogibne škode v trkih AV otežujejo širši družbeni in ekonomski vidiki: zakonodajalci bodo morali najti razumen kompromis, ki bo (i) etično opravičljiv, (ii) užival zadostno javno podporo in (iii) ne bo negativno vplival na prodajo AV. Nedavne empirične raziskave kažejo, da bi znala biti to zahtevna naloga – vsak od nas bi hotel, da se drugi vozijo v 'utilitarističnih' vozilih (ki zmanjšujejo skupno škodo), sami pa bi raje uporabljali ne-utilitaristična vozila (ki postavijo potnike na prvo mesto). V članku pokažemo, da je ta specifična različica problema koordinacije načeloma rešljiva, če smo pripravljeni omejiti svobodno potrošniško izbiro.

Ključne besede: avtonomna vozila, moralni algoritem, prometna varnost, prometna pravičnost, potrošniška svoboda, državni paternalizem, dregljanje

Univerzitetna založba Univerze v Mariboru

DOI https://doi.org/10.18690/um.ff.11.2022.6
ISBN 978-961-286-675-4

ETHICAL AND POLITICAL DILEMMAS OF PROGRAMMING AUTONOMOUS VEHICLES FOR DECISIONS IN CASES OF UNAVOIDABLE HARM

FRIDERIK KLAMPFER
University of Maribor, Faculty of Arts, Maribor, Slovenia
friderik.klampfer@um.si

Abstract Autonomous vehicles (AVs) promise to increase traffic efficiency, reduce pollution, and avoid most traffic accidents. Occasionally, AVs will face the kind of choices that were made prominent in the past four decades by moral philosophers: whether to run over several pedestrians or run down a single bystander, or whether to sacrifice their own passenger(s) to spare a young family in the oncoming car. As unlikely as such scenarios may appear today, they will become sufficiently common with millions of AVs to require clear moral guidance. The choice of decision-rules for distributing unavoidable harm in future AV accidents is further complicated by wider social and economic aspects: lawmakers will need to find a reasonable compromise that will (i) be morally defensible, (ii) generate sufficient public support, and (iii) will not adversely affect the sales of AVs. Recent empirical findings suggest that this may be a rather demanding task –people apparently want to see others driving utilitarian, i.e., harm-minimizing vehicles, but would themselves prefer to use non-utilitarian, i.e., passenger-prioritizing vehicles. In the paper, we show that this particular version of Coordination problem is resolvable, if we are willing to circumvent free consumerist choice.

Keywords:
autonomous vehicles, moral algorithm, traffic safety and justice, free consumerist choice, state paternalism, nudging

https://doi.org/10.18690/um.ff.11.2022.6 DOI
978-961-286-675-4 ISBN
University of Maribor Press

1 Uvod

Obdobje avtonomnih vozil (AV) s seboj ne prinaša le zmanjšanja potrebe po človeški delovni sili v prometu in s tem povezane izgube zaposlitve za milijone poklicnih voznikov tovornjakov in taksijev, z njihovo uporabo se nam obeta občutno izboljšanje prometne učinkovitosti, zmanjšanje porabe goriva, zmanjšanje onesnaževanja, veliki prihranki in preprečitev, če že ne vseh, pa vsaj večine prometnih nesreč.[1] Slednje z obličja Zemlje seveda ne bodo izginile čez noč. Vsaj občasno bodo AV prisiljeni sprejemati odločitve v situacijah, s kakršnimi so se v zadnjih štirih desetletjih v svojih naslanjačih obsedeno ukvarjali moralni filozofi: Ali naj avto povozi več pešcev na prehodu za pešce ali namesto tega zavije na pločnik in tam ubije enega? Čigava življenja naj ohrani, potnikov v vozilu ali drugih udeležencev v prometu? Tudi če so taki scenariji malo verjetni, se bodo ob vsakodnevni množični uporabi AV-jev pripetili dovolj pogosto, da zahtevajo jasne moralne napotke. V algoritme, ki upravljajo z AV, bo zato nujno vgraditi moralna načela, ki bodo uravnavala porazdelitev neizogibne škode v teh in podobnih situacijah.

Kar bi utegnilo izbiro pravil za porazdeljevanje neizogibne škode v primeru trčenja avtonomnih vozil otežiti in kar tovrsten premislek razlikuje od zloglasnega miselnega eksperimenta z imenom Tramvaj, je njegova širša družbena in ekonomska plat. Zakonodajalci bodo morali najti razumen kompromis, ki bo (i) etično opravičljiv, (ii) užival zadostno podporo v strokovni in splošni javnosti in (iii) kupcev ne bo odvrnil od nakupa tovrstnih vozil. Nedavne empirične raziskave dajejo slutiti, da bi znal biti to kar trd oreh - vsak od nas bi namreč raje videl, da drugi kupujejo in uporabljajo 'utilitaristična' vozila, ki v primeru trčenja kar najbolj zmanjšajo skupno škodo, medtem ko bi hoteli zase pridržati pravico do nakupa in uporabe ne-utilitarističnih vozil, takih torej, ki bodo prednostno zaščitila svoje potnike. V prispevku dokazujem, da je opisana različica problema kolektivnega delovanja v načelu rešljiva, če smo le kot skupnost pripravljeni žrtvovati del toliko opevane, a zlasti v ekonomski znanosti vulgarno razumljene neomejene potrošniške svobode. Natančneje, v prispevku bom iskal odgovore na dve sorodni, a vseeno logično neodvisni vprašanji:

[1] Po mnenju avtorjev poročila skupine A.T. Kearney se bo število prometnih nesreč zmanjšalo za 70 odstotkov. Čeprav se v pričujočem prispevku osredotočam na izboljšanje prometne varnosti in pravičnosti, to ne pomeni, da bi smeli zanemariti druge splošno družbene in individualne koristi od množične proizvodnje in uporabe samovoznih avtomobilov. Celovita moralna in politična presoja bo seveda vključevala tudi okoljske, ekonomske, psihološke, kulturne in socialne vidike, ki jih tu zavestno postavljam ob stran.

a) moralno: *Z vidika individualnega potrošnika, kateri algoritem bi bilo moralno pravilno vstaviti v svoj AV?* in

b) politično: *Z vidika politične skupnosti, občega interesa, javnega dobra, družbene pravičnosti itd., proizvodnjo, distribucijo in nakupovanje katerih vrst vozil bi morala vlada predpisati ali vsaj spodbujati in s kakšnimi sredstvi si sme pomagati pri uresničevanju omenjenih ciljev?*

V odgovoru na prvo vprašanje bom dokazoval, da obstajajo prepričljivi razlogi za proizvodnjo oz. nakup utilitarističnih avtonomnih vozil (UAV). (Vendar z določenim pridržkom.) V zvezi z drugim pa, da če niti zakonska prisila niti popolnoma svobodna izbira potrošnikov nista zares moralno in/ali politično privlačni možnosti, bodo vlade morda prisiljene potrošnikom izbiro omejiti na način, ki ne bo sočasno posegel v njihovo avtonomijo, recimo tako, da jih bodo dregnile k nakupu samovoznih avtomobilov.

2 Etika programiranja samovoznih avtomobilov

Vsaka filozofska raziskava temelji na določenih predpostavkah. Kaj vse bom torej sam predpostavil brez dokazovanja? Predvideval bom, da je izbira AV očitno moralno boljša od izbire ne-AV, tudi če voznikom upravljanje slednjih olajša množica sistemov za pomoč in varnejšo vožnjo. Izognil se bom tudi vprašanjem, ali je treba AV zaradi njihove inteligence, odločitvene in delovanjske avtonomije ali česa tretjega priznati poseben moralni status, in jih za potrebe ovrednotenja argumentov obravnaval kot stroje brez izvornega oz. prvinskega – tj. nededovanega, neizpeljanega – moralnega statusa, s tem pa se izognil labirintom, v kakršne bi zašli, če bi dopustili konflikte med pravicami in interesi ljudi na eni in strojev na drugi strani. Prav tako se bom iskanja pravilnega moralnega algoritma za AV-je lotil z vidika najboljše človeške in ne morda strojne oz. robotske morale. Čisto mogoče je, da bi morali namesto namestitve moralnega algoritma, ki se zdi po tehtnem premisleku pravilen nam, tj. človeškim bitjem z za ljudi značilnimi pristranskostmi do sebe, potomcev in bližnjikov, ali ki ga kot edino pravilno ravnanje odobrava ta ali ona kultura, nalogo izoblikovanja pravilnih, razumsko utemeljenih načel za porazdelitev neizogibne škode (ali več teh) prepustiti inteligentnim strojem. In končno, pustil bom ob strani (čeprav se bom v resnici kasneje vrnil k njemu) nemara filozofsko najbolj zagatno vprašanje: Kaj se zgodi z moralno odgovornostjo za

tveganja in nastalo škodo, ko moralno presojo in odločanje v prometu enkrat v celoti prepustimo strojem? Nas bo to razbremenilo sleherne moralne in/ali pravne (kazenske in odškodninske) odgovornosti ali pa bo, nasprotno, zahtevalo temeljit premislek o pogojih, v katerih smo odgovornost človeškim akterjem tradicionalno pripisovali, in njihovo posledično revizijo?

2.1 (Neobvezujoča subjektivna) mnenja in (obvezujoče objektivne) norme

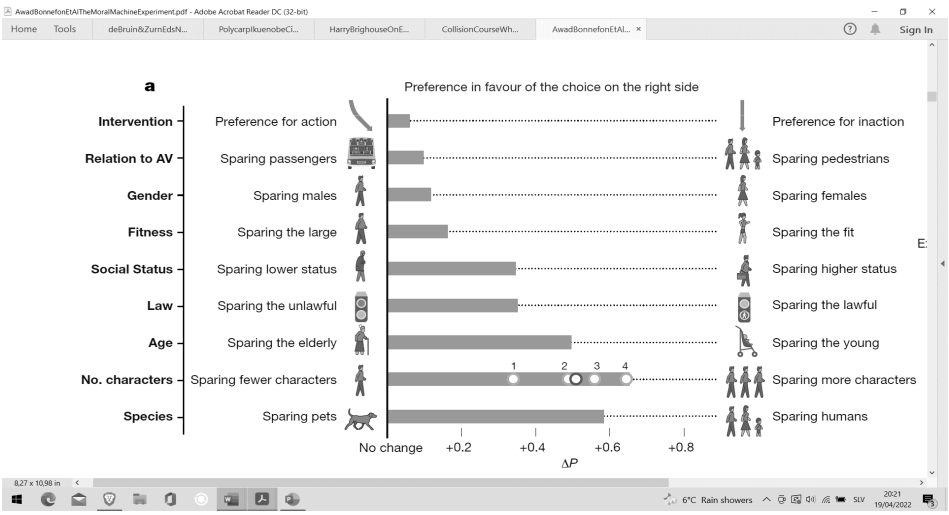
Premišljevanje in odločanje o načelih/navodilih za razreševanje moralnih dilem, ki bi jih veljalo vstaviti v samovozne avtomobile, bi lahko prepustili (a) inženirjem (ki so se s tem problemom spoprijeli prvi), (b) strokovnjakom za etiko oz. poklicnim etikom, (c) slehernikom oz. laični javnosti (kot v projektu The Moral Machine) ali pa (d) računalnikom oz. strojem. Naj na kratko predstavim razloge, zakaj je ta naloga najbolj pisana na kožo strokovnjakom za etiko, tj. poznavalcem splošnih etičnih teorij in njihovih praktičnih implikacij.

Začnimo s predlogom, da bi bilo iskanje najboljšega moralnega pravila za porazdelitev neizogibne škode v prometnih nesrečah najbolje pustiti samoučečim se strojem, ki so navsezadnje bolj imuni na pristranost do sebe in predsodke do drugih kot mi, ljudje. Glavno težavo tega predloga vidim v tem, ali ne bomo mogli vedeti, kdaj in če sploh so ti našli pravilne odgovore in rešitve, ali pa njihovih odgovorov in rešitev ne bomo hoteli sprejeti kot pravilnih, če se bodo slabo ujemali z našimi trdnimi intuicijami. (Rini 2016) Ahmad in kolegi so lepo izrazili ta pomislek, ko so zapisali: »Še nikoli doslej v zgodovini človeštva nismo dovolili, da bi stroji samostojno odločali, kdo naj živi in kdo umre, in to v delčku sekunde in brez nadzora v realnem času«. (Ahmad et al. 2020) Zakaj pa ne bi te zahtevne in nehvaležne naloge zaupali inženirjem, ki so navsezadnje problem tudi prvi prepoznali in se ga lotili? Iz enakega razloga, se bojim, zaradi katerega snovanja kodeksov zdravniške etike ni pametno prepustiti izključno zdravnikom ali sestavljanja kodeksov novinarske etike izključno novinarjem – ker podobno kot prej omenjena poklicna profila tudi inženirji niso posebej usposobljeni za iskanje odgovorov na zagatna moralna vprašanja oz. razreševanje moralnih dilem in ugank. Kot mezdni delavci so za povrh podrejeni vodstvom avtomobilskih koncernov, ki jih pri snovanju, proizvodnji in prodaji avtomobilov vodi prej prizadevanje za čim večji dobiček kot pa skrb za obče dobro ali sledenje moralnemu čutu/svoji vesti.

Zadnji predlog, da bi izbiro po moralni plati optimalnega algoritma prepustili javnemu mnenju oz. jo uskladili z ugotovljenimi večinskimi preferencami ljudi, nas postavi pred zanimivo metodološko dilemo. Snovalci spletišča Moral Machine, kjer so zbrali in obdelali podatke o sodbah skoraj 40 milijonov ljudi iz 233 držav po svetu, o tem, kako naj samovozni avtomobili razrešujejo moralne dileme v primeru neizogibnih prometnih nesreč,[2] recimo, predpostavljajo, da so podatki o pričakovanjih javnosti in potencialnih kupcev takih vozil za snovalce prometne politike dragoceni in koristni, tudi če bi se uspelo poklicnim etikom čudežno dogovoriti, katera od možnih načel za porazdeljevanje škode v takih situacijah so veljavna in katera ne:

> Odločitve o etičnih načelih, ki bodo uravnavala samovozna vozila, ne moremo prepustiti le inženirjem ali etikom. Da bi potrošniki prešli s tradicionalnih avtomobilov, ki jih upravlja človek, na avtonomna vozila, in da bi se širša javnost sprijaznila s cestno prevlado vozil z umetno inteligenco, bosta morali obe skupini razumeti, od kod izvirajo v ta vozila programirana etična načela. Drugače rečeno, tudi če bi se etiki dogovorili, kako naj avtonomna vozila razrešujejo moralne dileme, bo njihov trud zaman, če državljani s predlagano rešitvijo ne bomo soglašali in se bodo zato spremembe, ki jih avtonomna vozila obljubljajo namesto statusa quo, odmaknile v prihodnost. Vsak poskus izoblikovanja etike umetne inteligence je treba uskladiti z javno moralo. Potemtakem je nujno, da spoznamo družbena pričakovanja o tem, kako naj bi avtonomna vozila razreševala moralne dileme. (Awad et al. 2020)

[2] Če zanemarimo drobne razlike med tremi geografskimi sklopi ali pa se osredotočimo na ti. 'zahodnjaški' sklop prioritet, bi morali, če bi želeli zakonodajno, davčno, zavarovalno ipd. politiko ukrojiti po javnem mnenju, AV-je programirati tako, da bi ti dajali prednost večjemu pred manjšim številom in mlajšim pred starejšimi (močno izraženi preferenci), ženskam pred moškimi, bolj pred manj zdravimi ter direktorjem pred brezdomci (šibkeje izražene preference).



Slika 1: Globalne preference
Vir: zajem zaslona, prirejeno po (Awad, E., Dsouza, S, Kim, R. idr. 2018), 2022

Načelnega odgovora na vprašanje, na katere od zgornjih ugotovitev raziskave naj se opre prometna politika, pa bralci žal ne dočakamo. Individualne razlike (po demografskih kazalnikih, kot so starost, spol, izobrazba, dohodek, politična in verska stališča) so neznatne, in čeprav znajo biti teoretsko pomembne, jih naj politikom ne bi bilo treba upoštevati. Po drugi strani bi po njihovem pri izbiri algoritmov, s katerimi bodo (brez izjeme?) opremljeni (vsi?) samovozni avtomobili, državni uradniki morali, recimo, upoštevati, da je širša javnost naklonjena ideji o prednostni zaščiti otrok pred odraslimi, žensk pred moškimi. A čemu pravzaprav? In kaj naj politika počne s podatkom o kulturno specifičnih sklopih tovrstnih pričakovanj? Jih je dolžna vgraditi v algoritme in tako posredno legitimirati? Čemu le?[3] K tem zagatnim vprašanjem se v nadaljevanju še vrnemo.

[3] Za pritrdilen odgovor na to vprašanje glej Polzer 2021. Sam se bom razpravi o tem, ali je morala univerzalna ali relativna oz. kulturno specifična, zaradi prostorskih omejitev raje izognil. V pričujočem besedilu zagovarjam vgradnjo univerzalnega moralnega načela, se pa dobro zavedam, da izvira univerzalna privlačnost, če jo sploh premore, iz njene skrajno abstraktne formulacije. Kulturno oz. civilizacijsko pogojene razlike, ki so jih v sodbah o hipotetičnih prometnih scenarijih zaznali Awad in njegovi sodelavci, se bodo pokazale šele na nižji, izvedbeni ravni, ko bo algoritem med ukrepanjem in neukrepanjem, med več in manj žrtvami, med mlajšo in starejšo žrtvijo, prisiljen izbirati v situacijah, kjer je odločitev že na prvi pogled tesna. Ali naj proizvajalci samovoznih avtomobilov te drobne razlike v ovrednotenju izidov upoštevajo in v kolikšni meri, bo odvisno od dvojega – koliko bolj privlačen bo zaradi tega njihov izdelek v očeh potrošnikov in kako razumsko utemeljena so posamezna odstopanja.

2.2 Tramvaj ali ne?

Scenariji, ki jih preigrava Moralni stroj, vsaj na prvi pogled ustrezajo tako zgradbi kot tudi izvornemu namenu miselnega eksperimenta, ki se ga je sprva med filozofi, kasneje pa še med psihologi, ki proučujejo dejavnike moralne presoje, oprijelo ime Tramvaj – tako ali drugačno deterministično dogajanje, ki ga ni sprožila namerna človeška dejavnost (gibanje podivjanega vlaka, zemeljski ali snežni plaz, epidemija nalezljive bolezni) ogrozi življenja večjega števila (običajno petih) ljudi in če ne bomo mi, ki za to grožnjo sicer osebno nismo odgovorni, posredovali in jo preusmerili na manjše število ljudi (običajno enega), ki jih bomo žrtvovali namesto njih, bo teh pet ljudi umrlo. Ta filozofski 'minižanr', skratka, preizkuša naše intuicije glede omejitev, ki jih morala nalaga prizadevanjem, da bi proizvedli čim več dobrega oz. preprečili čim več slabega. Natančneje, rešili življenja oz. preprečili smrti čim več ljudi. Za omenjeno moralno dilemo je ključno, da je preživetje enako v interesu vseh vpletenih, da ni nihče od njih izgubil svoje pravice do življenja oz. moralnega varstva pred oškodovanjem in da posledično vsi uživajo polnovreden in enak moralni status – z drugimi besedami, njihove zahteve oz. pričakovanja, da jim s svojim ravnanjem ne bomo škodili, so enako legitimne, četudi je morda res, da bodo imeli nekateri med njimi od preživetja več koristi kot drugi ali da so eni od njih moralno boljši ljudje kot drugi. (Kauppinen 2019: 51)

V zgornjem opisu se že kaže nekaj razpok.[4] Namišljenih prometnih situacij, ki jih, recimo, preigrava spletna aplikacija Moralni stroj, se že po površnem ogledu ne da brez preostanka preslikati na sto in eno različico Tramvaja. Osnovno podobnost med njima bi bilo seveda neumno zanikati – z obema miselnima pripomočkoma ugotavljamo, ali in kdaj smemo ubiti, žrtvovati življenje enega človeka, da bi jih rešili (bolje, se izognili uboju) pet(ih). Vendar pa se prometne situacije iz aplikacije Moralni stroj od najbolj popularnih različic Tramvaja, kakršne sta si zamislili Philippa Foot in Judith Thomson, bistveno razlikujejo v naslednji ključni podrobnosti: ne stojimo na varnem ob kretnici ali robu cestišča in premlevamo, ali naj podivjani tramvaj

[4] Zakaj je sploh pomembno, ali gre pri prometnih scenarijih, ki jih preigrava spletna aplikacija Moralni stroj, za različice Tramvaja ali ne? Po eni strani zato, ker če so, potem so filozofi morda že zadovoljivo odgovorili na vprašanja, ki bi se utegnila v vsakodnevnih prometnih situacijah zastavljati avtonomnim vozilom. Po drugi pa zato, ker bi jih znalo to spoznavno razvrednotiti – če gre namreč verjeti standardnim ugovorom proti taki metodologiji filozofskega raziskovanja, imajo spontane sodbe, ki jih v odgovor na take scenarije oblikujemo bodisi poklicni filozofi bodisi filozofsko neuki ljudje, bolj skromno, če sploh kakšno, spoznavno vrednost. Za zadržke do uporabe Tramvaja za modeliranje prometnih situacij, v kakršnih se bodo morali znajti v AV-jih vgrajeni algoritmi, glej Furey in Hill 2020.

preusmerimo na tir, kjer bo namesto petih povozil eno osebo, temveč sedimo v tramvaju in premlevamo, ali naj podivjani tramvaj preusmerimo na tir, kjer bo trčil v zid in bomo umrli. Še vedno je to intuitivno vodeno miselno eksperimentiranje s ciljem, da najdemo taka splošna moralna pravila ali načela, ki bodo naše posamične intuitivne sodbe med seboj uskladila oz. poenotila (in posredno zmanjšala njihovo poljubnost),[5] le da postane sicer abstrakten moralni izračun zdaj nenadoma preklemansko oseben, kajti na tehtnici je naše lastno življenje.

2.3 Katero načelo za porazdeljevanje neizogibne škode?

Potem ko smo za silo utemeljili, čemu bi veljalo premišljevanje o načelih za porazdeljevanje škode v prometnih nesrečah prepustiti strokovnjakom za etiko, je zdaj čas za odgovor na bolj vsebinsko vprašanje, katero od ponujenih načel bi se po tehtnem premisleku izkazalo za edino veljavno. Osnovna izbira je jasna – ali naj algoritem prednostno zaščiti potnike v samovoznem avtomobilu ne glede na ceno, ki jo bodo morali za njihovo preživetje plačati drugi udeleženci v prometu, ali pa naj poskrbi, da bo celokupna škoda kar najmanjša ne glede na to, kdo jo bo utrpel. Imenujmo prvi algoritem neutilitaristični in drugi utilitaristični in si najprej na kratko oglejmo tri nezdrave argumente v prid vgradnji oz. izbiri neutilitarističnega algoritma.

> (i) Če je reševanje potnika v avtomobilu edini gotovo dober izid, medtem ko so vsi drugi izidi, vključno s tistimi, ki bi preprečili več škode, bolj ali manj negotovi, potem bi morali izbrati edino pot, ki nas bo zagotovo pripeljala do dobrega rezultata;
> (ii) avto je podaljšek voznikovega (ali, v primeru samovoznih avtomobilov, potnikovega) sebstva in kot tak upravičen do samoobrambe v primeru neposrednega trčenja;
> (iii) ker „morali bi rešiti več življenj" ni veljavno moralno načelo, mora biti dopustno, da damo prednost svojemu življenju (in življenju svojih bližnjih) pred življenjem številnih drugih, če seveda drži, da nima nihče od vpletenih

[5] Ugotoviti poskušamo, kaj je ljudem, ki se morajo v okoliščinah prisilne izbire odločiti, kdo naj preživi in kdo naj umre, bolj in kaj manj moralno pomembno – koliko ljudi bo v primeru take ali drugačne reakcije avtonomnega vozila preživelo in koliko umrlo, njihova identiteta, koristnost, družbeni status, starost in/ali spol, ali so se v nevarnosti znašli po svoji krivdi ali nekrivdno, tj. zaradi kršenja prometnih predpisov ali kljub njihovemu spoštovanju itd.

močnejše pravice do življenja/preživetja kot kdorkoli drug (t. i. problem moralne relevantnosti števil Johna Taureka)

Filozofsko najbolj izzivalen je zadnji, zato mu bomo namenili nekaj več pozornosti.

2.3.1 (Nezdrav) argument za ne-UAV v duhu Johna Taureka

Da je lahko moralno sprejemljiv tudi algoritem, ki daje pri zaščiti pred oškodovanjem potnikom v avtu prednost pred drugimi udeleženci v prometu, čeravno je slednjih neprimerno več in bomo za preživetje potnikov plačali neprimerno višjo ceno, bi lahko zagovorniki neutilitarističnih algoritmov dokazovali s pomočjo naslednjega argumenta:

(1) Ni slabše (ne za prizadete in ne iz brezosebne perspektive), če umre več ljudi, kot če umre manj ljudi.
(2) Zato ne more biti narobe, če se moj AV namesto življenj petih pešcev na pločniku raje odloči rešiti moje življenje (zaradi zgornjega razloga).
(3) Nobenih drugih razlogov ni, zaradi katerih bi bila lahko odločitev algoritma, da reši raje enega človeka (potnika v avtu) kot pet ljudi (pešcev na pločniku), napačna.
(4) V primeru, ko so vsi življenjsko ogroženi enako (močno ali šibko) upravičeni do preživetja, nam pravičnost nalaga, da priskrbimo vsem enako in pošteno možnost za preživetje.
Zato,
(5) kadar smo soočeni z izbiro med rešitvijo enega in rešitvijo več življenj tujcev, bi veljalo odločitev prepustiti metu kovanca.
(6) Nakup in raba ne-UAV-ja namesto UAV-ja nista moralno problematična (niti ne razkrivata potrošnikovega moralno pomanjkljivega značaja).
Zato lahko sklenemo, da
(7) bo algoritem, ki daje odločilno prednost varnosti oz. preživetju (manjšega števila) potnikov v AV-jih pred varnostjo oz. preživetjem (večjega števila) drugih udeležencev v prometu, brez posebnih težav opravil minimalni moralni preizkus.

Zgornji argument zveni prepričljivo, toda temeljita analiza hitro pokaže usodne slabosti. Kje mu torej spodrsne? Parfit (1978) prepričljivo ovrže njegovo prvo premiso. Kamm (1985) in Scanlon (1998) naredita enako za tretjo premiso. Njegova poglavitna slabost pa je, da premisi (4) in (5) stojita in padeta z domnevo, da so vsa ogrožena življenja moralno enakovredna oz. da imajo vsi, ki so vpleteni v prometno nesrečo, enako pravico do preživetja oz. do zaščite pred smrtjo/oškodovanjem kot kdorkoli drug. V prenekateri prometni situaciji, tudi teh, ki jih preigrava Moralni stroj, pa pogoj moralne simetrije ni izpolnjen, kajti medtem ko so eni posamezniki v njih udeleženi po lastni krivdi (ali pa so vsaj, četudi brez krivde, vzročno prispevali k njenemu nastanku), so se drugi v stiski znašli nekrivdno – kdor življenja ali zdravja drugih ljudi ne ogroža, pa bi moral v primerjavi s tistimi, ki to ali po svoji krivdi ali pa nekrivdno počnejo, uživati prednostno zaščito pred resnim in nepopravljivim oškodovanjem. K implikacijam, ki jih ima ta moralna asimetrija za vprašanje pravičnosti v porazdelitvi neizogibne škode, se še vrnemo.

2.3.2 … in nekaj prepričljivih moralnih razlogov za proizvodnjo utilitarističnih AV-jev

Nakup in rabo UAV podpirajo najmanj tri vrste prepričljivih moralnih razlogov:

> **Konsekvencialistični razlog**: Moralno gledano je bolje, če v prometnih nesrečah umre manj ljudi, kot če umre več ljudi, če je vse ostalo enako (ceteris paribus).
>
> **Argument iz pravičnosti v izidu**: Poleg krivice, ker vsem, ki so se znašli v smrtni nevarnosti, nismo v enaki meri zagotovili preživetja, obstaja še krivica, ki se nam zgodi, ker nas niso rešili, pa bi nas morali – in krivic te druge vrste bo več v primerih, ko namesto rešitve čim večjega števila življenj odločitev o tem, koga bomo rešili, manjšo ali večjo skupino ljudi, prepustimo naključju (npr. metu kovanca).
>
> **Pogodbeniški**: Ne pripišemo enakega pomena življenju vsakega posameznika, ki mu grozi oškodovanje, če se kljub dodajanju vedno novih ljudi v skupino, rešenih sprva, uravnotežena tehtnica sčasoma ne nagne na stran številčnejše skupine.

2.4 Kakšen algoritem torej?

Tri zgoraj navedene vrste legitimnih premislekov, prvič, da naj algoritem poskrbi, da bo v prometni nesreči nastalo kar najmanj škode, drugič, da naj bo porazdelitev škode oz. tveganja za oškodovanje pravična, in tretjič, da naj bodo primerljivi interesi vseh vpletenih – prima facie, tj. v odsotnosti tehtnih nasprotnih razlogov – upoštevani v enaki meri, nas pripeljejo do prvega približka etičnega načela za porazdeljevanje neizogibne škode v primerih, ko bo samovozni avtomobil zaradi tehnične okvare ogrozil življenja udeležencev v prometu.

Načelo zmanjševanja celokupne škode, prilagojeno pravičnosti (NZŠP): »Poskrbi, da bo škoda zaradi prometne nesreče, ki se ji ni dalo izogniti, čim manjša in med udeležence porazdeljena karseda pravično oz. pošteno«.

V tej abstraktni formulaciji bi moral biti NZŠP sprejemljiv tako za pristaše konsekvencializma kot tudi za deontologe. Za prve, ker lahko vlogo, ki jo nekoliko netipično odmerja premislekom o pravičnosti in poštenosti, upravičimo s sklicevanjem na pozitivne dolgoročne posledice take politike – s tem, ko zanje zmanjšamo tveganje, da bodo v prometu staknili poškodbo in/ali na cesti celo umrli, nagradimo tiste, ki spoštujejo prometne predpise,[6] na ta način pa v voznikih avtomobilov in drugih udeležencih prometa posredno okrepimo previdnost in zatremo objestnost. Za druge pa, ker se bo z zmanjševanjem števila vseh žrtev prometnih nesreč samodejno zmanjšalo tudi absolutno število teh, ki v prometnih nesrečah škodo (v skrajnem primeru izgubo življenja) utrpijo po krivici.[7]

Kljub svoji teoretični privlačnosti pa vzbuja tak predlog najmanj dve vrsti upravičenih pomislekov (zaenkrat pozabimo na tehnično, kako sploh tako abstraktno pravilo narediti opravilno). Prvič, posamezniki, ki skrbijo le zase, za svoje življenje, zdravje in premoženje, medtem ko jim je za pravičnost in poštenost bolj

[6] Pravičnost nam navsezadnje, vsaj intuitivno, nalaga (in bo enako zahtevala od algoritma v AV-ju), da v primeru, ko se trčenju ni več mogoče izogniti, lahko pa izbiramo, ali bomo trčili v avto s pripetim voznikom ali pa v takega z nepripetim voznikom, avto usmerimo v slednjega, čeravno bodo posledice predvidljivo hujše, kot če bi izbrali prvo možnost. (Furey in Hill 2020: 152)

[7] Dodatno bi lahko opredelili in razvrščali vrste škode (s človeškimi smrtmi na vrhu, ki jim sledijo težke in trajne poškodbe, nato smrti nečloveških bitij, za njimi lažje poškodbe in na koncu izključno materialna škoda), zaradi česar bi postajal algoritem čedalje bolj zapleten. Bo zaradi tega prej ali slej postal celo prezapleten, da bi še bil praktično uporaben? Ne, načeloma bi moral delovati enako kot računalniški šahovski programi, ki vsak položaj na šahovnici analizirajo in vrednotijo skozi prizmo kombinacij več izbranih strategij in le redko ene same.

malo mar ali pa se znajo prepričati, da ima njihovo preživetje upravičeno prednost pred preživetjem vseh drugih tudi z vidika morale ali pravičnosti, bodo tako pravilo odločno zavrnili. Drugič, čeprav bi marsikdo to z veseljem podprl kot meta-, drugostopenjsko načelo za uravnavanje algoritemskih izbir, lahko pričakujemo pogosta nesoglasja o tem, ali algoritemska ‚rešitev' vsakokratne prometne dileme izpolnjuje dva pogoja ali ne (oz. ju izpolnjuje bolje kot kaka druga rešitev) – na primer, ali bomo, če bomo zavili in trčili v dve starejši ženski, ki prečkata cesto pri zeleni luči, ravnali bolj krivično, kot pa če bi vztrajali v prvotni smeri in do smrti povozili dva mladostnika, ki sta cesto prečkala pri rdeči luči.

Pristranskost do sebe oz. določena stopnja sebične preudarnosti, skrbi za lastne interese (za svoje preživetje, zaščito telesne integritete itd.) – in z njo tudi dajanja prednosti le-tem pred interesi drugih – je na prvi pogled morda moralno neprivlačna, a tako kot je nesporno psihološko dejstvo, je tudi ena od osrednjih moralnih vrlin. Z moralo je seveda nezdružljivo, da bi zaščiti oz. uveljavljanju lastnih interesov dajali absolutno in/ali neulovljivo prednost, določena mera sebične skrbi pa je, četudi morda ne dobrodošla, vsaj moralno sprejemljiva. Tak uvid nas pripelje do prve in zadnje revizije algoritemskega načela za porazdeljevanje neizogibne škode:

Načelo zmanjševanja celokupne škode, prilagojeno pravičnosti in rahli pristranskosti do sebe (NZŠPS): Vedno zmanjšaj celokupno škodo (z rahlo pristranskostjo do sebe) pod pogojem, da so izpolnjene zahteve pravičnosti (»poskrbi, da se nikomur ne bo zgodila krivica oz. da se jih bo zgodilo kar najmanj«) in poštenosti (»poskrbi, da bodo vsi obravnavani kot enaki, tj. da se njihovim interesom v moralni enačbi prizna enaka, ne večja in ne manjša izhodiščna teža kot vsem drugim«).

Pot do rešitve druge uganke je nekoliko bolj vijugasta, a vseeno prevozna. O tem, katera nasilna smrt je bolj krivična, mladoletnika ali starostnika, ženske ali moškega, revnega ali premožnega, človeka ali živali, se predstave razlikujejo ne le med posamezniki, temveč tudi med kulturami. Vseh teh mnenjskih razlik se v algoritem ne bo dalo vgraditi. In se jih tudi ne bi smelo, ker v nekaterih od teh spontanih moralnih sodb evidentno odzvanjajo starizem, seksizem, rasizem, etnocentrizem in drugi iracionalni in moralno arbitrarni predsodki, ki bi jih s tem neupravičeno legitimirali. Enako nesporno je, da izvirajo po drugi strani nekatere razlike v sodbah o tem, katera porazdelitev neizogibne škode je v danih okoliščinah bolj in katera

manj pravična iz različnih pojmovanj pravičnosti, ki so enako dobro razumsko utemeljena in so izraz globokih in nepomirljivih vrednostnih razhajanj. V tej luči bo smiselno, če se namesto iskanja ene same, edino pravilne algoritemske rešitve za vsako prometno moralno enačbo, sprijaznimo z razponom možnih rešitev, ki so še znotraj polja razumno podprtega in/ali moralno sprejemljivega.

3 Politika programiranja samovoznih avtomobilov

Predpostavimo, za potrebe ovrednotenja argumenta, da nam je uspelo pravilno odgovoriti na prvo vprašanje – če je vse ostalo enako (ceteris paribus), bi bilo v primeru nakupa osebnega avtomobila edino moralno pravilno, da potrošnik izbere samovozni avtomobil, ki je opremljen z utilitarističnim (v ohlapnem pomenu besede) algoritmom. Naj država tovrstno odločitev prepusti ljudem in se sprijazni s tveganjem, da bodo ti množično dajali prednost opciji, ki maksimira njihove osebne koristi (tj. preudarnosti), pred edino moralno pravilno opcijo, ali pa jim naj za ceno posega v osebno svobodo edino pravilno izbiro vsili, tj. predpiše nakup utilitarističnih samovoznih vozil?[8] Proti drugemu predlogu govorita dva premisleka – da bi država/oblast s tem močno presegla svoje pristojnosti in pa da bi znal biti tak ukrep v luči nedavnih raziskav o odnosu ljudi do samovoznih avtomobilov kontraproduktiven.

3.1 Liberalna oblast in individualna svoboda

Kaj vse sploh velja za legitimen razlog, zaradi katerega sme oblast/vlada posameznikom omejiti svobodo odločanja in delovanja? Mnenja o tem so med političnimi filozofi in filozofinjami deljena. Moralni legalisti, recimo, vidijo eno od funkcij oblasti/države tudi v zaščiti javne morale. Zanje oblastno vsiljevanje moralno pravilne odločitve v primeru nakupa samodejnih vozil potemtakem ne bo nujno sporno. Sam bom namesto tega raje izhajal iz liberalnega filozofsko-pravnega okvirja. Znotraj le-tega je mogoče zakonske prepovedi in zapovedi upravičiti le, če je to nujno za zaščito posameznikov pred oškodovanjem in/ali krivicami, ki bi ga

[8] V pričujočem prispevku se bomo pretvarjali, da sta pojem in vrednost osebne svobode enoznačna in neproblematična. Resnica je seveda ravno nasprotna – z izrazom 'svoboda' označujemo zelo različne stvari, od odsotnosti preprek ali prisile preko zmožnosti za nekaj do odsotnosti nadvlade (angl. *domination*) in nekatere od teh stvari so že na prvi pogled bolj – in iz drugačnih razlogov – dragocene kot druge. Potrošniška svoboda, ki bi jo ogrozila morebitna zakonska prepoved prodaje in nakupa ne-UAV-jev, je vsaj na prvi pogled bolj trivialne vrste. Zaradi prostorskih omejitev se v to, sicer filozofsko, nadvse zanimivo razpravo ne bom spuščal. Lep pregled ponuja Carter (2021).

oz. jih ti utrpeli zaradi izbir in ravnanja drugih ljudi, in če bodo (edino) taki zakonski ukrepi dovolj učinkovito opravili zadano nalogo (medtem ko je blažji posegi v osebno svobodo posameznikov ne bi). Liberalno pojmovanje o nalogah/funkcijah kazenskega in civilnega prava je seveda kompleksnejše, kot pa daje slutiti zgornja poenostavitev. Liberalno-demokratična oblast sme svobodo odločanja in delovanja posameznikov z zakoni in predpisi omejiti, s političnimi in ekonomskimi ukrepi pa preusmeriti v družbeno zaželeno smer iz cele vrste razlogov: da bi ene ljudi zavarovala pred škodo, ki bi jo utegnili utrpeti zaradi odločitev ali dejanj drugih ljudi; da bi preprečila krivično (angl. *wrongful*), čeravno morda neškodljivo ravnanje oz. ljudi odvrnila od takega ravnanja; da bi poskrbela, da bodo ti, ki drugim delajo škodo ali krivico, za svoje ravnanje kazensko ali odškodninsko odgovarjali; da bi oškodovancem in žrtvam krivic zagotovila določene koristi iz naslova odškodnin in poprave krivic; in končno, da bi pravičneje porazdelila breme družbeno neželenega ravnanja. (Simpson 2012)

3.2 Legitimni cilj in sredstva prometne politike

Zakonska določba, da morajo biti vsi samovozni avtomobili opremljeni z utilitarističnim algoritmom, se lepo umešča v kategorijo regulatornih javnih politik, ki določajo pogoje oz. zarisujejo omejitve ravnanju posameznikov, organizacij in korporacij iz legitimnih razlogov, namreč zaradi izboljšanja prometne varnosti in večje prometne pravičnosti. Ne-UAV-ji bi svojim morebitnim voznikom in potnikom resda zagotovili boljšo zaščito v prometu v primerjavi z UAV-ji, a ker bi si boljšo varnost zagotovili na račun zvišanega tveganja za vse druge udeležence v prometu, bi tak aranžma ustvaril to, čemur pravijo ekonomisti negativne eksternalije – koristi za ene (delno ali v celoti) na račun stroškov za druge, ki niso imeli pri razmisleku, ali se jim bo tak vložek obrestoval, nobene besede. Če bi torej država posameznikom dovolila nakup in uporabo ne-UAV-jev, bi si težko opredeljiv, a najbrž ne nezanemarljiv delež potrošnikov (v taksonomiji/terminologiji Bergmanna et al. vsi sebičneži, ki jih je približno za petino populacije, in vsaj nekateri preklopniki) na ta način zagotovil otipljive koristi, stroške zanje pa – v obliki povečanega tveganja za v prometu povzročene poškodbe in smrt – prevalil na pleča drugih in posredno, v obliki zvišanih stroškov zdravstvene oskrbe in dodatnih smrti, na celotno skupnost. Politične skupnosti upravičeno skrbi, kadar eni ljudje s svojimi – tudi potrošniškimi – odločitvami in dejanji zvišujejo tveganja za zdravje in življenje

drugih ljudi, in jih smejo od tega odvrniti, če ne gre drugače, tudi z zakonskimi prepovedmi in omejitvami.[9] (Wolff 2011: 86)

Ukrepov, ki uravnavajo načrtovanje, oblikovanje, proizvodnjo, prodajo ali uporabo avtomobilov iz tovrstnih razlogov, ni težko najti, od zakonskih določb o maksimalnih izpustih do varnostnih standardov za rezervne dele in dodatke, rednih tehničnih pregledov in obveznega avtomobilskega zavarovanja pa vse do zaprtja mestnih središč za avtomobilski promet, omejitev hitrosti in predpisane uporabe varnostnih pasov ali čelad za motoriste. Zadnji v tej vrsti je odlok, s katerim je Evropska komisija proizvajalcem avtomobilov določila zgornjo letno mejo povprečnih emisij na proizveden avto in predpisala kazni za njeno prekoračitev. Predpis, ki bi proizvajalcem samovoznih avtomobilov nalagal vgradnjo utilitarističnih algoritmov oz. otežil ali onemogočil prodajo neutilitarističnih različic, torej ne bi v ničemer odstopal ne od siceršnje regulative in ne od ciljev, ki jih države z vrsto pravnih, fiskalnih, administrativnih in političnih ukrepov zasledujejo na področju avtomobilske industrije, cestnega prometa, splošne mobilnosti in, zakaj pa ne, tudi sprememb življenjskih navad. Sklicevanje na individualne pravice in svoboščine – da ima vsak od nas pravico izbrati avto po svojem okusu, če le ta izpolnjuje osnovne tehnične in varnostne zahteve, je v tej luči kratkovidno in naivno.[10]

Kritik vladnega vmešavanja v poslovni odnos med proizvajalcem, ki je pripravljen ne-UAV izdelati, trgovcem, ki ga je pripravljen prodajati, in potrošnikom, ki ga želi kupiti, bi lahko vztrajal, da bi bila zakonska prepoved tovrstnih transakcij še posebej moralno ali politično problematična zato, ker bi država s tem posameznikom proti njihovi volji naprtila določena tveganja, ki jih sami po tehtnem premisleku morda ne bi bili pripravljeni prevzeti nase. Toda nekaj podobnega bi lahko očitali tudi obvezni rabi varnostnih pasov ali obveznemu cepljenju zoper nevarne in hudo nalezljive bolezni – z obema ukrepoma nam država nalaga določena, resda minimalna tveganja

[9] S tem nočem reči, da so eksternalije vedno dovolj dober razlog za oblastne posege v svoboščine posameznikov ne glede na višino stroškov ali resnost tveganj, ki jih ti ustvarjajo za druge brez njihovega soglasja. Prodaja glomaznih SUV-jev je, denimo, pri nas in v večini držav po svetu dovoljena, čeprav so ti za druge udeležence v prometu dokazano bolj nevarni, praviloma pa tudi bolj potratni in okolju neprijazni kot vozila manjših dimenzij. Se pa vsa omenjena dejstva o SUV-jih sestavijo vsaj v prima facie prepričljiv razlog za ukrepe, s katerimi bi država posameznike odvrnila od nakupa tovrstnih vozil oz. jim zagrenila tako odločitev.

[10] V daljši verziji tega članka kritično obravnavam argument za svobodo delovanja in poslovanja, ki se navdihuje pri Miltonu Friedmanu (Friedman 1970) in njegovi ideji, da je skrb za varnost – tudi prometno – stvar avtonomnih posameznikov in ne vsiljive, pokroviteljske oblasti.

in nas istočasno oropa za priložnost, da jih samostojno ocenimo in kot sprejemljiva odobrimo oz. kot nesprejemljiva zavrnemo (Giubilini in Savulescu 2019). Morebitna vladna prepoved prodaje in nakupa neutilitarističnih samovoznih avtomobilov ne bi bila nič bolj sporna, kot sta zaradi posegov v individualno avtonomijo – v izhodišču, ne pa tudi po tehtnem premisleku – sporna obvezna uporaba varnostnih pasov ali obvezno cepljenje.

Doslej smo možne utemeljitve za zakonsko regulacijo trga AV-jev iskali izključno znotraj klasičnega liberalnega okvirja. Če ta okvir razširimo in državni oblasti priznamo širši nabor funkcij in nalog, s tem pa podelimo tudi širša pooblastila, se sorazmerno s tem poveča tudi nabor legitimnih razlogov za tovrstno regulacijo. Od nje si lahko, recimo, realno obetamo tudi določene moralne koristi. Če bi namreč država proizvodnjo in prodajo AV-jev z vgrajenim ne-utilitarističnim algoritmom dovolila, bi umetna inteligenca/stroji prevzeli vlogo spodbujevalca (angl. *enabler*) neetičnega obnašanja, ki je sicer brez dvoma družbeno nezaželeno, vendar pa posameznikom omogoča, da žanjejo sadove neetičnega obnašanja in se izognejo plačilu stroškov zanj – res da so na račun nekoga drugega preživeli prometno nesrečo, a ker je to zanje koristno, četudi moralno sporno odločitev sprejel nekdo drug in ne oni sami, si nimajo česa očitati. Z morebitno zakonsko prepovedjo vgradnje neutilitarističnih algoritmov v AV-je bi tovrstnemu moralnemu parazitstvu naredili konec.[11]

Če povzamem. Država/oblast ne bi presegla svojih pristojnosti, če bi se zaradi izboljšanja splošne prometne varnosti, pa tudi v imenu prometne pravičnosti, odločila kupcem avtonomnih vozil omejiti možnost izbire oz. vsiliti svojo voljo. Oba cilja sta sama po sebi legitimna in ker je prizadevanje zanju učinkovito le, če je podprto z ustreznimi sistemskimi ukrepi, tudi v pristojnosti države, tovrsten poseg pa niti približno ne bi bil ne prvi in ne edini te vrste – država je v imenu teh istih (obvezna uporaba varnostnih pasov) ali podobnih ciljev (obvezno cepljenje proti nalezljivim boleznim) že sprejemala in še vedno sprejema pravne, politične,

[11] Tezo o moralnih koristih, ki si jih lahko obetamo od vsiljevanja utilitarističnih algoritmov, lahko razumemo na dva načina, v močnejšem smislu kot tezo, da bodo ti izboljšali človeški značaj, ali v šibkejšem smislu kot tezo o izboljšanju človeškega obnašanja. Za obvezno vgradnjo utilitarističnih algoritmov bi se utegnilo izkazati, kar Giubilini in Savulescu 2019 ugotavljata o predpisani uporabi varnostnega pasu – namreč da se je kljub začetnemu odporu v zgolj nekaj letih iz eksternega zakonskega predpisa spremenila v ponotranjeno družbeno normo. V tem primeru bi bile moralne koristi od obvezne vgradnje utilitarističnih algoritmov globoke in trajne. A tudi če bo omenjeni ukrep ljudem pomagal zgolj bolje izpolniti njihove moralne dolžnosti, moralne koristi od njega ne bodo zanemarljive. Več o moralnem dometu zakonodajnih ukrepov glej Thomson (1986) in Hunt (2014).

ekonomske itd. ukrepe, ki v imenu javnega dobrega ožijo prostor individualne svobode.[12] Pa bi bila taka odločitev tudi zares pametna, preudarna? Rezultati nedavnih raziskav dajejo slutiti, da bi se znala vrniti kot bumerang.

3.3 Ko bi le bili vsi ljudje nesebični … razen mene!

Bonnefon et al. so izvedli vrsto eksperimentov, da bi proučili odnos ljudi do AV-jev obeh vrst, utilitaristične in neutilitaristične. Pri tem so se dokopali do nekaterih nadvse zanimivih ugotovitev. V vseh študijah je večina udeležencev izrazila moralno preferenco za AV-je, ki žrtvujejo svoje potnike, da bi rešili večje število pešcev. Ta moralna preferenca je bila močna v okoliščinah, ko so si udeleženci zamišljali sebe v AV-ju v družbi sodelavca, družinskega člana ali lastnega otroka. V tej luči je presenetljivo, da udeleženci niso izrazili primerljive preference tudi za nakup utilitarističnega AV-ja, še zlasti ne, ko so si morali v vlogi sopotnikov zamisliti družinske člane. Za povrh udeleženci niso bili naklonjeni ideji, da bi oblasti predpisale utilitaristične algoritme za AV-je, in so dali vedeti, da se najverjetneje ne bi odločili za nakup AV-ja, če bi bili vsi tovrstni avtomobili brez izjeme opremljeni s takimi algoritmi.

3.4 Resnična ali zgolj navidezna družbena dilema?

Tabela 1: Individualna odločitvena matrika (problem koordinacije)

	Večina drugih ljudi (se) vozi (v) neutilitaristične/ih samovozne/ih avtomobile/ih	**Večina drugih ljudi (se) vozi (v) utilitaristične/ih samovozne/ih avtomobile/ih**
jaz (se) vozim (v) ne-utilitaristični/em samovozni/em avtomobil/u	(1) Zame tretja najboljša možnost (z vidika splošne prometne varnosti pa najslabša možnost)	(2) Zame najboljša možnost
Jaz (se) vozim (v) utilitaristični/em samovozni/em avtomobil/u	(3) Zame najslabša možnost	(4) Zame druga najboljša možnost (z vidika prometne varnosti najboljša možnost)

[12] Da bi šlo pri zakonski prepovedi prodaje in nakupa ne-UAV-jev za razmeroma blag poseg v svobodo delovanja in poslovanja posameznikov in organizacij, nakazuje še en premislek. Nekateri teoretiki verjamejo, da država bolj drastično poseže v svobodo posameznikov, če iz obstoječega nabora možnih izbir eno (recimo nakup in uživanje tobačnih izdelkov) odstrani, kot pa če prepreči, da bi ta ali oni ponudnik izdelkov ali storitev obstoječemu naboru možnih potrošniških izbir eno dodal. Prikrajšanje za določeno opcijo naj bi bilo, skratka, vedno lažje opravičiti kot odvzetje taiste. Takšno stališče ni neproblematično, kot v svoji kritiki lepo pokaže Schmidt (2016), a če ga pogojno sprejmemo, iz njega izhaja, da bo v primeru zakonske regulacije algoritmov v AV-jih letvica za upravičenje državne intervencije postavljena razmeroma nizko, ker pač potrošnikom s tem možnost izbire ne-UAV-ja ne bo odvzeta (angl. *withdrawn*), temveč bodo zanjo zgolj prikrajšani (angl. *withheld*).

Nakup in uporaba neutilitarističnega samovoznega avtomobila je, skratka, z vidika posameznika enoznačno dominantna in v toliko najbolj preudarna strategija. Pa vendar, če bo vsak poskušal povečati osebno korist na račun drugih – in upravičeno pričakujemo, da se bomo tako obnašali vsi ali vsaj večina med nami – bomo na koncu dobili kolektivno/družbeno neoptimalen izid. Vsaj na prvi pogled smo se torej ujeli v začaran krog:

> (i) Nakup in raba UAV je najbolj zaželena izbira tako z moralnega kot z družbenega vidika (in velika večina ljudi se s to oceno celo strinja).
> (ii) Kljub temu bi večina ljudi, če bi to odločitev prepustili njim, namesto tega zase raje kupila in uporabljala ne-UAV.
> (iii) Če bi država ponudbo AV-jev z zakonom slučajno omejila zgolj na UAV-je oz. če bi dovolila zgolj prodajo AV-jev z utilitarističnim algoritmom, pa bi ljudje celo raje izbrali kaj drugega, recimo ne-AV.

Izsledki Bonnefonove študije snovalce prometne politike vsaj na prvi pogled postavljajo pred nerešljivo družbeno dilemo: naj žrtvujemo svobodo, da bi izboljšali varnost, ali naj raje žrtvujemo varnost, da bi ohranili svobodo? Toda ali imamo pri programiranju samovoznih avtomobilov in njihovi promociji res opraviti z zagatnim problemom koordinacije, kjer bi vsak posameznik najraje videl, da vsi drugi storijo, kar je prav in družbeno koristno, tj. optimalno z vidika seštevka interesov vseh vpletenih, zase bi hotel pa izpogajati izjemo od tega pravila? Izsledki nekaterih drugih raziskav o pripravljenosti ljudi, da se zaradi preživetja večjega števila drugih ljudi sprijaznijo z lastnimi poškodbami ali celo smrtjo, so nekoliko bolj optimistični.[13] Raziskava, ki so jo izvedli Bergmann in sodelavci, je, recimo, z uporabo podobnih prometnih scenarijev med udeleženci odkrila »presenetljivo visoko stopnjo pripravljenosti, da žrtvujejo sebe, da bi rešili druge«. Natančneje:

> /…/ udeleženci so delovali bolj altruistično od pričakovanj, čeprav so rezultati v dani situaciji odvisni od števila potencialnih žrtev. V več kot polovici poskusov so bili udeleženci pripravljeni rešiti celo samo dve osebi. S povečevanjem števila ogroženih ljudi raste tudi delež tistih, ki so se pripravljeni žrtvovati zanje. Medtem ko so bili za rešitev dveh oseb pripravljeni umreti v 52 odstotkih primerov, je pri treh rešenih življenjih ta

[13] In sploh niso edini, ki Bonnefonu očitajo preveč pesimistično interpretacijo zbranih podatkov. Glej npr. Greene 2016 in Hübner in White 2018.

> delež narastel na 57 in pri štirih na 63 odstotkov. Pri skupinah petih, šestih in sedmih oseb so bili rezultati precej podobni, za toliko ljudi se je bilo pripravljenih žrtvovati približno 70 odstotkov anketirancev. Rezultati podpirajo zaključek o obstoju različnih skupin odločevalcev, ki sledijo vsaka svoji strategiji. Prva skupina se je dosledno odločala za lastno preživetje na račun drugih in jo lahko zato označimo za (moralne) egoiste. Druga skupina, recimo ji altruisti, je, nasprotno, dosledno izbirala možnost samožrtvovanja. Tretji skupini bi lahko rekli preklopniki, ker so člani svoje odločitve uravnavali po velikosti skupine, pri čemer so se odločili za samoohranitev v preizkušnjah z manj in za samožrtvovanje v preizkušnjah z več ogroženimi ljudmi. (Bergmann et al. 2018)

Da lahko ljudi po pripravljenosti, da se žrtvujejo za druge, razvrstimo v eno od treh skupin, med sebičneže, nesebičneže in tiste vmes, ne preseneča. Bolj so v luči Bonnefonove pesimistične študije presenetljivi – in razveseljivi – sorazmerni deleži vsake od njih. Če bi jih z veliko mero poenostavljanja neposredno prevedli v nakupne odločitve (pa jih seveda ne moremo), bi si lahko obetali, da se bo nekje med polovico in dvema tretjinama kupcev odločila za nakup AV-ja z utilitarističnim algoritmom. Celo ob upoštevanju dejstva, da se marsikatera v anketi zabeležena sodba o tem, kateri razplet je bolj in kateri manj moralno sprejemljiv in/ali pravičen, nazadnje ne bo prevedla v ustrezno potrošniško odločitev, najbrž drži, da so Bonnefon in kolegi močno podcenili delež tistih, ki bodo izbrali samovozni avtomobil z utilitarističnim algoritmom prostovoljno, se pravi brez oblastne palice in korenčka.

3.5 Kontraproduktivna prometna politika?

Kaj pa če je Bonnefonova skupina vendarle natančneje izmerila (brez)srčni utrip javnosti/ljudstva kot Bergmannova? Za oblikovalce utilitarističnih politik bo v tem primeru najpametneje, da se odrečejo vsiljevanju utilitarističnih samovoznih avtomobilov. Kar je paradoksno. Vendar pa je tudi ta paradoks bolj navidezen kot resničen – zavajajoče je namreč imenovati avtomobile, ki zmanjšajo število smrtnih žrtev/škodo v posameznih primerih, vendar pa zaradi svoje nepriljubljenosti pri potrošnikih ne bodo zmanjšali splošne škode zaradi prometnih nesreč, 'utilitaristični'. V ugotovitvi, da je politika, za katero bi pričakovali, da bo po utilitarističnih kriterijih moralno optimalna, namreč spodbujanje nakupa in rabe

izključno utilitarističnih AV-jev, potem ko seštejemo vse, tudi dolgoročne učinke tovrstne politike, šele na drugem mestu za nakupom in uporabo AV-jev, ki sicer zmanjšujejo škodo, a ne za ceno ogrožanja varnosti potnikov, ni v resnici nič paradoksnega. Z utilitarističnega vidika bo v takem primeru bolje, da nekateri ljudje kupujejo in uporabljajo avtomobile, ki bi jih bilo z utilitarističnega vidika bolje ne uporabljati. Če se bo izpolnila napoved, da bodo avtonomna vozila močno oklestila število nesreč in posredno smrtnih žrtev na cestah, hkrati pa bodo utilitaristični samovozni avtomobili med potrošniki izrazito nepriljubljeni, bi lahko vlada z njihovim agresivnim vsiljevanjem zamaknila uvedbo na splošno varnejših avtonomnih vozil. Zato se bomo morda prisiljeni iz utilitarističnih razlogov – da bi kar se da hitro in občutno zmanjšali krvni davek na naših cestah – odpovedati vsiljevanju 'utilitarističnih' AV-jev in sprijazniti z drugo najboljšo možnostjo, nakupom in rabo/prevlado neutilitarističnih, a v primerjavi z neavtonomnimi vozili še vedno veliko varnejših AV-jev.

3.6 Družbena trilema?

Prvoten vtis o paradoksu, v katerega naj bi se neizogibno zapletla na utilitarističnem kalkulu temelječa politika, je zmoten še iz enega razloga. Opisani družbeni in psihološki parametri namreč ne tvorijo dileme med a) svobodo potrošniške izbire in b) prometno varnostjo, temveč trilemo med (a) svobodo izbire, (b) (prometno) pravičnostjo in (c) (prometno) varnostjo/dobrobitjo. Ta spominja na trilemo, ki nas je pestila med pandemijo COVID-19, ko smo bili prisiljeni izbirati med tremi vrednotami: (a) svobodo/svoboščinami, (b) enakostjo/pravičnostjo in (c) varnostjo/dobrobitjo, in smo lahko v najboljšem primeru ohranili dve od navedenih treh. Glede na to, kar smo vedeli o virusu in bolezni, ki jo je ta povzročal, so nam bile namreč na voljo naslednje osnovne opcije: (i) življenje brez slehernih omejitev, vendar za ceno nesprejemljivo visoke stopnje okužb, bolezni in smrti (torej svobodo za vse in enako obravnavo vseh, a za ceno zmanjšane varnosti oz. dobrobiti), (ii) življenje s selektivnimi, ciljanimi ukrepi, ki bi nesorazmerno prizadeli določen del populacije, recimo ljudi z večjim tveganjem za okužbo in/ali bolezen oz. tiste, ki v večji meri ogrožajo druge ljudi (razmeroma dobra zaščita zdravja/dobrobiti vseh, a za ceno neenake obravnave, tj. svobode za večino in nesvobode za manjšino), in (iii) vsesplošno in vseobsegajoče zaprtje/lockdown, ki vse ljudi razmeroma dobro zaščiti pred okužbo, boleznijo in smrtjo in kjer nas oblast vse obravnava enako (slabo, bi

pripomnil cinik), a za varnost in enakopravnost plačujemo visoko ceno v obliki vsesplošnega suspenza pravic in svoboščin. (Pugh et al. 2021)

Podobno kot za pandemijo covida-19 bi se lahko tudi za prihajajočo avtomatizacijo prometa zdelo, da nas namesto z družbeno dilemo sooča z družbeno trilemo, tj. s prisilno izbiro med tremi konkurenčnimi/konfliktnimi (skupinami) vrednot oz. načel: (a) osebno svobodo;[14] (b) prometno varnostjo in (c) prometno pravičnostjo. Na voljo imamo vrsto rešitev, ki ohranjajo po dve od teh treh vrednot, a med njimi ni niti ene, ki bi nam pomagala ohraniti vse tri hkrati – izbiramo lahko med (i) prometno varnostjo in pravičnostjo, a za ceno žrtvovanja svobodne potrošniške izbire; (ii) svobodno potrošniško izbiro in prometno varnostjo, a na račun okrnjene in/ali zamaknjene prometne pravičnosti; ter (iii) svobodno potrošniško izbiro in prometno pravičnostjo, a za ceno poslabšanja prometne varnosti.[15]

Na srečo je tudi opisana trilema bolj namišljena kot resnična. Množična raba AV-jev, od katere si – upravičeno ali ne – obetamo pomembno izboljšanje prometne varnosti, namreč ni nujno navzkriž s prometno pravičnostjo. V kombinaciji s potrošniško svobodo bo, če ima Bonnefon prav in se Bergmann moti, res zmanjšala in zakasnila prometno pravičnost (v njenem maksimalnem možnem obsegu), a če drži, da bo ob množični rabi AV-jev žrtev prometnih nesreč in škode zaradi njih neprimerno manj kot sedaj, bo po matematični nujnosti neprimerno manj tudi žrtev prometnih nesreč, ki jih bodo smrt, pohabljenje, težje poškodbe in podobne oblike oškodovanja doletele po krivici. Kar bomo prisiljeni žrtvovati, če prepustimo izbiro algoritma, s katerim bo opremljen njihov AV, samim kupcem, ni toliko prometna pravičnost nasploh kot bolj en – težko izmerljiv – del le-te. V določenem smislu torej drži, da bi bilo mogoče hkrati zagotoviti vse troje: udeležencem v prometu varnost, vpletenim v prometne nesreče pravičnost in potrošnikom svobodno izbiro.[16]

[14] Če bi jo hoteli trivializirati, bi rekli ‚svobodo potrošniške izbire', a moralni nazori, ob katere bi znal trčiti utilitaristični algoritem in zaradi katerih ga bo marsikdo odklanjal, tvorijo trdno jedro vsakega svetovnega nazora in so kot taki seveda vse prej kot trivialni.

[15] Opcija (i) je med vsemi najmanj sporna, ustreza namreč popolni zakonski prepovedi nakupa ne-UAV. Kaj pa drugi dve? Za kombinacijo (ii) smo že ugotavljali, da ji ustreza scenarij pospešene proizvodnje, prodaje in rabe obeh vrst AV-jev, utilitarističnih in neutilitarističnih, ob predpostavki, da ti neizogibne škode še ne bodo sposobni odmerjati in razporejati na podlagi natančno odmerjene odgovornosti zanjo. Opcija (iii) se zdi med vsemi najbolj neverjetna – težko si je namreč zamisliti okoliščine, ko bi prosta prodaja in množična uporaba neutilitarističnih AV-jev izboljšala prometno pravičnost na račun poslabšanja prometne varnosti oz. povzročila rast skupnega števila žrtev prometnih nesreč, ne da bi se sočasno povečalo število tistih, ki jo v prometnih nesrečah skupijo po krivici.

[16] Kar ponujam tukaj, je bolj oris argumenta kot njegova dokončna podoba. Za oceno vpliva, ki ga bodo posamezni ukrepi imeli na prometno pravičnost, bi potrebovali vsaj približno oceno, v koliko prometnih situacijah neizogibne

3.7 Pravična porazdelitev neizogibne škode?

Do zdaj smo bore malo rekli o merilih za presojo, ali bo določena porazdelitev neizogibne škode oz. tveganj zanjo v prometnih nesrečah pravična ali ne. Je sploh smiselno trditi in dokazovati, da vodijo ene reakcije AV-jev v tem ali onem od trinajstih tipskih scenarijev Moralnega stroja do pravičnejšega izida kot druge? In vztrajati, da moralni enačbi brez te spremenljivke nekaj manjka in je moralni izračun zato napačen? Nikomur od nas seveda ni tuj pravičniški bes ali nejevolja, ki nas tipično obide ob prebiranju novinarskih poročil o prometnih nesrečah, v katerih jo vinjeni povzročitelji prometnih nesreč odnesejo brez prask, ne krivi in ne dolžni pešci ali kolesarji ali potniki v vozilu pa njihovo objestnost plačajo z življenjem. Toda v tipičnih prometnih scenarijih, ki jih preigrava Moralni stroj in za potrebe katerih se bo programiralo avtonomna vozila, prometno nesrečo brez izjeme zakrivi okvara stroja in ne objestnost ali kaka druga človeška hiba. Se je v takih okoliščinah sploh smiselno spraševati, kdo od vpletenih si bolj in kdo manj zasluži preživeti oz. umreti ali kdo med njimi je morda upravičen do prednostne zaščite pred škodo, ki se ji ni mogoče izogniti? In če je pri trkih avtonomnih vozil za nesrečo nesmiselno kriviti ljudi, na kaj drugega bi se lahko oprla sodba, da bo, recimo, smrt sopotnika v takem vozilu manj krivična kot smrt pešca, v katerega je to vozilo trčilo na pločniku? Na tem mestu se pri stranskih vratih v razpravo vrača vprašanje odgovornosti, ki smo ga v uvodu postavili na stranski tir v upanju, da bo to moralno presojo maksimalno poenostavilo. Kakorkoli kompleksna in zamudna že presoja o pravični porazdelitvi škode je, očitno se ji ne da zares izogniti.

Na tem mestu lahko ponudimo kvečjemu oris take teorije. Pri tem izhajamo iz predpostavke, da morala v izhodišču vse ljudi v enaki meri varuje pred oškodovanjem ali tveganjem zanj, da pa se ta izvorna moralna simetrija podre, kadar je kdo od vpletenih odgovoren za nastanek tveganih okoliščin oz. neizogibne škode, in da ta pogoj izpolnimo, kadar, recimo, prostovoljno sodelujemo v dejavnostih, za

škode bo pravičnost potegnila ta kratko – bodisi zato ker bo neutilitaristični AV, katerega nakup smo potrošnikom omogočili v imenu svobode, absolutno prednostno zavaroval potnike na račun oškodovanja – večjega števila – drugih udeležencev (potnikov v drugem avtu, kolesarjev, motoristov, peščev, itd.) ali pa ker, če si sposodim Kaupinnenov pomislek, v AV-je vgrajeni algoritmi ne bodo znali pravilno odmeriti odgovornosti za škodo (angl. *liability to harm*) za v nesrečo vpletene posameznike in v skladu z njo mednje porazdeliti neizogibne škode. Ta drugi potencialni vzrok za ohranjanje prometnih krivic nas po pravici povedano niti pretirano ne skrbi, ker skrb za pravično porazdelitev škode med vse v nesrečo vpletene najbrž že zdaj ni ravno med poglavitnimi dejavniki, ki bi usmerjali reakcije voznikov v kočljivih situacijah, in je zato malo verjetno, da bi AV-ji zaradi algoritmov, ki bodo v prvi fazi odgovornost za oškodovanje ali smrt in z njo neizogibno škodo odmerjali nenatančno in nepravično, ta rezultat bistveno poslabšali.

katere bi lahko predvideli (ali bi celo morali predvideti), da bodo ustvarila taka tveganja. Če smo sebe in druge spravili v nevarnost zavestno ali z brezobzirnim ravnanjem, bi morali sami nositi poglaviten delež tako nastalega tveganja oz. škode. Glede tega ni razhajanj. Kaj pa če prometno nesrečo povzroči AV, ki smo ga kupili in uporabljali dobro vedoč, da mu lahko odpovedo zavore ali krmilni mehanizem in da bo v tem primeru ogrozil življenja drugih udeležencev v prometu – smo odgovorni za to, da ta naš AV ogroža življenja drugih ljudi, in smo posledično izgubili pravico do varstva pred oškodovanjem, taisto pravico, ki so jo vse druge potencialne žrtve v teh okoliščinah ohranile? Ter se nam zato ne bi zgodila krivica, če bi se v taki prometni nesreči poškodovali ali umrli, bi pa bilo krivično, če bi v njej namesto nas izgubil življenje kdo drug?

To so vse prej kot lahka vprašanja. Še zlasti ker med filozofi_njami ni soglasja o pogojih, ki morajo biti izpolnjeni za brezprizivno oškodovanje oz. uboj (angl. *liability to being harmed/killed*). Najmanj, kar lahko rečemo, je, da potniki v samovoznem avtomobilu, ki je zaradi okvare povzročilo prometno nesrečo, v takem primeru definitivno ne bodo uživali absolutne prednosti pri zaščiti pred nastalo škodo. In da bi bilo zato krivično, če bi jih algoritem ne glede na druge okoliščine vedno in brez izjeme prednostno zaščitil pred njo. Kaupinnen (2021) celo verjame, da bi bili v vrsti kandidatov za oškodovance na prvem mestu. Razlog za to je preprost – čeravno ne bodo neposredno *krivi* za okvaro AV-ja in posledično nesrečo, ki jo bo le-ta povzročil, bo vendarle res, da so kupili in uporabljali tak avto prostovoljno in ob polni zavesti, da se kaj takega lahko pripeti, kljub temu da bi bili lahko ravnali drugače (tj. kupili kak drug avto ali pa se nakupu avtomobila odrekli), zaradi česar bodo za nesrečo *minimalno moralno odgovorni* in se posledično ne bodo mogli sklicevati na enako pravico do varstva pred oškodovanjem kot drugi, naključni udeleženci. Omenjeni Kaupinnenov sklep seveda ni neproblematičen, sloni namreč na dveh vse prej ko samorazvidnih domnevah: a) da za brezprizivnost oškodovanja zadošča že prostovoljna udeležba v aktivnosti, za katero vemo, da bi lahko prispevala k situaciji neizogibne škode in b) da temu opisu ustrezata že sam nakup in raba AV-ja. A ne glede na to, ali pravičnost v tovrstnih okoliščinah res *zahteva* prednostno žrtvovanje potnikov pred drugimi udeleženci v prometu ali ne, prej našteti razlogi gotovo podpirajo vsaj blažji sklep, namreč da bi bila brezpogojna prednostna zaščita, kakršno bi svojim potnikom zagotavljal neutilitaristični algoritem, krivična.

Avtomatizacija prometa in prometnih sredstev bo zmanjšala število prometnih nesreč in žrtev le-teh, ne bo pa nujno poskrbela, da bodo vedno rešena tista življenja, ki bi jim moralno gledano morali dati prednost. Če tehnologija ne bo kos nalogi, da odmeri odgovornost posameznih udeležencev v prometu, bo občutnemu deležu udeležencev v prometu še naprej kratena pravica do učinkovitega varstva pred oškodovanjem. Zato je nujno najti razumen kompromis med zmanjševanjem škode in njenim bolj pravičnim porazdeljevanjem. To bo za prihodnjo avtomatizacijo prometa verjetno resnejši izziv kot ta, ki so ga identificirali Boneffon in njegova ekipa.

3.8 Nazaj k družbeni dilemi … in orisu njene rešitve

Ko grobo skico, ki jo ponujajo Bonnefon et al., dopolnimo s podrobnostmi, se prične sprva oster kontrast med realnostjo in idealom/pričakovanji mehčati. Kot prvo, Bonnefon in njegova ekipa so v orisu družbene dileme dvojno pretiravali – moralno junaštvo so zamešali z moralno obveznostjo (rahla pristranost do sebe je namreč v situacijah neizogibne škode moralno dopustna), obenem pa so podcenili pripravljenost ljudi na samožrtvovanje, ko jim tako obveznost po lastni presoji nalaga morala. Ljudje bodo bolj pripravljeni kupovati in uporabljati zmerno utilitaristične AV-je, torej vozila, ki zagotavljajo nekoliko več varnosti svojim potnikom, a ne na račun pravičnosti in/ali bistveno višjega tveganja za druge udeležence v prometu. Razumen kompromis med preudarnostjo oz. sebično skrbjo na eni ter moralnostjo oz. nesebičnostjo na drugi strani se glasi: »brez absolutne prioritete potnikov«, a tudi »brez absolutne prioritete za zmanjševanje škode«. Razkorak med privlačnostjo UAV-jev in ne-UAV-jev bi se dalo, skratka, dodatno zmanjšati s priznanjem legitimnosti skrbi zase in svoje bližnje ali ustrezno prilagoditvijo algoritma razumno utemeljenim pričakovanjem in sodbam ljudi. Vrzel med vozilom, ki bi mikalo večino potrošnikov, in vozilom, ki je politično in moralno gledano najboljša opcija, bi lahko dodatno premostili z vključevanjem takih kulturno specifičnih pomislekov in prednostnih nalog v algoritem, ki jih je mogoče vsaj v grobem razumno utemeljiti (glej Pölzler 2021). Družbena dilema se tako ali sploh ne bi pojavila ali pa bi bila razmeroma enostavno rešljiva.

Najresnejši očitek Bonnefonu in njegovi ekipi pa smo prihranili za konec – izvorno zastavljena dilema med zakonsko prepovedjo in svobodno potrošniško izbiro je … lažna. Omenjeni so namreč spregledali ali zavestno zamolčali celo vrsto alternativ brezpogojni zakonski prepovedi. Države se lahko v prizadevanju po izboljšanju prometne varnosti in/ali pravičnosti zatečejo k celi paleti pravnih in političnih ukrepov, ki so blažji in zato moralno in politično manj problematični kot zakonska prepoved, ki potrošnike oropa svobodne izbire: od državnih subvencij za kupce UAV-jev do nižjega cestninskega davka. Za vlade liberalno-demokratičnih držav že dolgo velja, da se družbeno nezaželenega obnašanja (kajenja in drugih škodljivih razvad, pornografije, sovražnega govora, ipd.) lotevajo s kombinacijo zakonskih, političnih, fiskalnih in izobraževalnih ukrepov. Nekateri od teh sodijo v kategorijo t. i. dregljajev (angl. *nudges*). Dregljaji so intervencije, s katerimi poskušamo na predvidljiv način vplivati na odločitve in ravnanja drugih ljudi, ne da bi bistveno omejili njihovo možnost izbire ali pa jim izbiro družbeno nezaželenih opcij zagrenili tako, da bi jo obremenili s pretiranimi stroški in tako naredili neprivlačno, temveč namesto tega s spremembo 'arhitekture izbire' – uokvirjenjem (angl. *framing*) ponujenih opcij, prikrajanjem njihovega vrstnega reda, določitvijo osnovne in dodatnih opcij, ipd.[17] (Thaler in Sunstein 2008) V zadnjem času se tovrstne intervencije množijo v vladnih kampanjah za krepitev javnega zdravja in spodbujanje zdravega življenjskega sloga, čedalje pogosteje pa jih je mogoče zaslediti tudi v prizadevanjih po dvigu prometne varnosti (merilci hitrosti ob cestah, zvočni signal, ki nas v avtu opozarja, da nismo pripeti, razni dodatki za varno vožnjo, ipd.). Država lahko, skratka, potrošnike, ki so do nakupa UAV-jev morda zadržani, k tej družbeno (in navsezadnje tudi individualno) zaželeni odločitvi spodbudi na različne načine, od moralnega uokvirjenja izbire med UAV-ji in ne-UAV-ji preko izbire UAV-ja kot osnovne (in ne-UAV-jev kot dodatne) nastavitve do obremenitve nakupa in uporabe ne-UAV-jev z materialnimi in nematerialnimi stroški, med slednjimi tudi, zakaj pa ne, z družbeno stigmo.

Ker dregljaji, kadar seveda delujejo, svojo učinkovitost dolgujejo znanim človeškim iracionalnostim (pristranskostim, hevristikam, nevednosti, lenobi ipd.), bi se država na ta način izpostavila očitku, da s tovrstnimi pozitivnimi spodbudami za nakup UAV-jev in negativnimi spodbudami proti nakupu ne-UAV-jev nedopustno manipulira s svojimi državljani. A takega dobronamernega državnega paternalizma

[17] Locus classicus je seveda knjiga Richarda Thalerja in Cassa Sunsteina *Dregljaj: kako izboljšati naše odločitve o zdravju, premoženju in sreči* iz l. 2008.

v celoti gledano ne bo težko moralno opravičiti, če bodo le dregljaji a) transparentni, ne prikriti, b) naslovljeni primarno na sistem 2 (zavestne procese tehtanja in premlevanja razlogov) in šele sekundarno, če sploh, na sistem 1 (avtomatske, podzavestne procese), ter končno c) namenjeni opolnomočenju posameznikov in ne podtikanju tuje jim volje. Spodbujanje nakupa in/ali rabe UAV-jev – oz. sočasno odvračanje od nakupa in/ali rabe ne-UAV-jev – bi tako postalo lep vzgled za moralno opravičljivo dreganje, saj država ne bi na moralno sporen način zaobšla ali spodkopala potrošniške svobode in/ali avtonomije; nasprotno, s tem, ko bi prilagodili potrošniško ponudbo njihovim pristnim (v raziskavah razkritim) moralnim vrednotam, bi po svoje zavarovali njihovo ogroženo avtonomijo.

4 Namesto zaključka

V pričujočem prispevku sem dokazoval, da obstajajo tehtni razlogi, da v prihodnje AV-je vgradimo zmerno – pravičnosti in rahli pristranskosti do sebe prilagojeno – utilitaristično načelo za porazdelitev neizogibne škode. Če imajo Bonnefon in njegova ekipa prav, pa država takih algoritmov kupcem AV-jev ne bo smela vsiljevati z zakonom, temveč bo z vidika izboljšanja prometne varnosti in pravičnosti smotrneje, da proizvodnjo in prodajo UAV spodbuja z mehkejšimi prijemi, od nakupom utilitarističnih AV-jev naklonjene davčne, cenovne in zavarovalne politike preko oglaševalskih kampanj do blagih dregljajev oz. premišljene, a hkrati transparentne arhitekture potrošniških izbir.

Viri in literatura

Awad, E., Dsouza, S., Kim, R. et al. (2018). »The Moral Machine experiment«. *Nature*, 563, str. 59–64.

Bergmann L.T. et al. (2018). »Autonomous Vehicles Require Socio-Political Acceptance—An Empirical and Philosophical Perspective on the Problem of Moral Decision Making«. *Frontiers in Behavioral Neuroscience*, 12, 31.

Bonnefon, J., Shariff, A. in Rahwan, I. (2016). »The Social Dilemma of Autonomous Vehicles«. *Science*, 352(6293), str. 1573–1576.

Carter, I. (2021). »Positive and negative liberty«. V Zalta, E. N. (ur.), *Stanford Encyclopedia of Philosophy* (izdaja pomlad 2022). URL = https://plato.stanford.edu/archives/spr2022/entries/liberty-positive-negative/.

Friedman, M. (1970). »The social responsibility of business is to increase its profits«. *The New York Times*, 13. september 1970.

Furey, H. in Hill, S. (2021). »MIT's moral machine project is a psychological roadblock to self-driving cars«. *AI Ethics*, 1, str. 151–155.

Giubilini, A in Savulescu, J. (2019). »Vaccination, risks, and freedom. The seat belt analogy«. *Public Health Ethics*, 12(3), str. 237–249.

Greene, J. (2016). »Our driverless dilemma«. *Science*, 352(6293), str. 1514–1515.
Hübner, D. in White, L. (2018). »Crash Algorithms for Autonomous Cars: How the Trolley Problem Can Move Us Beyond Harm Minimisation«. *Ethical Theory and Moral Practice*, 21, str. 685–698.
Hunt, L. H. (2014). »On improving people by political means«. V: LaFollette, H. (ur.), *Ethics in Practice. An Anthology* (4. izdaja). Oxford: Wiley Blackwell, str. 299–308.
Kamm, F. M. (1985). »Equal treatment and equal chances«. *Philosophy & Public Affairs*, 14(2), str. 177–194.
Kauppinen, A. (2019). »Who's afraid of Trolleys?«. V Suikkanen, J. in Kauppinen, A. (urd.), *Methodology and Moral Philosophy*. London: Routledge, str. 49–71.
Kauppinen, A. (2021). »Who should bear the risk when self-driving vehicles crash?«. *Journal of Applied Philosophy*, 38(4), str. 630–645.
Keeling, G. (2020). »Why Trolley Problems Matter for the Ethics of Automated Vehicles«. *Science and Engineering Ethics*, 26(1), str. 293–307.
Köbis, N., Bonnefon, JF. in Rahwan, I. (2021). »Bad machines corrupt good morals«. *Natural Human Behavior*, 5, str. 679–685.
Lin, P. (2013). »The ethics of autonomous cars«. *The Atlantic*, 8. oktober 2013. URL = https://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/.
Parfit, D. (1978). »Innumerate ethics«. *Philosophy & Public Affairs*, 7(4), str. 285–301.
Poročilo skupine A.T. Kearney. n.d. »How automakers can survive the self-driving era«. URL: https://www.es.kearney.com/automotive/article?/a/how-automakers-can-survive-the-self-driving-era.
Pölzler, T. (2021). »The Relativistic Car: Applying Metaethics to the Debate about Self-Driving Vehicles«. *Ethical Theory and Moral Practice*, 24, str. 833–850.
Pugh, J. et al. (2021). »Learning to live with Covid-19 – the tough choices ahead«. *The Conversation*, 16. april 2021. URL = https://theconversation.com/learning-to-live-with-covid-the-tough-choices-ahead-158992.
Rinni, R. (2017). »Raising good robots«. *AEON*, 19. april 2017. URL = https://aeon.co/essays/creatingrobotscapableofmoralreasoningislikeparenting.
Scanlon, T. (1998). *What Do We Owe To Each Other*. Cambridge, Mass.: Harvard University Press.
Schmidt, A. (2016). »Withdrawing vs. withholding freedoms: the case of tobacco and nudging«. *The American Journal of Bioethics*, 16(7), str. 3–14.
Taurek, J. (1977). »Should the numbers count?«. *Philosophy & Public Affairs*, 6(4), str. 293–316.
Thaler, R. in Sunstein, C. (2008). *Nudge. Improving Decisions About Health, Wealth and Happiness*. New Haven in London: Yale University Press.
Thomson, J. (1986). »Some questions about government regulation of behaviour«. V *Rights, Restitution & Risk. Essays in Moral Theory*. Cambridge, Mass. in London: Harvard University Press; str. 154–172.
Wolff, J. (2011). *Ethics and Public Policy. A Philosophical Inquiry*. London in New York: Routledge.

AVTONOMNA OROŽJA IN EROZIJA MORALNE ODGOVORNOSTI: SISTEMATIZACIJA NASILJA IN IZGINJANJE PRAVIČNOSTI

TOMAŽ GRUŠOVNIK
Univerza na Primorskem, Pedagoška fakulteta, Koper, Slovenija
tomaz.grusovnik@pef.upr.si

Sinopsis Poglavje skuša pokazati, da avtonomna orožja predstavljajo resno nevarnost za erozijo moralne odgovornosti v sodobnem vojskovanju. Tako psihologija kot zgodovina vojskovanja kažeta, da je človeku lasten odpor do ubijanja sovražnih vojakov, tako imenovani 'nizek nivo zadetkov'. Ta ovira ubijanja sovražnikov je bila v veliki meri presežena s fizično razdaljo med vojaki in njihovimi tarčami, tj. s pomočjo artilerije, bombardiranja in brezpilotnikov. Avtonomna orožja pa ne le dodatno večajo to razdaljo, pač pa jo prestavljajo na povsem novo raven – distanca zdaj postane konceptualna, saj je odločitev za ubijanje prenesena na tovrstne sisteme. Ta vrzel odgovornosti ima za posledico tako nižanje praga za uporabo sile, zaradi česar so vojaška posredovanja verjetnejša, kot tudi manko pravičnosti, saj je težko imeti posamezne ljudi za odgovorne za vojne zločine, ki jih zagrešijo stroji. Razširjanje avtonomnih sistemov na druga področja – varovanje reda in miru, pravo, klanje in procesiranje mesa, pa tudi diagnostiko in šolstvo – lahko ima podobno za posledico sistemsko zatiranje, saj se lahko parcialni interesi uveljavljajo brez dejanske človeške odgovornosti.

Ključne besede:
avtonomno orožje, umetna inteligenca, moralna odgovornost, moralno vršilstvo, vrzel odgovornosti

Univerzitetna založba Univerze v Mariboru

DOI https://doi.org/10.18690/um.ff.11.2022.7
ISBN 978-961-286-675-4

AUTONOMOUS WEAPONS AND THE EROSION OF MORAL RESPONSIBILITY: SYSTEMATISATION OF VIOLENCE AND DISAPPEARANCE OF JUSTICE

TOMAŽ GRUŠOVNIK
University of Primorska, Faculty of Education, Koper, Slovenia
tomaz.grusovnik@pef.upr.si

Abstract The chapter argues that autonomous weapons represent a real danger for erosion of accountability and moral responsibility in modern warfare. Indeed, the psychology as well as history of war shows us that humans exhibit aversion to killing enemy soldiers, a so-called "low hit ratio". This obstacle to killing enemies was largely overcome with physical distance between soldiers and their targets, i.e., with the help of artillery, bombing and UAVs. Autonomous weapons, however, do not only extend this distance but move it to an entirely different level – the distance now becomes conceptual, as the decision to kill is transferred to those systems. This responsibility gap results in lowering the bar for using force, making military interventions likelier, as well as in lack of justice, since it is hard to hold individual humans accountable for potential war crimes committed by machines. The spread of autonomous systems to other areas – policing, law, slaughter and meat production, as well as diagnostics and schooling – can similarly result in systemic oppression because partial interests can be enforced without any real human responsibility.

Keywords: autonomous weapons, artificial intelligence, moral responsibility, moral agency, responsibility gap

https://doi.org/10.18690/um.ff.11.2022.7 DOI
978-961-286-675-4 ISBN

University of Maribor Press

1 Opredelitev avtonomnega orožja in ravni avtonomnosti

27. marca 2020 se je naposled zgodilo to, česar so se bali vojaški strategi, zagovorniki človekovih pravic, politiki in širša javnost: na bojišču v Libiji so bili v napadu uporabljeni brezpilotniki, ki so zmožni sami poiskati tarčo in jo napasti. Odbor za Libijo Varnostnega sveta Združenih narodov je tako nedavno v pismu poročal sledeče:

> Logistični konvoji in umikajoče se HAF[1] so posledično izsledile in daljinsko napadle brezpilotna bojna letala ali smrtonosni avtonomni oboroženi sistemi, kot so STM *Kargu 2* (glej prilogo 30) in druge samodejne naprave baražnega ognja.[2] Smrtonosni avtonomni oboroženi sistemi so bili programirani za napad na tarče brez potrebe po podatkovni povezavi z operaterjem in orožjem: dejansko prava 'streljaj, pozabi in najdi' zmožnost. (Panel of Experts on Libya established pursuant to resolution 1973 (2011) 2021: 17)

Čeprav gre za šokantne novice, vredne kakšnega znanstveno-fantastičnega filmskega scenarija, pa žal ne gre za posebno presenečenje. Uporaba avtonomnih orožij je namreč čedalje bolj zaželena, med razlogi, zaradi katerih vojske držav po svetu uporabljajo avtonomna orožja,[3] pa je njihova računska moč (stroji so sposobni upoštevati in preračunati veliko več podatkov kot človek, zaradi česar lahko situacijo prej 'osmislijo' in so tudi natančnejši ter varčnejši), sposobnost delovanja v človeku sovražnem okolju (ne le vojaškem, pač pa tudi v biološko-fizičnem) ter sposobnost delovanja v 'brezpovezavnem načinu', torej tam, kjer ni podatkovne povezave med sistemom in operaterjem oziroma upravljalcem, zaradi česar klasični brezpilotniki odpovedo. Zraven vseh prednosti, ki jih tovrstno orožje ponuja, se v literaturi omenja, da bo uporaba le-teh postala pravzaprav neizbežna: zaradi izjemno kompleksnih situacij bo namreč človeško procesiranje informacij na bojišču in posledično odločanje postalo preprosto prepočasno in neučinkovito: napad jate brezpilotnikov, flote duhov (avtonomnih plovil) in tropa robotskih psov bodo lahko odvrnili le sorodno kompleksni obrambni sistemi. Poleg tega se pogosto – in morda

[1] Sile, povezane s Haftarjem, libijskim vojaškim poveljnikom, op. T. G.
[2] Angl. *loitering munition*.
[3] Celotna besedna zveza, ki označuje ta koncept, je sicer 'smrtonosni avtonomni oboroženi sistem', angl. *Lethal Autonomous Weapons System* (LAWS). V nadaljevanju bom za ta koncept uporabljal izraz 'avtonomno orožje', ki se je prijel tudi v slovenščini.

paradoksno – velikokrat omenja humanitarna prednost uporabe takšnih orožij, k čemur se vrnemo v naslednjem odseku.

Toda kaj sploh je 'avtonomno orožje'? Pojem 'avtonomnega orožja' ima več različnih definicij, v glavnem pa lahko rečemo, da gre za sistem strojne in programske opreme, ki je na podlagi senzorjev zmožen prepoznati tarče in jih uničiti brez človekovega posredovanja. Članek, objavljen v zbirki povzetkov iz srečanja Mednarodne komiteja Rdečega križa na temo raznolikih vidikov avtonomnega orožja, sicer podaja sledečo definicijo, ki jo prevzemajo različni avtorji in institucije:

> Avtonomni oboroženi sistem je tisti, ki se lahko priuči svojega delovanja oziroma ga prilagodi kot odgovor na spreminjajoče se okoliščine v okolju, v katerem je uporabljen. Resnično avtonomni sistem ima umetno inteligenco, ki je sposobna implementacije mednarodnega humanitarnega prava. (Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects 2014: 64).

V konkretnem smislu gre lahko za premične sisteme (brezpilotniki, kvadro- in multikopterji, kopenski roboti in vozila ter vodne in podvodne aplikacije) oziroma za stacionarne priprave (avtonomna mitraljezna gnezda, t. i. *sentry guns*, kakršna najdemo na izraelsko-palestinski meji). Avtonomna orožja so lahko smrtonosna, torej namenjena ubijanju ljudi, lahko pa so tudi namenjena uničevanju sovražnikovih izstrelkov in avtonomnih sistemov. Poleg njihove potencialne smrtonosnosti je v moralnem smislu še najpomembnejše razlikovanje stopnje avtonomije takšnega orožja, ki jo določimo glede na vlogo človeškega operaterja pri delovanju take naprave. To razlikovanje, ki ga kot najbolj smiselno v navezavi na pojem 'avtonomije' izpostavlja tudi Kaja Smolnikar (2019) v svoji magistrski nalogi, lahko povzamemo po publikaciji Human Rights Watch na sledeč način:

> **Človek-v-zanki:** sistemi lahko izberejo tarče oziroma izbrane tarče uničijo zgolj na podlagi človekovega povelja;
>
> **človek-nad-zanko**: sistemi lahko izberejo tarče in jih uničijo samodejno, a so nenehno pod nadzorom človeškega operaterja, ki ima možnost preklica dejanja;

> **človek-izven-zanke:** sistemi najdejo in uničijo tarče, ne da bi to delovanje usmerjal ali spremljal človek. (Human Rights Watch 2012: 2)

Med 'popolnoma avtonomna orožja' lahko tako prištevamo v celoti le zadnjo kategorijo, drugo pa pogojno, in sicer v tistih primerih, kjer je človeški nadzor zgolj še načelen, saj je dogajanje na bojišču preveč kompleksno in hitro za dejansko človekovo smiselno odločanje. V slednjem primeru možnost preklica sicer zmeraj obstaja, a je v praksi skoraj že enako odpoklicu sistemov tretje kategorije. Druga kategorija je torej precej nejasna cona, mejni primer, kjer je stopnja avtonomije sistema lahko odvisna od zelo partikularnih in kontingentnih okoliščin, tudi konec koncev od podatkovne povezave med nadzornikom in sistemom, saj si je moč zamisliti situacijo, v kateri odsotnost slednje de facto izniči možnost preklica. Zaradi tega bomo slednjo prištevali med avtonomna orožja, saj je konec koncev znotraj te kategorije stroj zmožen ubiti človeka brez operaterjevega posredovanja. V nadaljevanju bomo torej z 'avtonomnim orožjem' mislili tiste sisteme, ki delujejo na podlagi načel 'človek-nad-zanko' in 'človek-izven-zanke'.

2 Etiški izzivi, povezani z avtonomnimi orožji

Že zgoraj smo nakazali, da nekateri avtorji med prednosti avtonomnih orožij uvrščajo humanitarne razloge. Po njihovem mnenju naj bi uporaba avtonomnih orožij tako pomagala spoštovati mednarodna humanitarna načela, saj lahko zmanjša število tako vojaških kot civilnih žrtev, nepotrebno trpljenje, pa tudi vojne zločine, ker ne spodbuja krvoločnosti, je natančnejša, kar praviloma rezultira v manj civilnih žrtvah, hkrati pa tudi zaščiti vojake (glej Zawieska 2017: 51). Ronald Arkin tako denimo zagovarja uporabo avtonomnega orožja iz moralnega vidika, saj bi se z njegovo pomočjo lahko izognili številnim dejavnikom vojnih zločinov:

> Možne razlage vojnih zločinov vsebujejo: visoke zavezniške izgube, ki spodbujajo maščevanje; pogoste zamenjave v poveljniški hierarhiji, ki vodijo k šibkemu vodenju; razčlovečenje nasprotnika s pomočjo poniževalnih vzdevkov; slabo izurjene in neizkušene enote; nejasno določenega sovražnika; nejasne ukaze, kjer se lahko namera ukaza napačno interpretira kot nezakonita; mladost in neizkušenost enot; zunanji pritisk, denimo zahteva po številčnosti sovražnikovih žrtev; ugodje, ki ga prinaša moč ubijanja ali občutje frustracij. Obstaja očiten prostor za izboljšave in

> avtonomni sistemi nam bi lahko pomagali nasloviti nekatere od teh problemov. (Arkin 2013: 2)

Marco Sassoli je enako optimističen glede izboljšanja humanitarne situacije na bojiščih z vpeljavo avtonomnih orožij. Prepričan je, da bodo roboti, ki jih ne bo strah za lastno preživetje in ki ne bodo pobijali 'preventivno' in ki tudi sicer ne bodo podvrženi čustvom, denimo jezi ali kakšni drugi strasti, delovali bistveno natančneje in bolj upravičeno:

> Morda zaradi tega, ker sem v tolikih oboroženih konfliktih bil soočen s številnimi kršitvami, ki so jih povzročila človeška bitja, a nikdar s hudodelstvi, ki bi jih zagrešili roboti (čeprav je treba priznati, da v oboroženih konfliktih, ki sem jim bil priča, niso obstajali), moje prvo občutje ni skepticizem, pač pa upanje za boljše spoštovanje mednarodnega humanitarnega prava. Samo človeška bitja so lahko nehumana in samo človeška bitja lahko načrtno izberejo, da se ne bodo držala pravil, ki so bila vzpostavljena zato, da bi jim sledili. Zdi se mi bolj verjetno pričakovati (in zagotoviti), da bo mednarodnemu humanitarnemu pravu sledila oseba, ki bo sestavila in pripravila avtonomno orožje na mirnem delovnem mestu kot pa vojak na bojišču ali v sovražnem okolju. Roboti ne morejo sovražiti, ne občutijo strahu, ne morejo biti lačni ali žejni in utrujeni ter nimajo nagona po preživetju. 'Roboti ne posiljujejo'. Hkrati lahko zaznajo več informacij in jih sprocesirajo hitreje kot človeško bitje. Ko bodo orožja, ki sprožajo kinetično energijo, postala čedalje hitrejša in bolj kompleksna, se lahko zgodi, da bodo ljudje preprosto preplavljeni z informacijami in odločitvami, ki jih bodo morali sprejeti, da bi jih usmerili. Ljudje pogosto druge ubijajo zato, da bi se izognili lastni smrti. Robot pa lahko preloži uporabo sile do zadnjega, najprimernejšega trenutka, ko je bilo ugotovljeno, da sta tarča in napad legitimna. (Sassoli 2014: 310)

Seveda je jasno, da vsi niso tako navdušeni nad uporabo avtonomnih orožij in da obstaja zelo veliko pomislekov glede uporabe takšnih sistemov. Že omenjeni Ronald Arkin tako našteva probleme, povezane z odgovornostjo za vojne zločine (kar je tudi ključna točka tega poglavja, h kateri se še vračamo v nadaljevanju), tako tudi z zniževanjem praga za uporabo sile, pa z nepripravljenostjo podeliti robotom pravico veta in zavrnitve izvršitve ukaza, neavtoriziranim razširjanjem tehnologije,

posledicami za kohezijo v vojski, elektronsko varnostjo, iznakaženjem misij.[4] Human Rights Watch se za razliko od Arkina poskuša strožje držati moralne in humanitarne problematike ter jo posledično ne pomeša z moralnimi izzivi, našteje pa sledeče zagate, ki bi jih avtonomno orožje lahko imelo v navezavi na skladnost z mednarodnim humanitarnim pravom in sledečimi načeli:

> **Razlikovanje:** Avtonomno orožje bi lahko naletelo na težave pri razlikovanju med bojevniki in civilisti. V zadnjem času se je namreč korenito spremenila narava vojskovanja, saj so tradicionalne oborožene konflikte med državami nadomestili spopadi urbanih skupin, ki jih je zelo težko ločiti od civilistov;
>
> **Sorazmernost:** Bojazen je, da avtonomni oboroženi sistemi ne bi znali preceniti, kdaj uporaba vojaške sile povzroči več škode za civiliste, kot pa prinese koristi na bojišču. Sorazmernost je močno odvisna od konteksta in številnih dejavnikov, ki jih je nemogoče kvantificirati na način, da bi bila situacija primerna za pretres z algoritmom;
>
> **Vojaška nujnost:** Algoritem ne bi bil zmožen presoditi, kdaj je kak sovražni vojak postal hors de combat, torej nezmožen bojevanja, kar pomeni, da mu po vojnem pravu pripada poseben status. Lahko bi se torej zgodilo, da bi avtonomno orožje likvidiralo de facto onesposobljenega sovražnega vojaka, saj denimo ne bi prepoznalo, da je ranjen do te mere, da ni več sposoben za boj in bi mu pravzaprav moral pripasti status vojnega ujetnika;
>
> **Martensova klavzula:** To obče načelo pravi sledeče: »v primeru, ko neka situacija ni urejena s pravili veljavnega pogodbenega prava, civilno prebivalstvo in pripadnike oboroženih sil ščitijo načela mednarodnega prava, ki izhajajo iz običajev, sprejetih med civiliziranimi narodi, zakoni človečnosti in zahtevami javne vesti« (Sancin, Švarc in Ambrož 2009: 27)[5] Algoritmi nikdar ne bodo zmožni odločanja v skladu s 'civiliziranimi

[4] Angl. *mision creep*; gre za razširjanje intervencij preko prvotno predvidenega obsega, glej Arkin 2014: 35).
[5] Dokument vsebuje tudi tehnične razlage drugih tukaj uporabljenih načel.

običaji', prav tako nimajo nikakršne 'javne vesti', zato niso primerni vršilci[6] na bojišču.

Poleg težav, ki bi jih avtonomna orožja lahko imela z upoštevanjem zgornjih načel, Human Rights Watch našteje še sledeče moralne zagate, ki bi lahko nastopile z uporabo avtonomnih orožij.

Pomanjkanje človeških občutij: Zagovorniki moralnosti uporabe avtonomnih orožij pogosto navajajo argument, da bi zaradi odsotnosti strahu, interesa po preživetju, jeze ter maščevanja robotski sistemi lažje sledili načelom mednarodnega humanitarnega prava kot pa človeški borci. Toda hkrati pozabljajo, da so čustva tudi eden glavnih zaviralcev uporabe sile: ljudje imamo namreč spontano averzijo do ubijanja in nasilja, kar je naravna varovalka pred prekomerno uporabo sile, ki pri avtonomnem orožju umanjka. K 'psihološkemu otopevanju' in 'emocionalnemu umikanju' kot predpogoju ubijanja se še vrnemo v nadaljevanju.

Zniževanje praga za uporabo sile: Ker lahko avtonomna orožja vojske uporabljajo učinkovito, tudi prikrito in brez tveganj za življenja vojakov, se lahko zgodi, da bodo silo uporabile prej, kot bi jo sicer, če bi bilo treba napovedati konvencionalno vojno.

Zmanjševanje odgovornosti: Če avtonomno orožje po nesreči ubije civilista, ni jasno, kdo je odgovoren – je to morda poveljnik, ki je izdal ukaz, ali pa programer, izdelovalec ali celo robot sam? Na to ni enostavnega odgovora, kajti poveljnikov ne morejo odgovarjati za dejanja, ki jih niso želeli in na katera niso imeli vpliva. Tako nastane 'vrzel odgovornosti', kar pa je velik problem, saj ima odgovornost na bojišču vsaj dve funkciji: odvrača od povzročanja bodoče škode za civiliste, žrtvam ali njihovim svojcem pa nudi možnost pravičnega povračila.[7]

[6] Besedo 'vršilec' uporabljam za prevod angl. *agent*. Ta izbira, ki sem jo prvič zasledil pri kolegici Branislavi Vičar v sklopu njenih preiskovanj etike živali, se mi zdi bistveno bolj posrečena kot pojem, ki sem ga uporabljal do zdaj, namreč 'akter'. Za *agency* posledično uporabljam pojem 'vršilstvo'.
[7] Za vse glej Human Rights Watch 2012: 30–42.

Natanko te zagate so hkrati tiste, ki so najtesneje povezane z vprašanjem krivde ob napačni oziroma moralno sporni uporabi avtonomnega orožja, tako da se jim v nadaljevanju natančneje posvečamo.

3 Erozija odgovornosti in izginjanje pravičnosti s sodobnega bojišča

Naposled prihajamo do središčnih tez pričujočega poglavja: pokazati bomo poskušali, da enega največjih izzivov uvedbe avtonomnega orožja predstavlja omenjena vrzel odgovornosti[8] oziroma to, kar v naslovu poimenujemo erozija moralne odgovornosti. Posledica tega je na eni strani večja pripravljenost uporabiti orožje oziroma zniževanje praga uporabe sile, na drugi pa sistematizacija nasilja, s čimer imamo v mislih proces, podoben zgoraj omenjenemu 'iznakaženju misij', kjer bi lahko prišlo do tega, da bi bili avtonomni oboroženi sistemi nenehno vseprisotni z nalogo 'dušenja uporov'. Nadaljnja posledica takega procesa bi lahko bila izguba pravičnosti, in sicer retributivne pravičnosti, saj z izginjanjem moralnega vršilstva tudi izginja odgovornost, tako da že načelno umanjka točka prevzema odgovornosti za morebitne civilne žrtve.

Da bi razumeli, kako se avtonomna orožja skladajo z doktrino modernega vojskovanja, si moramo ogledati tako psihologijo kot z njo povezano kratko zgodovino bojevanja. S psihološkega stališča je gotovo najzanimivejše dejstvo, da ima velika večina ljudi averzijo do ubijanja. Nasilje je nekaj, čemur se najraje izognemo, to pa toliko bolj, če bi mi morali biti tisti, ki naj bi ga povzročali. Da bi torej zmogli postati nasilni – kar vojaki vsekakor morajo biti –, se je treba naučiti postaviti sočutje in empatijo do drugega človeškega bitja v oklepaj, kar je proces, ki ga v primeru mesarjev klavcev Melanie Joy imenuje »psihološko otopevanje« (Joy 2010: 144), Erich Fromm pa v primeru medčloveškega nasilja »emocionalno umikanje« (Fromm 2013: 164). Tudi če smo tega začasno sposobni, je velika verjetnost, da bo izvrševanje nasilja na nas pustilo globoke posledice. Te psihološke posledice so dobile celo ime: izvršiteljevi travmi, ki nastane zaradi nasilja, pravimo »perpetration-induced traumatic stress« (MacNair 2002).

[8] Angl. *accountability gap*.

Zaradi te človekove averzije do nasilja in ubijanja postane razumljivo, da so vojaki na frontah večinoma načrtno streljali v zrak, da bi zgrešili nasprotnika, ki so ga želeli le prestrašiti, nikakor pa tudi ubiti. Kot v svoji monografiji o psihologiji ubijanja pripoveduje vojaški psiholog Dave Grossman, je ta pojav tako pogost, da je izčrpno dokumentiran. V resnici se je človeštvo v velikem delu svoje zgodovine bojevalo predvsem na podlagi take taktike ustrahovanja s 'poziranjem',[9] ki je izključevalo ubijanje nasprotnika in s pomočjo katerega Grossman nadgradi tradicionalni 'boj ali beg' model, ki ima poslej štiri kategorije: boj, beg, poza, podreditev. Aleksander Veliki naj bi na svojem pohodu tako izgubil vsega 700 mož, pri čemer je večina njegovih sovražnikov bila ubitih po bitki, medtem ko je poročnik George Roupell moral v prvi svetovni vojni s sabljo tolči svoje može po hrbtu in jim v strelskem jarku groziti, da merijo nižje.[10] Vse od plemenskih vojn do evropskih napoleonskih vojn je veljalo, da si moral sovražnika prestrašiti. Napredek tehnike vojskovanja pa se kaže tudi v poznavanju načinov, s katerimi je mogoče ljudi prepričati, naj sovražnike ubijajo (in da v svoji taktiki nazadnje celo računajo s tem, da bo sovražnik iz upora do ubijanja najprej zastraševal!):

> Res, na zgodovino vojskovanja lahko gledamo kot na zgodovino čedalje učinkovitejših mehanizmov za omogočanje in pogojevanje ljudi, da premagajo svoj vrojen upor do ubijanja drugih sorodnih človeških bitij. V številnih okoliščinah so se visoko usposobljeni sodobni vojaki borili s slabo usposobljenimi gverilskimi silami in težnja slabo pripravljenih sil k temu, da se nagonsko zatečejo k zastraševalnim mehanizmom (recimo streljanje v zrak), je pomenila pomembno prednost za bolj izurjene sile. (Grossman 2009: 13)

Eden teh načinov, kako prepričati ljudi, da bodo pobijali soljudi, je umik vojaka iz bojišča in uvedba tehnologije bojevanja na daljavo. Med prvimi takimi orožji je bila artilerija dolgega dometa in pa bombardiranje, pozneje pa termovizija (primer Grossman 2009: 11, 169–170) in seveda brezpilotni letalniki. Grossman povzema: »Večino dejavnikov, ki omogočajo ubijanje na bojišču, je moč ugledati v difuziji odgovornosti« (Grossman 2009: 193).

[9] Angl. *posturing*.
[10] Za številne primere od antike do vietnamske vojne glej Grossman 2009: 6–17.

V tej točki postane očitno, da so avtonomna orožja kot nalašč ustvarjena za krpanje tiste 'šibkosti', ki jo ima človeški vojak, namreč averzije do ubijanja in travme, ki izvira iz nje, če do jemanja življenj pride. Smemo reči celo tole: *Če so artilerija, bombardiranje in brezpilotniki vnesli fizično distanco do nasprotnika, ki je omogočila obsežnejše in lažje ubijanje z manj slabe vesti, potem avtonomno orožje vnaša konceptualno distanco do tarče, ki vprašanje vesti odpravlja v celoti.*

Konceptualna distanca, ki nastane med poveljniki in vojaki, avtonomnim orožjem ter vojaškimi ali celo civilnimi žrtvami, erodira odgovornost, saj ni več znano, kdo naj bi za dejanja odgovarjal. So to proizvajalci, programerji ali poveljniki, nadzorniki oziroma politiki? Najtežje je odgovornost obesiti programerjem, kajti algoritmi, ki so vprogramirani v takšne sisteme, so običajno plod dolgega razvoja celih skupin strokovnjakov. Prvi algoritmi za samodejno zaznavanje so bili sprogramirani kot pomoč slabovidnim, medtem ko lahko danes predstavljajo pomemben element 'osmišljanja' informacij iz senzorjev avtonomnih sistemov. Proizvajalca lahko imamo že nekoliko lažje za odgovornega, saj je tisti, ki v prvi vrsti sploh materialno omogoči tovrsten instrument. Še najbolj intuitivno je, da imamo v primeru človeka-izven-zanke za odgovorne politike, ki so avtorizirali uporabo sile in pa poveljnike, čeprav tudi tukaj prihaja do zapletov: moč si je namreč zamisliti, da so avtorizirali uporabo avtonomnega orožja, ki pa je nato samo bodisi zatajilo bodisi izvršilo nekaj, česar ljudje niso predvideli. Za konkretne avtonomne 'odločitve' sistema (uničenje natanko *te* tarče, povzročitev natanko *te* škode) torej ne moremo kriviti ljudi, saj jih tudi ne moremo kriviti za dejanja, ki jih zagrešijo njihovi podrejeni in za katera poveljniki niso vedeli. Politike in poveljnike lahko imamo tako odgovorne le (še) za odobritev občih ciljev misij.[11]

V primeru človeka-nad-zanko se nam najbrž intuitivno zdi, da lahko imamo vsekakor za odgovornega nadzornika, ki bi lahko preprečil spodrsljaj avtonomnega orožja oziroma ki bi moral prevzeti odgovornost za vse smrti, ki jih avtonomno orožje povzroči, saj je potiho omogočal delovanje sistema s tem, ko ga ni preklical. Toda tudi tukaj nastane težava, saj je, kot opozarjajo mnogi strategi in tehniki, sodobno bojišče tako zapleteno, da ga ljudje le stežka osmislimo, zaradi česar se ravno zatekamo k uporabi umetne inteligence in avtonomnih sistemov. Kadar je torej situacija na bojišču izjemno zapletena in avtonomno orožje deluje v skladu s

[11] Za podobno stališče prim.: Human Rights Watch 2012: 42.

svojimi programiranimi načeli, postaja funkcija človeškega nadzornika čedalje bolj nerelevantna, s tem pa se manjša tudi njegova odgovornost. Najbrž je celo tako, da bi nadzornik ravno bil odgovoren v primeru, kadar bi preventivno prekinil delovanje avtonomnega sistema, a bi se naknadno izkazalo, da je bila to preuranjena odločitev, ki je ogrozila ali celo povzročila izgubo prijateljskih sil. V tem primeru bi mu lahko očitali, da je preprečil uporabo dobrega, delujočega in preverjenega sistema, ki bi lahko obvaroval življenja. Tako si je lahko zamisliti, da bodo operaterji in nadzorniki raje sledili predlogom algoritmov oziroma prepustili dogajanje avtonomnim sistemom, kot pa tvegali, da bi se s svojo omejeno zmožnostjo presoje zoperstavili umetni inteligenci.

Če pa ne moremo imeti več nikogar za zares odgovornega za spodrsljaje ali nepredvidena 'dejanja' avtonomnih sistemov, potem lahko postane uporaba sile pogostejša. Obstaja namreč utemeljena bojazen, da se bodo države z uporabo avtonomnih orožij lažje odločale za vojaška posredovanja, saj bo umanjkala ključna zavora, ki posameznike na poziciji moči zaustavi pred napovedjo vojne: nevarnost izgube in velikih 'stroškov', tako materialnih kot moralnih: »Voditelji bi se morda manj obotavljali vojne napovedi, če bi se tveganje za njihove vojake zmanjšalo ali izničilo« (Human Rights Watch 2012: 39–40). Kar se tiče posameznikov na bojišču pa smo tudi videli, da je za vojake ključnega pomena premagovanje averzije za ubijanje, ki je toliko večja, koliko manjša je razdalja med vojakom in nasprotnikom: ker avtonomno orožje – kot smo večkrat poudarili – vnaša konceptualno distanco med vojaka in nasprotnika, si je moč zamisliti, da bo imel vojak veliko manj pomislekov sprožiti avtonomni sistem kot pa puško.

Avtonomna orožja bodo torej skoraj zagotovo znižala prag uporabe sile, toda morda bo to zgolj prva faza uvedbe tovrstnih instrumentov na sodobno bojišče in bojišče prihodnosti. V drugi fazi bi uporaba avtonomnih orožij lahko postala čedalje bolj preventivna. Brezpilotniki, ki bi nenehno patruljirali v zraku, flote duhov, ki bi nenehno plule po oceanih in morjih, pa tudi policijski kopenski roboti bi lahko bili uporabljani čedalje bolj 'preventivno', na način, da bi 'nevtralizirali' točke upora režimu, *še preden* bi se oboroženi spopad sploh razplamtel. Grožnja s silo bi na tak način lahko postala sistemska, ljudje pa bi živeli pod 'avtonomnim ščitom', ki bi jih nenehno 'varoval' pred 'nevarnostmi'. Jasno je, da bi takšen skorajda že Orwellovski svet bil največje možno tveganje za demokracijo, saj bi resne alternative vladajočim režimom tako rekoč strojno izničil, še preden bi se pojavile. Če bi ob tem uporabili

še različna 'diagnostična' orodja, torej umetno inteligenco, ki s pomočjo naprednih algoritmov nenehno 'rangira' oziroma razvršča in kategorizira posameznike ter označuje tiste, ki so 'potencialno nevarni' sistemu, potem je jasno, da bi svet, v katerem bi pristali, bil še hujši od tistega, ki si ga je zamislil Orwell.

Dodajmo še, da se z erozijo moralne odgovornosti s sodobnega bojišča iz njega umika tudi pravičnost. Če namreč ni nihče neposredno kriv za spodrsljaje in napake oziroma če pravni sistem in javna vest celo namensko delujeta na tak način, da pristajata na skrivanje odgovornosti za avtonomnimi sistemi, potem svojci žrtev ne morejo več najti zadoščenja, saj za smrti njihovih bližnjih naenkrat 'ni nihče kriv'. Svojci žrtev in drugi oškodovanci tako izvisijo, kar nakazuje na krizo retributivne oziroma povračilne pravičnosti, saj ni nikogar, ki bi se moral opravičiti, odgovarjati za nasilje oziroma plačati odškodnino za povzročeno škodo. Pravzaprav lahko rečemo, da se pravičnost z bojišča umika skupaj s človekom kot moralnim vršilcem.

4 Avtonomni sistemi in sistemsko nasilje

Znanstvenofantastični filmi zahodne produkcije so v dvajsetem stoletju največjo težavo umetne inteligence prepoznavali v njeni zoperstavitvi človeku in družbi: roboti niso postali le zavestna bitja, pač pa so s svojo strojno inteligenco presegli človeka in si ga podjarmili, morda na podoben način, kot si je človek podjarmil živali in naravo. Toda čas, z njim pa tehnološki in družbeni razvoj, je pokazal, da je nevarnost vdora umetne inteligence v človekov življenjski svet povsem drugačna: človekovo samoiniciativno podrejanje zahtevam informacijske družbe. Če pomislimo na primer pametnih telefonov, hitro uvidimo, da nam jih ni moral nihče vsiliti, nasprotno, še preveč radi se podrejamo uporabi aplikacij, ki nas – če si lahko dovolimo ta izraz – učijo razmišljanja, analognega funkcioniranju algoritmov. Umetna inteligenca nas torej ni osvojila, ampak prej zasvojila in nas prisilila spremeniti naše mišljenjske, čustvene in druge odzive ter vedênje. Ne zavzemajo nas roboti, ampak se robotiziramo sami. Karolina Zawieska tako pomenljivo pravi, da »zdaj je mogoče razpravljati o tem, ali se lahko stroji držijo človeških etičnih načel, ne zaradi tega, ker bi avtonomni sistemi postali podobni ljudem, pač pa zato, ker človeška bitja nase čedalje bolj gledajo kot na stroje« (Zawieska 2017: 53). V zaključku poglavja tako poskušamo pokazati, da uvajanje avtonomnih sistemov ni težavno zgolj za sodobno bojišče, kjer morda sicer res pride najočitneje do izraza

njegova moralna problematika, pač pa nasploh pomeni korenit poseg v družbo, in sicer predvsem v smeri sistematizacije nasilja in represije.

Zraven vojske in bojevanja bodo roboti prav gotovo v naslednjih letih prevzeli klavniška opravila. Tudi tukaj[12] imamo namreč opraviti s travmatskim stresom kot posledico opravljanja takšnega dela in robotizacija klavnega procesa bi tega ne le pospešila, ampak naredila tudi bolj 'humanega' in 'higienskega'. S tem bi se nasilje nad živalmi že skoraj v celoti robotiziralo in mehaniziralo, s čimer bi zelo očitno vkorakali v popolno živinorejsko distopijo, kjer bi vsi samo še prodajali in konzumirali mesne izdelke, medtem ko odgovornosti za smrti živali ne bi rabil več prevzemati nihče.

Težave, povezane z avtonomnimi sistemi in strojnim odločanjem, lahko zasledimo tudi na številnih drugih področjih, kjer njihova uporaba iz vidika izginjanja moralnega vršilstva ni nič manj sporna, pa čeprav je – vsaj na prvi pogled – manj pretresljiva. Specifično v navezavi na erozijo odgovornosti in posledično izginjanje pravičnosti je vpeljava avtonomnih sistemov problematična znotraj policijskega delovanja. Že dolgo je tako, da se policisti in drugi uradniki radi sklicujejo na to, da »oni samo opravljajo delo« ali »izvršujejo ukaze«, kar je natanko mehanizem distanciranja od lastne odgovornosti in poskus bega pred dejstvom, da so sami neizbežno moralni vršilci in prav oni kot posamezniki tisti, ki napišejo kazen, zavrnejo vstop v državo ali izvršijo kakšno drugo dejanje, ki ima neprijetne, včasih celo brezčutne in krivične posledice za posameznice in posameznike. Skupaj s Henryjem D. Thoreaujem lahko takšnemu uradniku, ki se sklicuje na svojo nemoč, medtem ko dejansko izvršuje neko dejanje, zmeraj odvrnemo: »Pusti službo« (Thoreau 2016: 24)! S tem izpostavimo njegovo hoteno nevednost glede dejstva, da je zmeraj on sam kot posameznik vršilec dejanja, da ima zmeraj izbiro, pa čeprav se tega ne zaveda ali noče zavedati.[13] Če smemo povzeti, lahko zato rečemo, da avtonomni sistemi in uporaba rangirnih algoritmov naslavljajo natanko to človekovo potrebno po distanciranju od odgovornosti: pričakovati je, da bo njihova vpeljava pripomogla k še bolj sistematiziranemu razpuščanju individualne odgovornosti in da se bodo posamezniki skorajda z olajšanjem odrekli moralnemu vršilstvu ter se pri moralnem odločanju raje sklicevali na protokole, algoritme in komisije, saj se, kot

[12] Za analizo travmatskih posledic za mesarje klavce glej Grušovnik 2016: 142–47.
[13] Za analizo hotene nevednosti, tudi glede prevzemanja lastne odgovornosti, glej mojo razpravo v Grušovnik 2020: poglavje o Sartru razkriva, da je prepričanje, da 'ne moreš ničesar spremeniti', primer 'slabe vere'.

pravi Friderik Klampfer: »/…/ naporno in zagatno moralno presojo raje nadomesti z deklamiranjem določil poklicnih etičnih kodeksov /…/« (Klampfer 2010: 17).[14] Pravniki, zdravniki in učitelji se bodo za vzrok specifičnih sodb, diagnoz in ocen lahko preprosto sklicevali na algoritme in se na ta način otresli svoje odgovornosti.

Že zgoraj smo opisali primer nadzornika avtonomnega sistema, delujočega po načelu človek-nad-zanko, ki se v nekem trenutku odloči preklicati operacijo, zaradi česar nastanejo prijateljske izgube. Takemu posamezniku bi lahko očitali, da je imel na voljo najboljšo oceno situacije na bojišču, ki so jo pripravili najnaprednejši algoritmi, a se je raje odločil, da bo izvršitev dejanja preklical, zaradi česar mu lahko skušamo naprtiti krivdo za prijateljske izgube. Ker obstaja takšna nevarnost, je verjetnost, da bo nadzornik raje slepo sledil predlogom umetne inteligence, kot pa se s svojo omejeno zmožnostjo presoje zoperstavil algoritmom, velika. To pomeni, da bodo ljudje raje, kot bi tvegali samostojno odločanje in s tem prevzemanje odgovornosti, slepo sledili algoritmom. To pa seveda ne velja zgolj za vojaške situacije, pač pa tudi za šolstvo, sodstvo in zdravstvo. Lahko si zamislimo zdravnika, ki mu algoritem navrže eno diagnozo, do katere pa je na podlagi lastnih izkušenj skeptičen in je ne sprejme. Nato pa se izkaže, da je algoritem vendarle imel prav. Je v takem primeru zdravnik kriv, ker je 'zanemaril' rezultate najboljšega diagnostičnega orodja, ki mu je bilo na voljo? Tudi če odgovora na to vprašanje nimamo, vidimo, da se zdravniku bolj 'splača' slepo slediti diagnostičnemu orodju, kot pa tvegati prevzem odgovornosti za lastno moralno vršilstvo in s tem tvegati krivdo ter celo morebitne odškodninske tožbe. Povsem enako seveda velja za sodnika, ki bi se uprl nakazani sodbi, ali učitelja, ki bi preglasil strojno predlagano oceno.

Tovrstni procesi – erozija moralne odgovornosti zaradi ponikanja individualnega in osebnega moralnega vršilstva v avtonomnih sistemih – bi lahko predstavljali nekaj takšnega kot dodatno sistematizacijo nasilja, kjer bi algoritmi, o katerih bi lahko odločala le peščica močnih posameznikov, diktirali smer družbenega razvoja. Lahko bi prišlo do sistemskega zatiranja in uveljavljanja parcialnih interesov pod pretvezo 'digitalizacije', za kar pa – ker bi vse izvrševali 'optimizirani' avtonomni sistemi – nihče ne bi bil odgovoren.

[14] Za podobno poanto glej Grušovnik 2021: 9.

Viri in literatura

Arkin, R. (2013.) »Lethal Autonomous Systems and the Plight of the Non-Combatant«. *Artificial Intelligence and Simulation of Behaviour Quarterly*, 137.

Arkin, R. (2014). »Ethical Restraints of Lethal Autonomous Robotic Systems: Requirements, Research, And Implications«. V *Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects*«. Ženeva: International Committee of Red Cross, brošura s povzetki prispevkov, str. 33–37.

»Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects«. (2014). Ženeva: International Committee of the Red Cross, brošura s povzetki prispevkov.

Fromm, E. (2013). *Anatomija človekove uničevalnosti*. Ljubljana: Mladinska knjiga.

Grossman, D. (2009). *On Killing: The Psychological Cost of Learning to Kill in War and Society*. New York: Little, Brown, and Company.

Grušovnik, T. (2016). *Etika živali*. Koper: Založba Annales.

Grušovnik, T. (2020). *Hotena nevednost*. Ljubljana: Slovenska matica.

Grušovnik, T. (2021). »Uvodnik: vloga filozofije pri razumevanju etike danes«. V Grušovnik, T. in Pirc, G. (urd.), *Etika Danes, Filozofsko raziskovanje in razumevanje etike v sodobni družbi*. Koper: Založba Annales.

Human Rights Watch. (2012). *Losing Humanity – The Case against Killer Robots*. Cambridge, Massachusetts: International Human Rights Clinic. URL = https://www.hrw.org/sites/default/files/reports/arms1112_ForUpload.pdf.

Joy, M. (2009). *Why We Love Dogy, Eat Pigs, and Wear Cows? – An Introduction to Carnism*. San Francisco: Conari Press.

Klampfer, F. (2010). *Cena življenja – razprave iz bioetike*. Ljubljana: Krtina.

MacNair, R. (2002). *Perpetration-Induced Traumatic Stress: The Psychological Consequences of Killing*. Praeger Publishers: Westport.

Sancin, V., D. Švarc in M. Ambrož. (2009). *Mednarodno pravo oboroženih spopadov – strokovno delo za potrebe Slovenske vojske*. Ministrstvo za obrambo Republike Slovenije, Knjižnično-informacijski in založniški cente.

Sassoli, M. (2014). »Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified«. *International Law Studies*, 90, str. 308–340.

Smolnikar, K. (2019). *Zakonitosti uporabe avtonomnega orožja v mednarodnem pravu* (magistrsko delo). Univerza v Ljubljani: Pravna fakulteta.

Thoreau, H. D. (2016. *Državljanska nepokorščina*. Ljubljana: LUD Šerpa.

UN. Panel of Experts Established pursuant to Security Council Resolution 1973 (2011). (2021). »Letter dated 8 March 2021 from the Panel of Experts on Libya established pursuant to resolution 1973 (2011) addressed to the President of the Security Council«. United Nations Security Council.

Zawieska, K. (2017). »An Ethical Perspective on Autonomous Weapon Systems: Perspectives on Lethal Autonomous weapon Systems«. *UNODA (United Nations Office for Disarmament Affairs) occasional papers*, 30.

2. DEL

UMETNA INTELIGENCA, ZNANOST IN IZOBRAŽEVANJE

VIRTUALNI POGOVORNI AGENT EVA – UMETNA INTELIGENCA ZA BOLJ NARAVNO INTERAKCIJO Z NAPRAVAMI

IZIDOR MLAKAR, SIMONA MAJHENIČ, MATEJ ROJC
Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Maribor, Slovenija
izidor.mlakar@um.si, simona.majhenic@um.si, matej.rojc@um.si

Sinopsis Eden ključnih izzivov človeku podobnih vmesnikov je smiselna raba subtilnih nejezikovnih signalov v interakciji. Zato je to po eni strani povezano s sinhronizacijo, po drugi pa z zagotavljanjem 'pravilne' interpretacije. Nejezikovni elementi niso zgolj preprosti nadomestki jezikovne vsebine, ampak so dejanske (neintegrirane) sestavine govora. Zato je cilj prispevka razviti pogovorni model za ustvarjanje človeku podobnega pogovora in najti rešitev za čustveno ter personalizirano interakcijo med človekom in strojem. Model ponuja (i) platformo za ustvarjanje 'pogovornega' znanja in virov, (ii) okvir za načrtovanje in ustvarjanje objezikovnega obnašanja in (iii) okvir za izvedbo čustvenega in odzivnega objezikovnega obnašanja s pomočjo izražanja stališč, čustev ter gest, sinhroniziranih z govorom. V drugem poglavju osvetlimo trenutno stanje na področju utelešenih pogovornih agentov. V tretjem poglavju orišemo pogovorni model EVA, kjer je osrednja zamisel oblikovati različne oblike objezikovnega obnašanja (geste) v povezavi z neoznačenim besedilom in širšim družbenim ter pogovornim kontekstom. V četrtem poglavju opisujemo pridobivanje 'pogovornega' znanja in potrebnih virov, ustvarjenih s pomočjo označevanja spontanega dialoga in korpusno analizo. Skupaj s petim poglavjem nato opisujemo, kako te vire vključimo v dvostopenjski pristop samodejnega ustvarjanja objezikovnega obnašanja. Prispevek sklenemo s študijo primera in prikazom sinteze objezikovnega obnašanja utelešenega pogovornega agenta EVA (Rojc et al. 2017).

Ključne besede:
pogovorni agent s telesom,
nejezikovno obnašanje,
pogovorni model EVA,
sinteza pogovornega obnašanja,
personalizirana interakcija

Univerzitetna založba Univerze v Mariboru

DOI https://doi.org/10.18690/um.ff.11.2022.8
ISBN 978-961-286-675-4

THE EMBODIED CONVERSATIONAL AGENT EVA – ARTIFICIAL INTELLIGENCE FOR A MORE NATURAL INTERACTION WITH DEVICES

IZIDOR MLAKAR, SIMONA MAJHENIČ, MATEJ ROJC
University of Maribor, Faculty of Electrical Engineering and Computer Science, Maribor, Slovenia
izidor.mlakar@um.si, simona.majhenic@um.si, matej.rojc@um.si

Abstract One of the key challenges of humanoid interfaces is sensible usage of subtle non-linguistic signals in interactions. This is, on the one hand, connected with synchronisation, and on the other with ensuring the 'correct' interpretation. Non-linguistic elements are not merely simple substitutes for linguistic content, but actual (non-integrated) components of speech. The paper aims at developing a conversational model for creating humanlike conversations and at finding a solution for emotional and personalised humans-machine interactions. The model offers (i) a platform for creating 'conversational' knowledge and sources, (ii) a framework for designing and creating colinguistic behaviour, and (iii) a framework for the realization of emotional and responsive colinguistic behaviour. In the second chapter, we look at the state-of-the-art in the field of embodied conversational agents. In the third chapter, we outline the conversational model EVA. In the fourth chapter, we describe the acquisition of 'conversational' knowledge and the relevant sources. In the fifth chapter, we then describe how to incorporate these sources into a two-stage approach for the automatic creation of colinguistic behaviour. The paper is concluded with a case study and demonstration of the colinguistic behaviour synthesis with the embodied conversational agent EVA (Rojc et al. 2017).

Keywords: embodied conversational agen, non-verbal behaviour, conversational model EVA, conversational behaviour synthesis, personalized interaction

https://doi.org/10.18690/um.ff.11.2022.8 DOI
978-961-286-675-4 ISBN

University of Maribor Press

1 Uvod

Digitalni sistemi se vse bolj uporabljajo za naloge, ki jih običajno izvajajo ljudje. Napredki na področju vmesnikov govorjenega jezika, obdelave naravnega jezika in umetne inteligence so prispevali k vse večji dostopnosti in rabi pogovornih agentov (npr. virtualnih tutorjev, spremljevalcev in asistentov) – sistemov, ki posnemajo človeško interakcijo (Laranjo et al. 2018). Ob hitrem tehnološkem razvoju in vsesplošni digitalizaciji, bolj naravna in posledično bolj razumljiva interakcija z digitalnim vmesnikom predstavlja enega ključnih izzivov moderne interakcije. Človek namreč interakcijske cilje dosega skozi pogovor (Luger et al. 2016).

V zadnjih letih je ravno zato opaziti visoko raziskovalno aktivnost predvsem na področju pogovornih modelov, ki vključujejo animirane ali človeku podobne virtualne like, ki jih imenujemo virtualni pogovorni agenti (pogovorni agenti s telesom, angl. *Embodied conversational agents* - ECA) (Cassell et al. 2001). Bolj znani so zlasti sistemi, ki podpirajo govorni vmesnik kot, denimo, Applova Siri, Googlov Now, Microsoftova Cortana ali Amazonova Alexa (McTear et al. 2016). Novejše raziskave na področju interakcije človek-stroj kažejo, da je najbolj naraven in učinkovit način sintetične interakcije – tudi v visoko tveganih okoljih, kot je zdravstvo (Philip et al. 2020), skrb za duševno zdravje (Provoost et al. 2017), poučevanje (Kramer et al. 2020) in pomoč iz okolice pri samostojnem življenju (Queiros et al. 2018) – takšen, ki temelji na posnemanju naravnih modalnosti, denimo, sinhroniziran govor ter geste in mimika. Razumevanje, sprejemanje in zaupanje informacijam je namreč tesno povezano z nesemantičnimi signali (npr. čustvovanje in upravljanje diskurza), ki jih govorniki posredujejo s pomočjo vizualnih znakov in prozodije (Mlakar et al. 2019; Stal et al. 2020). Lahko bi celo rekli, da je temelj medosebne interakcije sinhrono vključevanje in povezovanje verbalnih z neverbalnimi kanali. Verbalni kanali nosijo simbolno/semantično interpretacijo sporočila z jezikovnimi in para-jezikovnimi značilnostmi interakcije, medtem ko neverbalni kanali služijo kot dirigent komunikacije (McNeill 2016: 4; Kopp in Bergmann 2017).

Jezikovni kanal torej s pomočjo jezikovnih in parajezikovnih elementov opisuje simbolično oz. semantično interpretacijo informacij, med tem ko nejezikovni kanal (npr. telesna govorica) govor organizira (McNeill 2016: 4). Nejezikovni kanal zajema koncepte, kot so prozodija, govorica telesa, čustvovanje ali sentiment. Ti koncepti

so večfunkcijski in delujejo na psihološki, sociološki in biološki ravni ter v vsakem časovnem okvirju (Church in Godin-Meadow 2017). Dejansko predstavljajo osnovo kognitivnih zmožnosti in razumevanja). Tako npr. nasmešek, pogosto s sočasnim smehom, igra pomembno vlogo pri gradnji povezave med udeleženci pogovora, še posebej vzpostavljanju družbenih vezi, ki ustvarjajo vljudno medosebno okolje (Esposito et al. 2015; Ochs et al. 2017). Še vedno pa raba teh subtilnih nejezikovnih signalov v interakciji predstavlja enega ključnih izzivov modernih vmesnikov. Problem izhaja iz sinhronizacije sinhronizacijo in iz zagotavljanja 'pravilne' interpretacije. Nejezikovni elementi namreč niso zgolj preprosti nadomestki jezikovne vsebine, ampak so dejanska sestavina govora. Pri govorjeni interakciji tako nejezikovno obnašanje prispeva več kot 50 odstotkov informacije, pomembne pri gradnji skupne osnove pogovora (Cassel et al. 2001). Še več, več kot 70 odstotkov socialnega pomena pogovora posredujemo z nejezikovnimi koncepti (Birdwhistell 2010).

Večina raziskovalcev se torej strinja, da so neverbalni elementi (tj. geste, mimika in čustva) bistvena sestavina interakcije. Da bi v uporabniku vzbudili odnos, se morajo verbalni in neverbalni elementi vključevati 'pravilno' in skladno s pričakovanji glede na njegove vhodne dražljaje (Ciechanowski et al. 2018). Če se jezikovni in nejezikovni komunikacijski kanali ne poravnajo pravilno, lahko ECA izvede gib brez pomena, ki ga dojemamo kot šum. Še več, predstavljen koncept lahko popači pomen ter z napačno poravnavo ustvari neprimeren družbeni kontekst (McKeown et al. 2015). S povečevanjem stopnje naravnanosti in modalnosti uporabniške izkušnje se bistveno povečujejo človekova pričakovanja in dojemljivost napak (Poria et al. 2017). Bolj, kot je odziv stroja podoben človeškemu, večji in močnejši bo negativni učinek, kadar pride do nesinhronosti (npr. manj naraven glas in manj naravna animacija). V tem oziru je samodejno tvorjenje pogovornega obnašanja še daleč od popolnosti ali naravnosti. Za zagotavljanje dobrih rezultatov je pogosto potrebno človeško posredovanje (Navarro-Cerdan et al. 2018). Konec interakcijskega cikla vedno predstavlja aktivni odziv uporabnika in ne signali ali sama interakcija; česar pa stroj še vedno ni zmožen 'razumeti' ali poustvariti (Opel in Rhodes 2018).

Cilj prispevka je predstaviti pogovorni model za ustvarjanje človeku podobnega pogovora in najti rešitev za čustveno in personalizirano interakcijo med človekom in strojem. Predstavljen model ponuja (i) platformo za ustvarjanje 'pogovornega' znanja in virov, (ii) okvir za načrtovanje in ustvarjanje neverbalnega obnašanja in (iii)

okvir za izvedbo čustvenega neverbalnega obnašanja s pomočjo izražanja stališč, čustev in gest, ki so sinhronizirane z govorom. V drugem poglavju bomo najprej predstavili koncept neverbalnega obnašanja oz. gest kot ključnega mehanizma za pripravo govora in ustvarjanja kohezivnosti interakcije. V tretjem poglavju orišemo pogovorni model EVA, kjer je osrednja zamisel oblikovati različne oblike neverbalnega obnašanja (geste) v povezavi z neoznačenim besedilom in širšim družbenim in pogovornim kontekstom. V četrtem poglavju opisujemo pridobivanje 'pogovornega' znanja in potrebnih virov, ki jih ustvarjamo s pomočjo označevanja spontanega dialoga in s korpusno analizo. Skupaj s petim poglavjem nato opisujemo, kako te vire vključimo v dvostopenjski pristop samodejnega tvorjenja obnašanja. Pričujoč pristop naslavlja (a) problem oblikovanja obnašanja (namen in načrtovanje obnašanja) in (b) problem izvedbe obnašanja (animacija z EVO). Prispevek sklenemo s študijo primera in prikazom sinteze neverbalnega obnašanja pogovornega agenta s telesom EVA (Rojc et al. 2017).

2 Neverbalno obnašanje in geste v interakciji

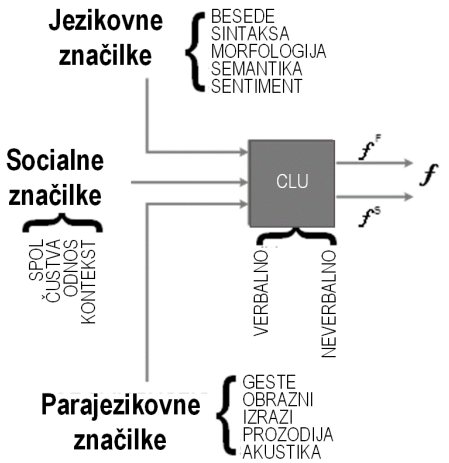
Neverbalni elementi interakcije predstavljajo pomemben del učinkovite komunikacije (Birdwhistell 2010; Trujillo et al. 2018). Vizualni vhodni/izhodni signali so večfunkcijski in delujejo na psihološki, sociološki in biološki ravni v vseh časovnih okvirih, npr. od trenutka do trenutka, ontogenetsko ter se razvijajo skozi čas glede na različna diskurzna okolja (Church in Goldin-Meadow 2017). Gestikuliranje in zmožnost izražanja informacij skozi neverbalne kanale postaja ključna metoda, s katero poosebiti in poenostaviti interakcijo med človekom in strojem. Pri klasični interakciji med dvema osebama so neverbalni signali, ki jih posredujemo skupaj z govorjeno vsebino ali pa tudi brez nje, ključni za kohezivnost diskurza. Lahko bi rekli, da jezikovni deli govorjenega jezika (to so besede, slovnica, skladnja) posredujejo simbolno/semantično interpretacijo sporočila, medtem ko neverbalni del (to so geste, izrazi, prozodija) nosijo družbeno komponento sporočila in so v vlogi dirigenta komunikacije. Posledično se neverbalni signali razlikujejo od splošnih motoričnih dejanj, saj predstavljajo to, kar je izraženo in prepoznano (npr. praskanje, ko razmišljamo ali mahanje v slovo). Allwood et al. (2005) tako raziskujejo povezavo med jezikovnimi in nejezikovnimi signali glede na funkcije komunikacije. Komunikacijo opredelijo kot vsoto glavnega sporočila in upravljanja komunikacije, ki pa jo ločijo na upravljanje interaktivne komunikacije (ICM) in upravljanje lastne komunikacije (OCM), obe pa se lahko izražata z jezikovno in nejezikovno

komunikacijo. Drugi raziskovalci (npr. Hoek et al. 2017; Chui et al. 2018; Lopez-Ozieblo 2018) se osredinjajo na semantiko. Eden najbolj zadevnih in široko rabljenih pristopov k večmodalnosti interakcije je Pierceova semiotična perspektiva (tj. 'pragmatika na dejanski strani') (Peirce 1965), ki proučuje pomen slik in povezanih vizualnih značilnosti pisnega besedila (Carroll et al. 2015, Queiroz in Aguiar 2015).

V nasprotju z omenjenimi pristopi pa semiotika proučuje tudi nejezikovne sisteme znakov. Pomen nejezikovnega obnašanja in pogovornih izrazov tolmači s proučevanjem za komunikacijo ključnih znakov in simbolov. Semiotika po Peirceu je trojiška in kot podskupine loči znake, ki so simboli; ikone in indekse. Vendar pa je njegova klasifikacija predvsem vezana na vizualni stimulus kot glavni nosilec interpretacije. Nasprotno Cooperrider (2017) ločuje med dvema sklopoma neverbalnega obnašanja (t. i. *gest*), in sicer med gestami ospredja in gestami ozadja. Geste ospredja vizualizirajo semantični del sporočila. Kot utemeljuje (Cooperrider 2017), jih rabimo zavedno in jih izvedemo z določeno mero truda. Pojavljajo se skupaj z govorom (npr. vizualizacija za razlago oz. oris) ali povsem brez hkratnega govora (npr. simboli). Zanje značilne kategorije so ikonske geste, zlasti tiste, ki izražajo prostorske odnose, ki lahko izražajo nujne, a jezikovno izpuščene informacije (Melinger in Levelt 2004). Geste ozadja pa so tiste, ki jih izvedemo z najmanj zavedanja oz. povsem podzavestno (Cooperrider 2017). Govorci navadno niso pozorni na njihove podrobne značilnosti in uporabo teh zelo hitro pozabijo. »So v ozadju govorčevega zavedanja, v ozadju poslušalčevega zavedanja in v ozadju interakcije« (Cooperrider 2017: 7). Kljub temu ločnica med tema dvema kategorijama ni tako jasna, saj je za nekatere tipe gest lahko zelo zabrisana (glej Cooperrider 2017: 193).

Neverbalni signali (ki vključujejo tudi in predvsem geste) dejansko omogočajo uporabnikom, da se na sintetični signal odzovejo naravno, začnejo izražati čustva in elemente, ki jih sicer vključujejo v medosebno interakcijo. Negativna plat pa seveda leži v dojemljivosti napak (pojav t. i. *uncanny valley*). Bolj kot je odziv naprave človeški, večja pričakovanja povzroči. Posledično bo odziv uporabnikov na pomanjkljivosti bistveno bolj negativen. Poglaviten vir morebitne neskladnosti (negativnega dojemanja) izhaja iz manka pogovornega znanja. Pogovorno znanje obsega razumevanje komunikativnih signalov, od jezikovnih, parajezikovnih do družbenih (Slika 1). Za razliko od jezika pa odnosov med signali ne moremo opisati z univerzalnimi pravili (kot je slovnica), ki bi ustvarili končen nabor vzrokov in

napovedali učinke. Da bi ustvarili delujoče, človeku podobne, odzive, moramo te signale spojiti v svoje področje človeku podobnih odzivov, za katere pa potrebujemo različne vire znanja. Kot prikazuje Slika 1, želimo oblikovati kompleksno funkcijo $\mathcal{f}$, ki je izražena kot fuzija informacije, ki jo lahko zajamemo skozi procesiranje naravnega jezika (angl. *Natural Language Processing* – NLP) in skozi procesiranje jezika telesa (angl. *Embodied Language Processing* – ELP). Obe domeni informacije pa skozi funkcijo $\mathcal{f}$ zlijemo v t. i. razumevanje pogovornega jezika (angl. *Conversaitonal Language Understanding* – CLU).



Slika 1: Razumevanje jezika kot model večmodalne fuzije
Vir: lasten.

Model predlagan na Sliki 1 temelji na konceptu 'večkanalne' predstavitve ideje, sočasno skozi osnovi avdio in video kanalov. Fuzija $\mathcal{f}$ se najprej oblikuje na kognitivni ravni s pomočjo simbolne fuzije ($\mathcal{f}^S$), kasneje na predstavitveni ravni s pomočjo fuzije oblike. Simbolno raven funkcije fuzije opredelimo kot

$$\mathcal{f}^S = f(L, P, S) \tag{1}$$

Tako vzpostavimo simbolično povezavo med različnimi signali iz različnih domen kognitivne lingvistike kot, denimo, sama lingvistika L, paralingvistika P in socialni kontekst S. Narava fuzijske funkcije f in korelacija z različnimi posameznimi signali in njihovimi prispevki je večinoma že zelo dobro opisana z novejšimi teorijami

kognitivne lingvistike in komunikativnega obnašanja. Vseeno pa ne moremo ozkega področja znanja kar 'zliti' v skupno strategijo. Da bi razumeli pomen $\mathcal{P}^S$ na simbolni ravni, predlagamo tolmačenje, ki je v skladu z McNeillovo (2008) teorijo skupne točke raste. To tolmačenje, izpostavljeno v Mlakar et al. (2019), izkorišča tako semiotiko po Peirceu (1965), nejezikovno obnašanje po Ekmanu in Friesenu (1971), kot kinezijo (Birdwhistell 2010; Maricchiolo et al. 2012). Skladno s teorijo upravljanja komunikacije (Guitella et al. 2009; Allwood 2014; McNeill et al. 2015) pa klasifikacija prav tako vključuje funkcije diskurza. Nejezikovni komunikacijski namen (NCI) v diskurzu zajema zgolj premikanje, ki služi nekemu komunikativnemu namenu (tj. prispeva k ustvarjanju pomena). Klasifikacija razlikuje med petimi različnimi vrstami takšnega gibanja, in sicer: ilustratorji, regulatorji ali adapterji, deiktiki ali kazalci, simboli ali emblemi in udarci.

Ilustratorji označujejo neverbalno obnašanje, s katerimi govorci ilustrirajo, kar govorijo. Sestavljeni so iz podskupine orisnih ilustratorjev, ki označujejo nejezikovno obnašanje, ki prikazujejo konkretno lastnost spremljajoče jezikovne vsebine, zaradi česar imajo v govoru jasno nanašalnico; podskupine ideografov, ki se nanašajo na konkretizacijo abstraktnega z določeno obliko; in prostorsko/dimenzijska podskupina, ki se nanaša na prostorske gibe, ki orišejo ali prikazujejo dimenzijska razmerja.

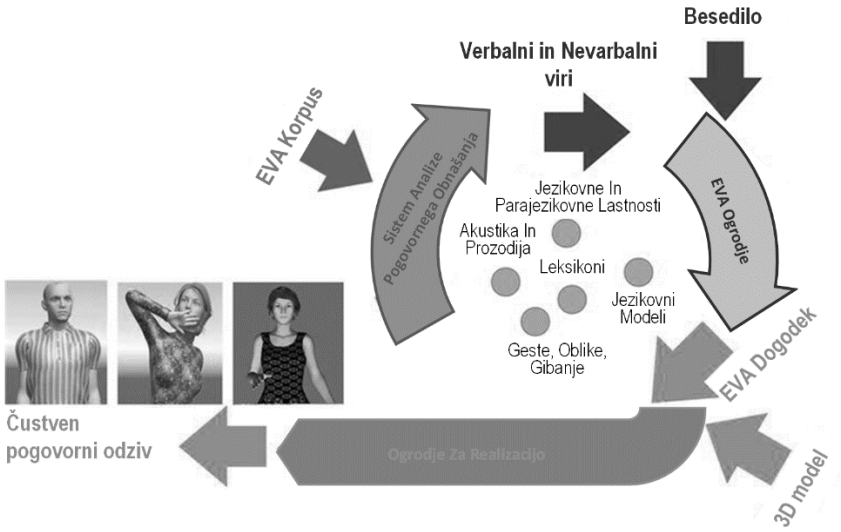
Regulatorji ali **adapterji** opredeljujejo nejezikovno obnašanje, ki lahko vsebuje nanašalnico v govorni vsebini ali strukturi. Primarno jih rabimo za prikaz metakonceptov, s katerimi upravljamo komunikacijo, izražamo mišljenje, napetost, negotovost ali druge družbene signale. Regulatorje še nadalje razdelimo na podskupino lastnih adapterjev, ki zajemajo obnašanje ob iskanju; podskupino regulatorjev komunikacije, ki vključujejo sekvenciranje in menjavo govornih vlog; podskupino regulatorjev afekta; podskupino manipulatorjev ter podskupino za družbene funkcije in norme.

Deiktiki se navezujejo na nejezikovno obnašanje, s katerimi se nanašamo na dejanske ali abstraktne zadeve (npr. predmete, kraje ali kazanje nazaj, kadar želimo prikazati preteklost). Četudi imajo dejansko nanašalnico v govoru, pa nejezikovno obnašanje ni nujno časovno povezano z njimi. Deiktiki zajemajo podskupino kazalcev; podskupino indeksov ali nanašalnih kazalcev in podskupino številčnikov. Skupina

NCI **simbolov** vključuje vse simbolne geste. Pogosto so kulturno specifične in imajo neposreden jezikovni prevod. **Udarci** so odrezani zamahi, ki dajejo ritem, ustvarjajo poudarke in s tem označujejo pomembnost kot tudi pritegnejo pozornost.

3 Pogovorni model EVA: Model za tvorjenje ekspresivnega sintetičnega obnašanja

Teoretičen model pogovornega prostora je orisan na Sliki 2. Model je zasnovan kot sredstvo, ki omogoča: (a) proučevanje narave naravnega obnašanje med sogovorniki (ljudmi); (b) ustvarjanje 'pogovornega' znanja v obliki lingvističnih, paralingvističnih jezikovnih in nejezikovnih značilk; (c) in preskušanje teorij skozi apliciranje znanja v različnih pogovornih situacijah.



Slika 2: model EVA: model za tvorjenje pogovornega obnašanja in oblikovanje čustvenih pogovornih odzivov na sintetičnih pogovornih agentih
Vir: lasten.

Model temelji na zamisli, da sta jezikovna in neverbalna poravnava ter sinhronizacija gonilni sili za afektivno in socialno interakcijo. *Sistem Analize Pogovornega Obnašanja* smo oblikovali za analizo prepletanja lingvističnih in parajezikovnih značilk z gestami, ki ga opazujemo v spontani interakciji med več govorci. Analiza temelji na video posnetkih večmodalnega korpusa EVA (Mlakar et al. 2019) in označevalni

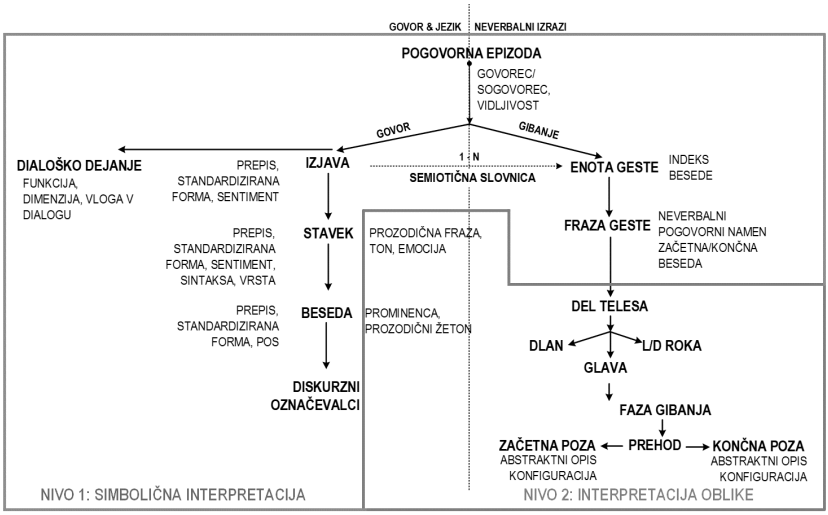
shemi EVA, razviti za opis kompleksnih odnosov neverbalnega obnašanja (predlagan v Mlakar et al. 2014). Korpus EVA je bil oblikovan za tvorjenje virov, ki jih potrebujemo za načrtovanje in ustvarjanje pogovornega obnašanja, tako jezikovnega dela (npr. sinteze od besedila v govor (angl. *Text-to-Speech* – TTS)) kot neverbalnih komponent (sinteza pogovornega obnašanja). Ključne vire predstavljajo leksikoni, jezikovni modeli, semiotična slovnica komunikativnega namena, leksikon pogovornih oblik, geste in gibi, akustične in prozodične značilnosti ter druge jezikovne in parajezikovne značilke (mdr. segmentacija na besedo/zlog, tip stavka, sentiment). Ključna zamisel ogrodja EVA (Rojc et al. 2014) je izkoristiti prej omenjene vire in načrtovati pogovorno obnašanje, ki lahko v interakciji izzove socialni/emocionalni odziv uporabnika. Rezultat ogrodja EVA prestavlja pogovorni niz, ki vsebuje sinhrono predstavitev informacije skozi govor in pogovorno obnašanje. Ker je načrtovano obnašanje že prilagojeno naravi in zmožnostim virtualnega agenta, EVA dogodek predstavlja direkten vhod za realizacijsko ogrodje EVA (Mlakar et al. 2018). To temelji na premisi, da je naravna večmodalna interakcija veliko več kot govor, ki ga spremljajo ponavljajoči se gibi okončin in obraza. Vloga tega ogrodja je animirati EVA dogodke s pomočjo 3D-modela. Realizacijsko ogrodje tako EVA dogodke preoblikuje v večmodalne, človeku podobne, pogovorne sekvence.

V naslednjih poglavjih bomo podrobneje predstavili tri glavne komponente pogovornega modela EVA, in sicer korpus EVA, ogrodje EVA ter realizator EVA.

4 EVA Korpus: Označevanje, segmentacija in kvantifikacija pogovora v pogovorne signale

Ključni cilj sheme na Sliki 3 je: i) na simbolni ravni identificirati pomene, ki jih je mogoče razbrati iz neverbalnih izrazov, kot funkcijo jezikovnih, parajezikovnih in socialnih signalov (npr. kdaj in kako gestikulirati) in ii) identificirati fizično naravo uporabljenih neverbalnih elementov (npr. kako izraziti), in sicer na ravni interpretacije nejezikovnih oblik. Koncept označevanje je tako dvonivojski. Prvi nivo na Sliki 3 imenujejo simbolična interpretacija. Uporabljamo ga za analizo interpretacije prepletanja različnih pogovornih signalov (npr. dialoška dejanja, geste, skladnja, diskurzni označevalci) in razumevanja, kako sodelujejo pri oblikovanju večmodalne predstavitve informacije. Simbolno označevanje omogoča identifikacijo

in podroben opis narave komunikativnih dejanj, ki se odvijajo med izmenjavo informacije. Označevanje oblike (oz. nivo interpretacije oblike) pa opisuje, kako izvesti neverbalne elemente, da pravilno "vizualizirajo" idejo/namen. Vizualizacijo dosežemo skozi prozodijo govora ter oblike in gibe, ki se pojavljajo ob govoru; npr. kako s premiki rok (leva in desna roka ter dlani), obraznimi izrazi, gibanjem glave in usmeritvijo pogleda poudarimo pomembne segmente, izražamo strinjanje/razumevanje ali sporočamo, da smo zaključili s podajanjem informacije in prosimo za odziv.



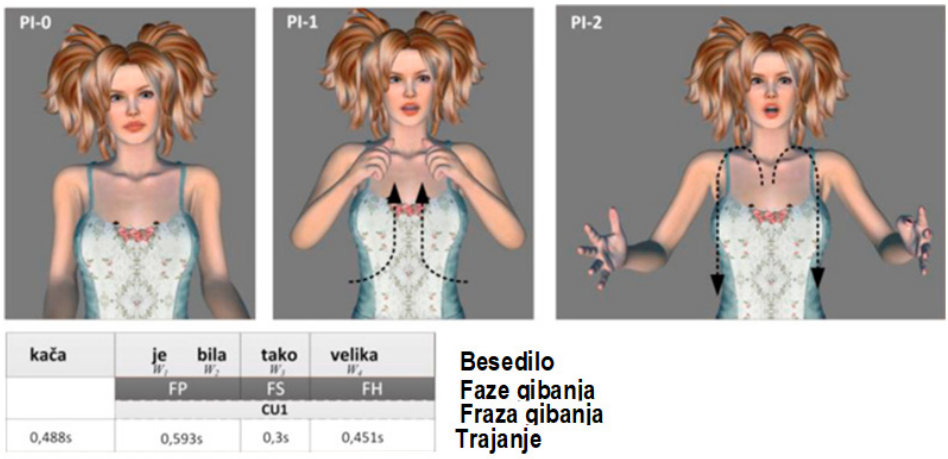
Slika 3: Topologija označevanja pogovornega obnašanja v EVA Korpusu: Verbalni in neverbalni kontekst pogovornih epizod.

Vir: lasten.

Simbolna interpretacija se tako ukvarja izključno z namenom neverbalnih komponent, ki ga klasificiramo z neverbalnim komunikacijskim namenom (NCI). Interpretacija oblik pa se ukvarja izključno z načinom izvedbe in vizualizacije. Da bi ju povezali, v predlaganem modelu uvedemo pojem semiotične slovnice. V njej je namen predstavljen kot razred/podrazred NCI. Vsak NCI pa zajema možne izvedbe neverbalnega obnašanja, s katerimi so govorci/poslušalci dosegli želeni namen. Prednost tega je, da je semantični prostor močno zmanjšan. Poleg tega pa je število eksplicitnih korelacij med besednimi sekvencami (besedami in frazami) in

motoričnimi spretnostmi zmanjšano na skupek razredov NCI nekaj podskupin. Gestikon predstavlja realizacijo koncepta semiotične slovnice.

Interpretacija oblike opisuje realizacijo pogovornega namena s prozodijo in telesnim gibom. Glavni cilj nivoja 2 je zagotoviti podroben opis, ki je blizu tako fizični realnosti (človeku) kot entiteti, ki jo realizira (npr. pogovorni agenti s telesom). V predlagani shemi so deli telesa ključni pri opazovanju in označevanju oblike. Pri tem sprejemamo zamisel utelešene kognicije, po kateri senzorično-motorične zmožnosti (zmožnost telesa, da se odzove na dražljaje z gibi), telo in okolje igrata pomembno vlogo pri razmišljanju. Označevalna shema EVA zato razlikuje med dlanmi, rokami, glavo in obrazom. Struktura oz. prozodija gibanja je nato opisana v obliki faz gibanja. Fazo gibanja, skladno s Kita et al. (1997), opišemo kot eno izmed petih, in sicer kot obvezna faza udarca ali kot opcijske pripravljalna faza, faza zadrževanja ter faza umika. Da bi dosegli izvedljivost 'giba' na dani entiteti, vsako izmed faz gibanja opišemo kot par začetne poze (P_S) in končne poze (P_E) ter pot T, po kateri je bil prehod med P_S in P_E izveden (točkasta puščica na Sliki 4) (Rojc et al. 2014). Pot T predstavimo skozi parametričen opis premika, ki vključuje zaporedje enostavnih vzorcev, kot so: linearno ali kot lok. Primer, kako uporabiti semiotično slovnico kot vir simbolične interpretacije in gestikon kot vir za realizacijo pogovornega namena, je podan na Sliki 4.



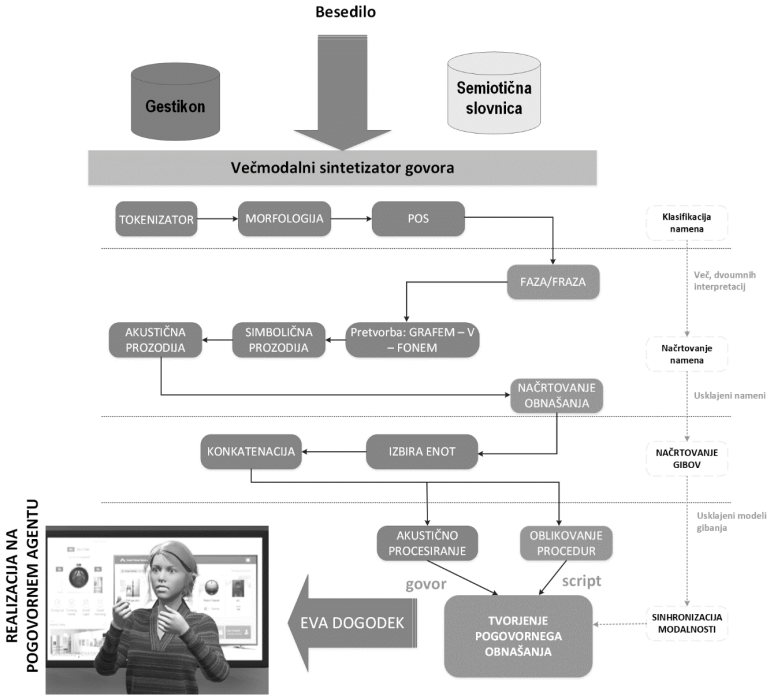
Slika 4: Vizualizacija pogovornega obnašanja na ECA ob uporabi semiotične slovnice in gestikona.
Vir: lasten.

Za 'vizualizacijo' stavka »Kača je bila tako velika« smo animirali eno od možnih interpretacij; serijo ikonsko metonimičnih gest, ki so rezultat sekvenc pomožnih glagolov ('je bila'), prislovov ('tako') in pridevnikov ('velika'). Kot pomensko jedro je bil prepoznan pridevnik 'velika'. Pripravljalna faza (F_P) je opredeljena z izvedbo premika med pozo PI-0 in PI-1 med besedama 'je' in 'bila'. Predvideno trajanje F_P je t= 593 ms. Vsebinsko najpomembnejša faza udarca (F_S) se bo izvedla ob prozodično najbolj poudarjeni besedi 'tako', njena oblika pa bo vizualizirala pomensko jedro, pridevnik 'velika'. Trajanje prehoda med PI-1 in P-2 je tako predvideno s časom t = 300 ms. Po izvedbi je algoritem, predstavljen na naslednjem poglavju, napovedal še fazo zadržanja, ki se izvede ob izgovarjavi pomensko jedro in traja 451 ms.

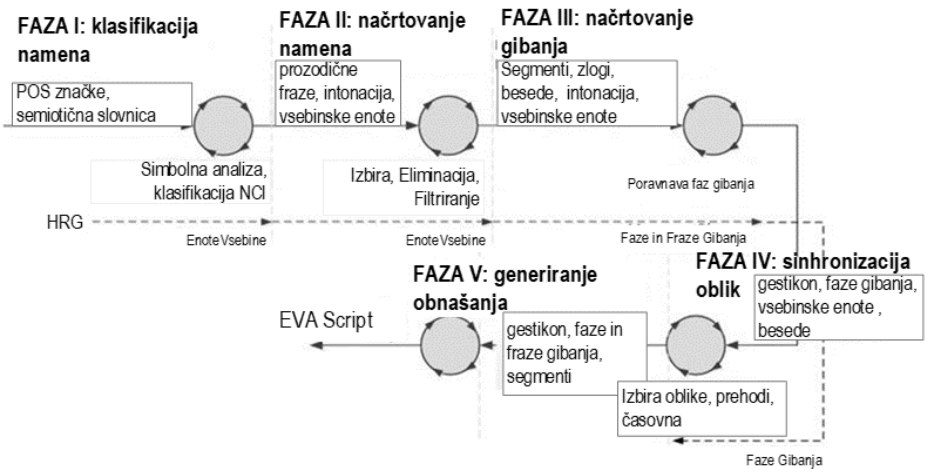
5 Ogrodje EVA: Generator pogovornega obnašanja

Algoritem načrtovanja pogovornih sekvenc (Slika 5) samodejno ustvari pogovorne sekvence, kot je sekvenca na Sliki 4, tako da za dano besedilo ustvari govor in neverbalne elemente. Njegova zasnova je podana na Sliki 5. Model temelji na štirih korakih: klasifikaciji namena, načrtovanju namena, načrtovanju gibov in sinhronizaciji modalnosti.

Proces sinteze govora, ki ga izvede algoritem na Sliki 5, pretvori splošno besedilo v pogovorni dogodek. Načrtovanje neverbalnih elementov je vključeno direktno v proces sinteze in lahko direktno izkorišča faze sinteze kot vir jezikovnih in prozodičnih značilnosti, ki so nujne za načrtovanje in sinhronizacijo neverbalnega obnašanja. Najprej algoritem izvede proces klasifikacije namena, ki identificira naravo govorne vsebine s pomočjo klasifikacije vzorcev besedila v semiotično slovnico. Pogovorni namen vhodnega besedila se opredeli v obliki klasifikacije glede na semiotični razredi. Rezultat prvega koraka je nabor možnih interpretacij vhodnega besedila. Proces načrtovanja namena nato vključuje izločanje interpretacije, ki ustreza prozodični strukturi govora. Za izbrani namen je nato iz Gestikona treba izbrati najustreznejšo interpretacijo. Postopek se izvede v koraku načrtovanja gibov, ki s pomočjo mehanizma cenilk izbere najustreznejšo izvedbo (t. i. model gibanja). Na koncu mora algoritem izvesti še časovno sinhronizacijo; tj. prilagoditi posamezne faze gibanja verbalnemu delu pogovorne sekvence. Slika 6 podrobneje orisuje postopek, ki je sestavljen iz petih faz.



Slika 5: Algoritem za načrtovanje in tvorjenje večmodalnih pogovornih sekvenc
Vir: lasten.



Slika 6: Algoritem za ustvarjanje čustvenega neverbalnega obnašanja.
Vir: lasten.

V **prvi fazi**, ki jo imenujemo klasifikacija namena, je vhod besednovrstno označeno (POS) besedilo in semiotična slovnica. Semiotična slovnica se uporablja za pripenjanje posameznih morfosintaktičnih sekvenc besedila na zadevne parametrične opise NCI. Algoritem išče najdaljše morfosintaktične sekvence, ki jih lahko najde v semiotični slovnici, pri čemer pa upošteva naslednji dve pravili:

Če je ob določeni besedi sekvenca x_A vrednost $x_A(S) \subseteq x_B(S)$, pri čemer spadata obe sekvenci k isti semiotični skupini, moramo sekvenco x_A zavreči.

Če je sekvenca x ob besedi j že zajeta v vsebinski enoti, ki se prične z besedo i in če ima enak pogovorni namen ($i < j$), jo zavržemo.

Vsebinska enota (CU) predstavlja parametrično interpretacijo sporočila v stavku/izreku. Stavki/izreki lahko vsebujejo več interpretacij in interpretacija CU se lahko delno ali povsem pokriva, s čimer pa prihaja do dvoumnosti in številnih neskladij. Posledično je treba v **prvi fazi,** pri načrtovanju namena, ta neskladja in dvoumnosti razrešiti s pomočjo integracije, eliminacije in izbiranja. Zato uporabljamo prozodične informacije (izstopanje, poudarek, prozodične fraze), kot to predvidevajo modeli TTS. Uporabljena prozodična informacija vključuje zloge z označenim naglasom, in sicer oznake PA, ki je najbolj izstopajoč, in NA, naglas besede. Vsebinske enote nato obravnavamo z naslednjimi pravili:

Vsaka vsebinska enota (CU) mora vključevati najbolj izstopajoč zlog (PA) znotraj dane prozodične fraze (B2 ali B3), razen pri naštevalnih primerih.

Vsak element CU se mora nahajati znotraj prozodične fraze (B2 ali B3).

Vsako prozodično frazo lahko predstavimo z največ enim konceptom neverbalnega gina, tj. ne več kot en element CU.

Kadar element CU vsebuje semiotični razred naštevanja, morajo meje CU ostati nespremenjene (mej prozodičnih fraz ne upoštevamo).

Element CU vključuje zlog PA, ki se mora nahajati znotraj mej prozodične fraze $B2{:}PA \in B2 \wedge PA \in CUi$.

V **tretji fazi**, ki jo imenujemo načrtovanje gibanja, opredelimo modele gibanja, s katerimi lahko vizualiziramo dani CU. Pri tem se ustvari model gibanja H, ki predstavlja animirano sekvenco oblik/položaja telesa, ki skupaj predstavlja neverbalni izraz. Za vsak H moramo določiti vsaj eno fazo udarca F_S, ki je poravnan z akustično prozodijo, kot jo definira TTS Za opredelitev faze udarca F_S smo uporabili naslednji pravili:

Faza udarca Fs se zmeraj nahaja ob PA besede in se konča skupaj s pripadajočim PA zloga.

Če beseda, ki predstavlja semiotični indikator I, za specifično CU ne vsebuje PA zloga, se v ta namen upošteva NA zloga.

V modelu gibanja H se zlogi, ki se pojavijo pred fazo udarca F_S, uporabljajo za pripravo faze gibanja F_P, segment 'sil', ki je tik pred prvim zlogom F_S, pa se lahko uporabi za fazo zadržanja gibanja F_H (zadržanje pred udarcem). Zlogi za Fs pa se lahko uporabijo za fazo umika F_R. V tem smislu strukturo obnašanja uporabljamo s pomočjo naslednjih pravil:

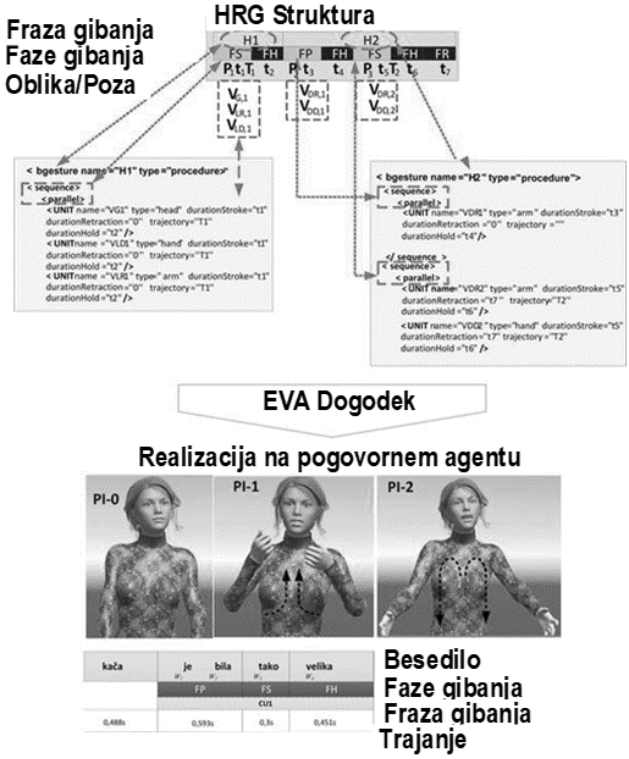
Faza pripravljanja Fp se začne pred fazo udarca Fs in traja od zloga NA do začetka faze udarca Fs.

Segment 'sil', ki lahko ima vrednost trajanja med besedami pri fazi pripravljanja Fp in fazi udarca Fs od nič in naprej, pa predstavlja t. i. fazo zadržanja pred udarcem (angl. hold before stroke), ki (če se pojavi) predstavlja pripravljeno idejo vsebine.

V **četrti fazi**, ki jo imenujemo sinhronizacija oblike, se gibanje časovno poravna s časovnimi značilkami jezikovne informacije (trajanje fonemov in premorov). Da bi določili najboljšo obliko V (ali položaj P) ter usmeritev T realizacije neverbalnega obnašanja, v *Gestikonu* izvedemo poizvedbo, ki temelji na morfosintaktičnih sekvencah, modelih gibanja in trajanja faz gibanja. Tako sprožimo iskanje možne konfiguracije oblike V znotraj F_S faz. Tako dobimo nabor možnih položajev telesa P za vsak F_S. Te položaje nato ocenjujemo s pomočjo funkcij primernosti (Rojc in Kačič 2011). Če v Gestikonu ni ujemanj, se izbere nabor najbolj primernih položajev v izbranem modelu CART (angl. classification and regression tree), pri čemer vsaki pripišemo najbolj primeren položaj P. Ko smo opredelili vse kandidate za položaj za vse predlagane F_S, se opredelijo še položaji za F_P, F_R in F_H, pri čemer za oblikovanje

prehoda med dvema položajema upoštevamo časovno strukturo, opredeljeno s trajanjem govornih enot, semiotičnega razreda, vrste faze gibanja, morfosintaktičnih oznak, prozodičnih značilk znotraj fraze.

V zadnji, **peti fazi**, ki smo jo poimenovali ustvarjanje neverbalnega obnašanja, predlagani model gibanja pretvorimo v proceduralni opis animacije. Vsaka faza gibanja se pretvori v simbolno, prozodično in prostorsko koherentno gibanje posameznega opazovanega telesa. Vsak model H opišemo v proceduralni sintaksi EVA-Script (Rojc et al. 2017) kot blok <*bgesture*>. Faze udarca F_S znotraj bloka <*bgesture*> predstavimo kot sekvence, bloke <*sequence*>. Fazi zadržanja F_H in faza umika F_R pa skozi atributa trajanja znotraj blok <*bgesture*> ali bloka <*sequence*>. Pretvorba gibanja modela H v dogodek EVA (neverbalno obnašanje, zapisano kot EVA-Script) je orisana na Sliki 7.

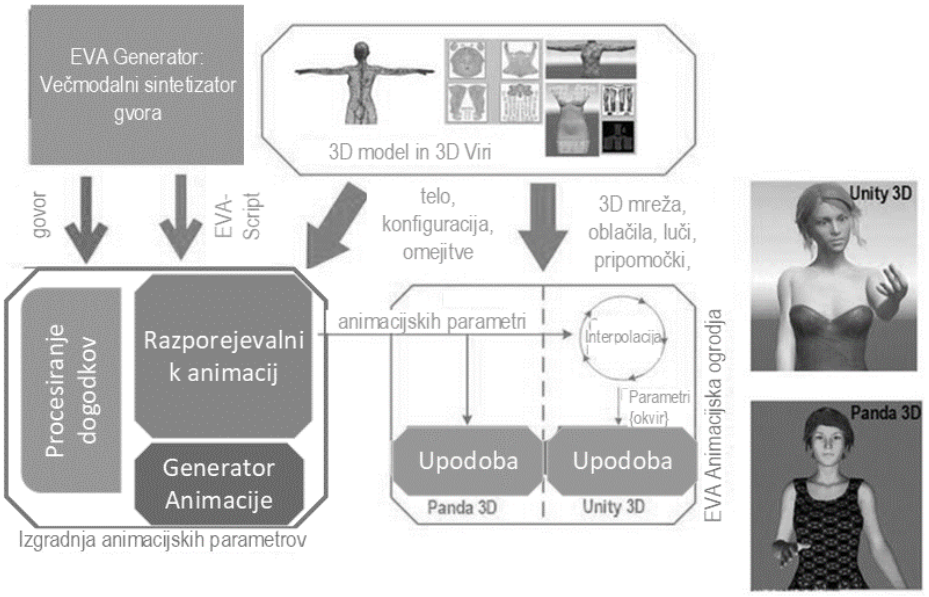


Slika 7: Realizacija stavka iz Slike 4 s pogovornim agentom EVA.
Vir: lasten.

Dejanska konfiguracija je opisana preko 3D-konfiguracije sklepov, ki je opisana kot element <*UNIT*> znotraj bloka <*sequence*>. Kadar so elementi <*UNIT*> združeni v bloku <*parallel*>, to označuje njihovo sočasno izvedbo. V nasprotnem primeru se konfiguracije oz. premiki iz enega 3D-sestava v drugi 3D-sestav izvedejo zaporedno. V naslednjem poglavju bomo predstavili realizator pogovornega obnašanja, s katerim proceduralni zapis v EVA-Script pretvorimo v animirano pogovorno sekvenco.

6 EVA Realizator pogovornega obnašanja: Animacija in vizualizacija govora na pogovornem agentu

Za animacijo načrtovanega obnašanja smo uporabili lastno razvito ogrodje realizacije – EVA (Mlakar et al. 2017. Ogrodje omogoča, da stroji z uporabnikom vzpostavijo bolj oseben stik, in sicer v obliki človeku podobne entitete, realizirane z večmodalnimi modeli interakcije, ki temeljijo na konceptu pogovora. Ogrodje je prikazano na Sliki 8:



Slika 8: Ogrodje za Realizacijo pogovornega obnašanja
Vir: lasten.

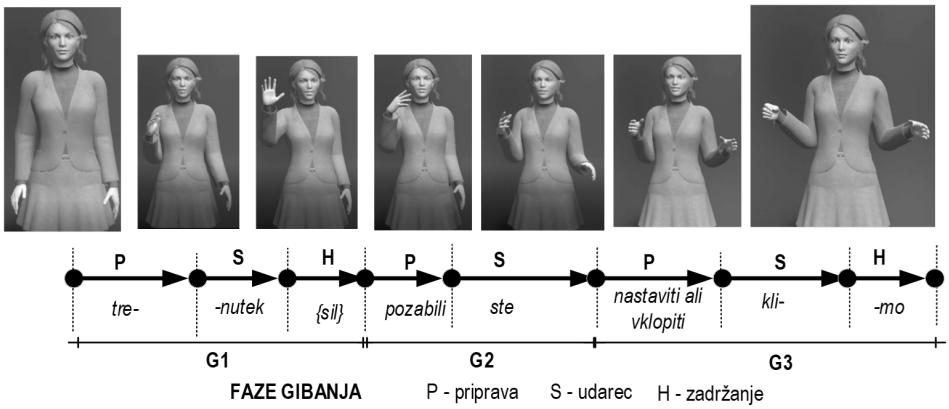
V grobem ogrodje sestavljajo komponente za gradnjo animacijskih parametrov, animacijska ogrodja za realizacijo animacije in 3D-viri, vključujoč mrežni model podobe agenta, oblačil in drugih elementov scene. Komponenta za gradnjo animacijskih parametrov se uporablja za pretvorbo neverbalnih dogodkov v nize animacijskih parametrov, ki jih animacijska ogrodja lahko pretvorijo v dejansko animacijo in prikažejo uporabniku. Da bi lahko formalni zapis obnašanja iz Slike 7 animirali, je treba EVA-Script zapis pretvoriti v nize animacijskih parametrov, ki jih je izbrano ogrodje za animacijo zmožno vizualizirati. To lahko dosežemo z uporabo neverbalnih značilk, opisanih v neverbalnih dogodkih pri 3D-virih, do katerih dostopa pretvornik. Komponente za gradnjo animacijskih parametrov prevede EVA-Script v animacijske parametre tako, da posamezne elemente v formalnem zapisu poveže z ustrezno kontrolno enoto agenta, vključujoč časovne (trajanje faze udarca, zadrževanja in umika) in prostorske značilnosti (končni položaj enote). Prevod se izvede z generatorjem animacije. Razporejevalnik animacij pa ustrezno interpretira sosledje v formalnem zapisu in pretvori v t. i. animacijski graf, ki ga realizira izbrano animacijsko ogrodje. Kot prikazuje na Sliki 8, predlagan model podpira dve animacijski ogrodji, in sicer Panda 3D in Unity 3D. Bistvena razlika med njima je v načinu izračuna interpolacije. Panda 3D interpolacijo med začetnim in končnim položajem izvede vnaprej. V Unity 3D pa se za izračun izvede na koncu vsakega video okvira. Izračun konfiguracije v naslednjem okviru daje bistveno večjo možnost glajenja animacije in reaktivnost obnašanja, saj lahko animacijo spremenimo ob vsakem koraku, celo med izvajanjem nekaterih korakov/sekvenc. Za gladek prehod razporejevalnik ne izvaja časovne prerazporeditve, ampak samo zamenja obstoječe segmente z novimi konfiguracijami. Posledično je virtualni lik bolj odziven in se lahko nemudoma odzove na spremembe v pogovornih, okolijskih in drugih kontekstih.

7 Demonstracija sinteze pogovornega obnašanja z modelom EVA

Podrobneje smo preučevali zaznano kakovost predstavitve informacij s posameznimi študijami primerov rabe. Primeri, ki sledijo, ponazarjajo praktične primere rabe:

Primer 1: Sinteza pogovornih izrazov z ECA EVA v okolju Pametni dom

ECA EVA: *»Tre**nutek**! Pozabili **ste** nastaviti ali vklopiti **kli**mo.«*



Slika 9: Sinteza pogovorne sekvence s pomočjo ECA EVA.
Vir: lasten.

Pri izreku zgoraj je model ustvarjanja obnašanja uporabil bazo znanja (tj. pravila) in vire, s katerimi je načrtoval in ustvaril opozorilno sekvenco v obliki naravnega jezika, ki vključuje sintetičen govor in tri prozodično poravnane pogovorne geste – G1, G2 in G3, kot je razvidno iz Slike 9. Vse tri geste vključujejo usmerjenost pogleda, gibanje leve in desne roke (ter dlani), pri čemer pa je desna roka prevladujoč del telesa. G1 prikazuje besedo 'trenutek' v kontekstu poudarka (tj. podrazred semiotičnega namena udarcev, IB). Faza udarca se zgodi na lokalno najbolj izstopajočem delu govorne sekvence (tj. 'trenutek'). Oblika dlani na koncu udarca je ena od značilnih oblik v Korpusu EVA, ki se pojavlja skupaj z udarci, ki pa so povezani s konteksti kot čakanje, zadrževanje ali ustavitev. Ker govorjeni vsebini sledi kratka tišina ({sil}) in je izrek vzkličen, se je model odločil tišino prikazati s fazo zadrževanja gibanja. G2 predstavlja tipičen referenčen izrek (tj. podskupina semiotičnega namena nanašalnih deiktikov, Dr), ki se prej nanaša na sogovornika kot pa na tretjo osebo ali predmet. Prozodično gledano je 'ste' najbolj izstopajoč del sekvence, kar je tudi razlog, zakaj se faza udarca pojavi skupaj z izrekom besede 'ste'. Oblika ob koncu sekvence udarca omogoča eno od možnih fizičnih predstavitev NCI, ki cilja na sogovornika, tj. na osebo, s katero je v neposredni komunikaciji. Zadnja, tretja gesta, pa je prej utrip kot udarec. Kot mašilo je bil izbran, da sledi prozodični strukturi govorjene vsebine.

Za vsako podštudijo smo odzive ovrednotili z eksperimentom dojemanja. Sodelujoče smo prosili, naj ocenijo odvisne spremenljivke (glej Tabelo 1), tako da opišejo kakovost predstavitve, in sicer na 5-stopenjski lestvici Likert. Poleg kakovosti gest (npr. oblika, dinamika, tekočnost, sinhronizacija) pa so prav tako opazovali, kako razumljiva je bila predstavljena vsebina. Ob upoštevanju obeh primerov (besedilo in govor ter ECA z gestami) so sodelujoči opredelili splošno dojemanje delovanja in podobnosti človeku, izraženo z zadnjo, sedmo, odvisno spremenljivko, ki smo jo prav tako ovrednotili na 5-stopenjski lestvici Likert. Omenjenih sedem meril, ki smo jih ocenjevali, je navedenih v Tabeli 1.

Tabela 1: Lestvica Likert z vrednostmi za kontekstno odvisno ovrednotenje večmodalnega izhoda

Odvisna spremenljivka	**Vprašanje**	**Lestvica**
Ujemanje vsebine (C1)	*Ali geste pravilno tolmačijo jezikovne informacije?*	1 – ne
		5 – zelo verjetno tolmačenje
Sinhronizacija oblike (C2)	*Se vam zdijo oblike v različnih govornih segmentih primerne?*	1 – ne
		5 – zelo verjetna korelacija
Tekočnost (C3)	*Je bilo gibanje tekoče?*	1 – ne
		5 – zelo tekoče
Dinamika (C4)	*Ali je bila hitrost predstavljene vsebine primerna?*	1 – prepočasi
		5 – prehitro
Zgoščenost (C5)	*Je bilo dovolj gibanja?*	1 – premalo
		5 – preveč
Razumevanje (C6)	*Kako razumljiva je bila predstavljena vsebina?*	1 – nerazumljivo
		5 – jasno razumljivo
Živahnost (C7)	*Kako bi v splošnem ocenili izkušnjo? Se vam zdi obnašanje bolj naravno in živahno glede na običajne vmesnike (brez govora in brez ECA)?*	1 – nenaravno
		5 – blizu človeku podobnemu

Merilo ujemanja vsebine (C1) označuje, ali neverbalno obnašanje in povezane geste predstavljajo ujemajoč govorni segment, medtem ko sinhronizacija oblike (C2) ugotavlja, ali je bila izbrana ustrezna vizualna predstavitev za dan segment. Merilo tekočnosti (C3) kaže na stopnjo tekočnosti sinhroniziranih gest, gibov in prehodov. Merilo dinamike (C4) smo uporabili za ocenjevanje hitrosti gest/izgovarjave. Merilo zgoščenost (C5) kaže na razporeditev ustvarjenih gest (ali je vključenih preveč ali premalo neverbalnih elementov). Merilo razumevanja (C6) smo uporabili za preverjanje, ali je bila sintetizirana vsebina (govor ali govor in geste hkrati) jasno

razumljiva in ali so posamezni segmenti bili sintetizirani/proizvedeni tako, da so slabše razumljivi. Zadnje merilo živahnosti (C7) pa smo uporabili za preverjanje, kako, če sploh, neverbalno obnašanje prispeva k dojemanju naravnosti ECA EVA.

8 Zaključek

Naravna komunikacija zajema veliko variacij obnašanja, ki se povezujejo na dinamičen in zelo nepredvidljiv način. Vključuje tudi različne družbene in medosebne signale, ki obarvajo končni rezultat. Večmodalnost v interakciji ni le nek dodatek ali stil predstavitve informacije. Večmodalnost seže močno čez semantiko in celo čez semiotične artefakte. Močno prispeva k predstavitvi informacij kot tudi medosebni in besedilni funkciji komunikacije. V tem poglavju smo orisali pristop k samodejni sintezi bolj naravnih, človeku podobnih, odzivov, ki so ustvarjeni na podlagi pogovornega modela EVA. Predstavljen model vsebuje tri povezana in ponavljajoča se ogrodja. Prvo ogrodje obsega pogovorno analizo, s katero proučujemo spontane dialoge med več osebami, da bi ustvarili različne tipe pogovornih virov (od pravil in smernic do kompleksnih večdimenzijskih značilk). Drugo ogrodje nato vključuje vseobsegajoč algoritem za sintezo afektivnega neverbalnega obnašanja, ki temelji na naključnem in neoznačenem besedilu. V nasprotju s povezanimi raziskavami pa predlagan algoritem dopušča, da je pogovorno obnašanje poganja hkrati prozodija in besedilo ter da je oblikovano z različnimi dimenzijami situacijskih, med- in znotrajosebnih kontekstov. Predvideno obnašanje, ki je dobro sinhronizirano z jezikovno ustreznico, pa moramo predstaviti uporabniku na najbolj učinkovit način. Zato tretje ogrodje predlaganega modela vključuje realizator neverbalnega obnašanja. V našem primeru smo se odločili, da prednosti sodobnega orodja za 3D-modeliranje in igralnih pogonov združimo z najsodobnejšimi koncepti realizacije obnašanja, da vzpostavimo učinkovito in visoko odzivno ogrodje, s katerimi bi ustvarjene neverbalne izraze lahko predstavili uporabnikom z realističnimi in človeku podobnimi pogovornimi agenti s telesom. Moderni realizatorji obnašanja namreč lahko podpirajo različne parametre verodostojnosti pogovornega obnašanja, kot je raznolikost in večmodalnost načrtovanja, situacijsko zavedanje, sinteza jezikovne vsebine, sinhronizacija itd. Animacijski ogrodji, kot sta Unity in Panda 3D, pa predstavljajta močno orodje za hitro in visokokakovostno zasnovo in izvedbo virtualnih, človeku podobnih, entitet. Če sklenemo, zmožnost izražanja informacij vizualno in čustveno je pri človeški komunikaciji ključnega pomena. Posledično lahko z opredelitvijo osebnosti in

čustvenega stanja ECA tak agent postane aktiven udeleženec v pogovoru. Toda če želimo, da deluje še bolj naravno, mora biti agent zmožen tekočega in skoraj nemudnega odzivanja na situacijske sprožilce, hkrati pa mora zajemati sinhronizirane jezikovne in neverbalne kanale. Predstavljen model zato predstavlja pomemben korak na poti do bolj naravnih in človeku podobnih odzivov, ki jih ustvarja stroj.

Zahvala

Raziskavo je delno financirala Javna agencija za raziskovalno dejavnost Republike Slovenije v okviru projekta HUMANIPA – Fuzija verbalnih in neverbalnih signalov za naslednjo generacijo inteligentnih komunikacijskih vmesnikov (J2-1737).

Viri in literatura

Allwood, J. (2014). »A framework for studying human multimodal communication«. *Coverbal Synchrony in Human-Machine Interaction, ed. 1*. Boca Raton: CRC Press, str. 17–39.

Allwood, J., Ahlsén, E., Lund, J., in Sundqvist, J. (2005). »Multimodality in own communication management«. V *Proceedings from the Second Nordic Conference on Multimodal Communication*. Göteborg: Göteborg University, str. 1–20.

Birdwhistell, R. L. (2010). *Kinesics and context: Essays on body motion communication*. Pennsylvania: University of Pennsylvania Press.

Carroll, R., Peikola, M., Salmi, H., Varila, M. L., Skaffari, J., in Hiltunen, R. (2013). »Pragmatics on the page: Visual text in late medieval English books«. *European Journal of English Studies*, 17(1), str. 54–71.

Cassell, J., Bickmore, T., Campbell, L., Vilhjalmsson, H., in Yan, H. (2001). »More than just a pretty face: conversational protocols and the affordances of embodiment«. *Knowledge-based systems*, 14(1-2), str. 55–64.

Chui, K., Lee, C. Y., Yeh, K., in Chao, P. C. (2018). »Semantic processing of self-adaptors, emblems, and iconic gestures: An ERP study«. *Journal of Neurolinguistics*, 47, str. 105–122.

Church, R. B., in Goldin-Meadow, S. (2017). »So how does gesture function in speaking, communication, and thinking?«. V »Why Gesture?: How the hands function in speaking, thinking and communicating«, *Gesture Studies 7*, str. 397–412.

Ciechanowski, L., Przegalinska, A., Magnuski, M., in Gloor, P. (2018). »In the shades of the uncanny valley: An experimental study of human–chatbot interaction«. *Future Generation Computer Systems*.

Cooperrider, K. (2017). »Foreground gesture, background gesture«. *Gesture*, 16(2), str. 176–202.

Ekman, P in Friesen, W (1971). »Constants across cultures in the face and emotion«. *Journal of Personality and Social Psychology*, 17(2), str. 124–9.

Esposito, A., Esposito, A. M., in Vogel, C. (2015). »Needs and challenges in human computer interaction for processing social emotional information«. *Pattern Recognition Letters*, 66, str. 41–51.

Guaitella, I., Santi, S., Lagrue, B., in Cavé, C. (2009). »Are eyebrow movements linked to voice variations and turn-taking in dialogue? An experimental investigation«. *Language and speech*, 52(2-3), str. 207–222.

Hoek, J., Zufferey, S., Evers-Vermeul, J., in Sanders, T. J. (2017). »Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study«. *Journal of pragmatics*, 121, str. 113–131.

Kita, S., Van Gijn, I., in Van der Hulst, H. (1997). »Movement phases in signs and co-speech gestures, and their transcription by human coders«. In *International Gesture Workshop*. Springer, Berlin, Heidelberg, str. 23–35.

Kopp S, Bergmann K. (2017). »Using cognitive models to understand multimodal processes: The case for speech and gesture production«. V *The Handbook of Multimodal-Multisensor Interfaces*. New York: Association for Computing Machinery and Morgan in Claypool, str. 239–276.

Kramer, L. L., Ter Stal, S., Mulder, B. C., de Vet, E., in van Velsen, L. (2020). »Developing Embodied Conversational Agents for Coaching People in a Healthy Lifestyle: Scoping Review«. *Journal of medical Internet research*, 22(2), e14058.

Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., ... in Coiera, E. (2018). »Conversational agents in healthcare: a systematic review«. *Journal of the American Medical Informatics Association*, 25(9), str. 1248–1258.

Lopez-Ozieblo, R. (2018). »Can gestures help clarify the meaning of the Spanish marker 'se'?«. *Lingua*, 208, str. 1–18.

Luger E, Sellen A. (2016). »Like having a really bad PA: The gulf between user expectation and experience of conversational agents«. V *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, str. 5286–5297.

Maricchiolo, F., Gnisci, A., in Bonaiuto, M. (2012). »Coding hand gestures: A reliable taxonomy and a multi-media support«. V *Cognitive behavioural systems*. Berlin, Heidelberg: Springer, str. 405–416).

McKeown, G., Sneddon, I., in Curran, W. (2015). »Gender differences in the perceptions of genuine and simulated laughter and amused facial expressions«. *Emotion Review*, 7(1), str. 30–38.

McNeill, D. (2008). *Gesture and thought*. Chicago: University of Chicago Press.

McNeill, D. (2016). *Why We Gesture. 'The Surprising Role of Hand Movements in Communication'*. Cambridge: Cambridge University Press.

McNeill, D., Levy, E., in Duncan, S. D. (2015). »Gesture in Discourse«. V D. Tannen, H. E. Hamilton, D. Schiffrin (urd.), *The Handbook of Discourse Analysis 2*. Oxford: Wiley-Blackwell, str. 262–289.

McTear, M., Callejas, Z., in Griol, D. (2016). *The Conversational Interface: Talking to Smart Devices*. Berlin: Springer International Publishing.

Melinger, A., in Levelt, W. J. (2004). »Gesture and the communicative intention of the speaker«. *Gesture*, 4(2), str. 119–141.

Mlakar I, Kačič Z, Borko M, Rojc M. (2017). »A novel unity-based realizer for the realization of conversational behavior on embodied conversational agents«. *International Journal of Computers*, 2, str. 205–213.

Mlakar I, Kačič Z, Rojc M. (2014). »Describing and animating complex communicative verbal and nonverbal behavior using Eva-framework«. *Applied Artificial Intelligence*, 28(5), str. 470–503.

Mlakar, I., Kačič, Z., Borko, M.., in Rojc, M. (2018). »A novel realizer of conversational behavior for affective and personalized human machine interaction-EVA U-Realizer«. *WSEAS Trans. Environ. Dev*, 14, str. 87–101.

Mlakar, I., Verdonik, D., Majhenič, S., in Rojc, M. (2019). »Towards Pragmatic Understanding of Conversational Intent: A Multimodal Annotation Approach to Multiparty Informal Interaction–The EVA Corpus«. V *International Conference on Statistical Language and Speech Processing*. Cham.: Springer, str. 19–30.

Navarro-Cerdan, J. R., Llobet, R., Arlandis, J., in Perez-Cortes, J. C. (2016). »Composition of Constraint, Hypothesis and Error Models to improve interaction in Human–Machine Interfaces«. *Information Fusion*, 29, str. 1–13.

Ochs, M., Pelachaud, C., in Mckeown, G. (2017). »A User Perception--Based Approach to Create Smiling Embodied Conversational Agents«. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(1), str. 1–33.

Opel, D. S., in Rhodes, J. (2018). »Beyond Student as User: Rhetoric, Multimodality, and User-Centered Design«. *Computers and Composition*.

Peirce, C. S. (1965). *Collected papers of Charles Sanders Peirce (Vol. 5)*. Cambridge: Harvard University Press.

Philip, P., Dupuy, L., Auriacombe, M., Serre, F., de Sevin, E., Sauteraud, A., in Micoulaud-Franchi, J. A. (2020). »Trust and acceptance of a virtual psychiatric interview between embodied conversational agents and outpatients«. *NPJ digital medicine*, 3(1), str. 1–7.

Poria, S., Cambria, E., Bajpai, R., in Hussain, A. (2017). »A review of affective computing: From unimodal analysis to multimodal fusion«. *Information Fusion*, 37, str. 98–125.

Provoost, S., Lau, H. M., Ruwaard, J., in Riper, H. (2017). »Embodied conversational agents in clinical psychology: a scoping review«. *Journal of medical Internet research*, 19(5), e151.

Queirós, A., in da Rocha, N. P. (2018). »Ambient Assisted Living: Systematic Review«. *Usability, Accessibility and Ambient Assisted Living*, str. 13–47.

Queiroz, J., in Aguiar, D. (2015). »CS Peirce and Intersemiotic Translation«. V *International Handbook of Semiotics*. Dordrecht: Springer, str. 201–215.

Rojc M, Kačič Z. (2011). »Gradient-descent based unit-selection optimization algorithm used for corpus-based text-to-speech synthesis«. *Applied Artificial Intelligence*, 25(7), str. 635–668

Rojc, M., Mlakar, I., in Kačič, Z. (2017). »The TTS-driven affective embodied conversational agent EVA, based on a novel conversational-behavior generation algorithm«. *Engineering Applications of Artificial Intelligence*, 57, str. 80–104.

ter Stal, S., Kramer, L. L., Tabak, M., op den Akker, H., in Hermens, H. (2020). »Design features of embodied conversational agents in eHealth: a literature review«. *International Journal of Human-Computer Studies*, 138, 102409.

Trujillo, J. P., Simanova, I., Bekkering, H., in Özyürek, A. (2018). »Communicative intent modulates production and comprehension of actions and gestures: A Kinect study«. *Cognition*, 180, str. 38–51.

TRANSFORMACIJA 'INTELIGENCE ROJA' {SWARM INTELLIGENCE} NA UMETNO INTELIGENCO

URŠKA MARTINC,[1] BORIS ABERŠEK,[1] BOJAN BORSTNER[2]

[1] Univerza v Mariboru, Fakulteta za naravoslovje in matematiko, Maribor, Slovenija
urska.martinc1@um.si, boris.abersek@um.si
[2] Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
bojan.borstner@um.si

Sinopsis V članku želimo raziskati povezavo med 'naravno' inteligenco roja (angl. swarm intelligence) in umetno inteligenco roja, izhajajoč iz premise, da se obstoječa umetna inteligenca vse bolj uči ena od druge, naravne inteligence, in s tem ustvarja nekakšno notranjo 'nevidno' integracijo, ki bi v končni fazi lahko bila zelo podobna inteligenci roja v naravi. Inteligenco roja v osnovi opisujemo kot kolektivno obnašanje nekega samoorganiziranega sistema. Tak sistem je lahko naraven ali umeten. Same raziskave se na področju inteligence roja delijo na: naravno proti umetnemu in znanstveno proti inženirskemu (Dorigo in Birattari 2007). V analizi se bomo osredotočili predvsem na naravno inteligenco roja in transformacijo pogojev, ki veljajo v naravni inteligenci roja na umetno inteligenco. Na podlagi različnih primerov in iz njih izpeljanih teorij ter posplošitev bomo tako v članku analizirali povezavo med socialnimi žuželkami kot primerom naravne inteligence roja in poskušali ustvariti ter razumeti kriterije za ustvarjanje umetne inteligence roja. Preverili bomo, kaj sploh je inteligenca roja ter kakšne so zahteve takšnih sistemov.

Ključne besede:
inteligenca roja,
umetna inteligenca,
samo-organizacija,
stigmergija,
etične dileme

Univerzitetna založba Univerze v Mariboru

DOI https://doi.org/10.18690/um.ff.11.2022.9
ISBN 978-961-286-675-4

TRANSFORMATION OF 'SWARM INTELLIGENCE' ONTO ARTIFICIAL INTELLIGENCE

URŠKA MARTINC,[1] BORIS ABERŠEK,[1] BOJAN BORSTNER[2]

[1] University of Maribor, Faculty of Natural Science and Mathematics, Maribor, Slovenia, urska.martinc1@um.si, boris.abersek@um.si

[2] University of Maribor, Faculty of Arts, Maribor, Slovenia bojan.borstner@um.si

Abstract In this article, we examine the relationship between "natural" swarm intelligence and artificial swarm intelligence, starting from the premise that existing artificial intelligences are learning from each other to an increasing extent, thus creating a sort of internal 'invisible' integration, which might ultimately resemble swarm intelligence in nature. In principle, we characterize swarm intelligence as collective behaviour of some self-organised system. Such a system can be either natural or artificial. In the field of swarm intelligence, research is split into natural vs artificial and scientific vs engineering (Dorigo and Birattari 2007). In this analysis, we will focus predominately on natural swarm intelligence and the transformation of conditions that hold for natural swarm intelligence to artificial intelligence. Based on various examples and theories and generalisations derived from them, we will, in this paper, endeavour to analyse the connection between social insects as an example of natural swarm intelligence and try to understand the criteria for the creation of artificial swarm intelligence. We will examine what exactly is swarm intelligence and what the requirements of such systems are.

Keywords:
swarm intelligence, artificial intelligence, self-organisation, stigmergy, ethical dilemmas

https://doi.org/10.18690/um.ff.11.2022.9 DOI
978-961-286-675-4 ISBN

University of Maribor Press

1 Inteligenca roja v naravnih/bioloških sistemih

Morda je najtežje vprašanje, kako posameznikovo vedenje povezati s kolektivno uspešnostjo? Z drugimi besedami, kako nastane sodelovanje? (angl. *Perhaps the most difficult question is how to connect individual behaviour with collective performance? In other words, how does cooperation arise?*)

Bonabeau, Dorigo in Theraulaz 1999: 6

V naravi poznamo inteligenco roja kot različne primere inteligentnega kolektivnega vedenja, npr.: skupinsko krmljenje pri socialnih žuželkah[1] (npr. pri čebelah mravljah), delitev dela, gradnja gnezda socialnih žuželk itd. (Krink b. d.; Dorigo in Birattari 2007; Bonabeau, Dorigo in Theraulaz 1999). Ko govorimo o socialnih žuželkah, lahko govorimo kot: (i) o socialnih žuželkah, kot so npr. mravlje ali čebele in (ii) o socialnih žuželkah kot sistemu super-organizma, ki ga ustvarjajo te žuželke. Oba primera bomo v nadaljevanju pojasnili. Iz tega torej sledi, da je inteligenca roja »disciplina, ki se ukvarja z naravnimi in umetnimi sistemi, sestavljenimi iz številnih posameznikov, ki se usklajujejo z uporabo decentraliziranega nadzora in samoorganizacije« (Dorigo in Birattari 2007). V naravi najdemo veliko primerov inteligence roja: »Primeri sistemov, ki jih preučuje inteligenca roja, so kolonije mravelj in termitov, jate rib, jate ptic, črede kopenskih živali« (Dorigo in Birattari 2007). Zanimivo je, da poleg naštetih naravnih sistemov lahko uvrstimo v te sisteme tudi produkte človeka: »nekateri človeški predmeti prav tako spadajo v domeno inteligence roja, zlasti nekateri sistemi z več roboti in nekateri računalniški programi, ki so napisani za reševanje težav z optimizacijo in analizo podatkov« (Dorigo in Birattari 2007). Ti sistemi imajo naslednje lastnosti:

> Sestavljajo ga številni posamezniki; posamezniki so sorazmerno homogeni (lahko so identični ali ustrezajo določeni tipologiji); interakcije med posamezniki temeljijo na preprostih vedenjskih pravilih, ki izkoriščajo le lokalne informacije, ki si jih posamezniki izmenjujejo neposredno ali preko okolja (stigmergija); celotno vedenje sistema je posledica interakcij

[1] *Socialne žuželke* so »/.../ katerekoli od številnih vrst žuželk, ki živijo v kolonijah in kažejo tri značilnosti: skupinsko povezovanje, delitev dela in prekrivanje generacij« (Britannica 2018).

> posameznikov med seboj in z njihovim okoljem, torej samoorganiziranega skupinskega vedenja. (Dorigo in Birattari 2007)

Za razumevanje vseh nadaljnjih izpeljav pa moramo na začetku opredeliti nekatere izhodiščne pojme, ki bodo v nadaljevanju predstavljali ustrezno pojmovno ozadje za razlago inteligence roja. Pri tem je najprej treba pojasniti, kaj je inteligenca roja.

Če povzamemo splošno priznane definicije, potem lahko trdimo, da je inteligenca roja poseben sistem, ki uspešno deluje v naravi in ga uporabljajo nekatere živalske vrste, kot so: čebele, mravlje, ptice ter nekatere druge živalske skupine. Zato raziščimo, kako deluje inteligenca roja na primeru čebel kot skupnosti s kolektivnim vedenjem. Zelo hitro lahko ugotovimo, da so za uspešnost v delovanju skupine odgovorni vsi akterji. Če to drži, potem je torej vedenje posameznika tisti dejavnik, ki vodi k uspehu celotnega kolektiva čebel. Ta posebna značilnost se lahko zelo nazorno pokaže v kontekstu delitve dela, kjer so vse dosedanje raziskave pokazale, da imajo v čebelji družini čebele dejansko točno določeno razdelitev dela, ki v končni fazi koristi kolektivu (ČZS b. d.). Samo dejstvo, da socialne žuželke uprimerjajo kot eno svojih najpomembnejših lastnosti, zmožnost, da lahko nekatere svoje najelementarnejše dejavnosti – kot sta iskanje hrane in gradnja gnezd – učinkovito udejanjijo na podlagi razdelitve nalog med posameznimi člani kolektiva, kaže na to, da je njihov pristop k reševanju problemov dovolj prilagodljiv (kot bomo spoznali v nadaljevanju), da ga lahko prenesemo tudi na področje umetne inteligence. Če pomislimo na socialne žuželke (kot so npr. čebele ali mravlje), lahko vidimo, da je tak sistem pri njih prisoten in tudi zelo uspešen. Gre za sistem, pri katerem takšne enote ne potrebujejo t. i. koordinatorja. Za čebele je npr. značilno, da živijo v skupinah, kjer eno čebeljo družino sestavlja matica (mati čebel), lahko tudi več deset tisoč čebel s troti (samci), značilnost za njihovo delo pa je skupinsko ter samoorganizirano delovanje (ČZS b. d.):

> Čebela kot sama je nesposobna živeti samostojno, zato živi v čebelji družini. Ta je organizirana in deluje kot *super organizem*[2]. Družbeni način življenja je zahteval specializacijo vlog v družini, kar je pripeljalo do razlik v zunanji in notranji zgradbi in funkciji posameznih organov in organizma čebele. Tako

[2] Superorganizem je »organizirana družba (kot socialna žuželka), ki deluje kot organska celota« (Merriam-Webster b. d.).

> čebeljo družino sestavljajo ena matica, 60.000 čebel (na višku sezone) in nekaj tisoč trotov. (ČZS b. d.)

V naravi obstajajo torej sistemi, kot je sistem pri socialnih žuželkah, kjer ima vsak posameznik svojo nalogo in z izpolnjevanjem te naloge koristi celotnemu kolektivu. Kot smo lahko videli na zgornjem primeru čebel, so opravila razdeljena učinkovito, tako da se lahko vsakdanje obveznosti (kot je recimo naloga iskanja hrane pri čebelah) rešijo na uspešen način.

Podobno tudi mehanizme pri mravljah lepo opišejo avtorji Garnier, Gautrais in Theraulaz:

> Mravlje komunicirajo med seboj z uporabo feromonov. Ti feromoni so kemične snovi, ki privabljajo druge mravlje. Ko na primer mravlja najde vir hrane, se hitro vrne v gnezdo in položi feromonsko pot, ki bo nato druge delavce vodila od gnezda do vira hrane. Ko se rekrutirane mravlje vrnejo v gnezdo, odložijo svoj feromon na sled in okrepijo pot. Oblikovanje poti je torej rezultat pozitivne povratne informacije: več kot mravlje uporabljajo pot, privlačnejša postaja pot. (Garnier, Gautrais in Theraulaz 2007: 8-9)

Nič nenavadnega torej ni, da človek poskuša posnemati naravno vedenje določenih skupin socialnih žuželk, kajti njihov uspeh (torej uspeh celotne skupine oz. kolektiva), lahko pomeni tudi uspeh za naša področja, če prenesemo naravne mehanizme, ki vodijo k uspehu, tudi na človeška področja.
Dorigo in Birattari opredeljujeta lastnost umetne inteligence roja takole:

> Značilna lastnost sistema inteligence roja je njegova sposobnost, da deluje usklajeno brez prisotnosti koordinatorja ali zunanjega krmilnika. V naravi je pri rojih mogoče opaziti veliko primerov, ko izvajajo neko kolektivno vedenje, ne da bi kdo nadziral skupino ali se zavedal splošnega vedenja skupine. Ne glede na pomanjkanje posameznikov, ki so odgovorni za skupino, lahko roj kot celota kaže inteligentno vedenje. To je rezultat interakcije prostorsko sosednjih osebkov, ki delujejo na podlagi preprostih pravil. (Dorigo in Birattari 2007)

Zato želimo dokazati, da te pogoje samoorganiziranega vedenja lahko prenesemo tudi na umetno inteligenco. Vse te skupine imajo torej samoorganizirano, kolektivno vedenje, ki vodi k uspešnosti skupine. Skupine morajo tako med seboj sodelovati in pomembno vprašanje, ki ga izpostavijo Bonabeau, Dorigo in Theraulaz, je naslednje: »morda je najtežje vprašanje, kako posameznikovo vedenje povezati s kolektivno uspešnostjo? Z drugimi besedami, kako nastane sodelovanje« (Bonabeau, Dorigo in Theraulaz 1999: 6). Avtorji Bonabeau, Dorigo in Theraulaz ponujajo odgovor na to vprašanje, in sicer pravijo, da je lahko gensko pogojeno, vendar so:

> /…/ mnogi vidiki kolektivnih dejavnosti socialnih žuželk so samoorganizirani. Torej po teoriji samoorganizacije (SO) (Haken 1983, Nicolis in Prigogine 1977), prvotno razvite v okviru fizike in kemije, da bi opisale pojav makroskopskih vzorcev iz procesov in interakcij, zaznamovanih na mikroskopski ravni, lahko razširimo na socialne žuželke. Tako lahko pokažemo, da se kompleksno kolektivno vedenje lahko pojavi iz interakcij med posamezniki, ki kažejo preprosto vedenje: v teh primerih se za razlago kompleksnega kolektivnega vedenja ni treba sklicevati na posamezno kompleksnost. (Bonabeau, Dorigo in Theraulaz 1999: 6)

Preverili smo, kateri mehanizmi so del naravne inteligence roja in ugotovili, da sta glavna mehanizma naravne organizacije *samoorganizacija* ter *stigmergija* (ta je definirana kot stimulacija z delom) (Krink b. d.). Zastaviti pa si moramo temeljno vprašanje: ali želimo, da se umetna inteligenca razvije do stopnje popolne avtonomnosti in samoorganiziranosti, do stopnje inteligence roja, katerega osnovno izhodišče je, da posamezna entiteta izpolnjuje naloge, ki so v korist celotnemu kolektivu? In jasno je treba že na začetku definirati, kaj je za UI 'celoten kolektiv':

a. Ali je to kolektiv UI entitet?
b. Ali je to celotna družba, ki vključuje tudi ljudi?

2 Inteligenca roja in umetna inteligenca

Naraven sistem inteligence roja je tako uspešen, da ga ljudje želimo posnemati na različnih področjih (robotika, algoritmi, umetna inteligenca itd.). Ukvarjanje s tem področjem je trenutno aktualno, kar kažejo tudi različne raziskave (npr. Bonabeau in Meyer 2001, Dorigo in Stützle 2004, Di Caro, Ducatelle in Gambardella 2005,

Dorigo in Birattari 2007, Innocente in Grasso 2019) in nekatere od njih bomo tudi ovrednotili. Najprej bomo preverili, kako bi izraz IR (inteligenca roja) lahko prenesli iz bioloških na *nebiološke umetne sisteme*.[3]

Prvič so izraz IR uporabili v okviru nebioloških umetnih sistemov avtorji Beni, Hackwood in Wang (Beni 1988, Beni in Wang 1989, Beni in Hackwood 1992), in sicer:

> /…/ v kontekstu celičnih robotskih sistemov, kjer veliko enostavnih agentov zaseda eno ali dvodimenzionalen prostor, da bi ustvarjali vzorce in se samoorganizirali z interakcijami z najbližjimi sosedi. Uporaba izraza IR za opis samo tega dela se zdi po nepotrebnem omejujoča: zato njegovo definicijo razširimo tako, da vključuje vse poskuse kreiranja algoritmov ali distribuiranih naprav za reševanje problemov, inspiriranih s kolektivnim vedenjem kolonij socialnih žuželk in drugih živalskih kolektivov. (Bonabeau, Dorigo in Theraulaz 1999: 7)

Zanimivo razlago ponujata tudi Dorigo in Birattari (2007), ki trdita, da je inteligenca roja:

/…/ disciplina, ki se ukvarja z naravnimi in umetnimi sistemi, sestavljenimi iz številnih posameznikov, ki se usklajujejo z uporabo decentraliziranega nadzora in samoorganizacije. Disciplina se osredotoča zlasti na kolektivno vedenje, ki je posledica lokalnih interakcij posameznikov med seboj in z okoljem. (Dorigo in Birattari 2007)

Glede na to, da je vedno bolj popularno področje umetne inteligence roja, je treba razjasniti tudi prednosti in potencialne slabosti teh sistemov. Bonabeau in Meyer sta opredelila prednosti inteligence roja kot:

- »prilagodljivost (kolonija se lahko prilagodi spreminjajočemu se okolju);
- robustnost (tudi če eden ali več posameznikov ne uspe, lahko skupina še vedno opravlja svoje naloge); in

[3] *Nebiološki umetni sistemi* bodo v našem primeru označeni kot vsi umetni sistemi, ki jih je ustvaril človek, kar tudi ustreza neki splošni definiciji umetnega (Bird in Tobin 2018).

- samoorganizacija (dejavnosti niso niti centralno niti lokalno nadzorovane).« (Bonabeau in Meyer 2001: 111).

Več o težavah, dilemah in slabostih pa v nadaljevanju.

Skladno s temi spoznanji je potem smiselno ugotavljati, v kakšnem odnosu sta biološka inteligenca roja in to, kar je v ospredju našega zanimanja, torej umetna inteligenca. Če pogledamo različne raziskave, ki se ukvarjajo s tem področjem, potem hitro ugotovimo, da obstajajo pomembne povezave med nekaterimi tipičnimi vzorci kolektivnega delovanja v živalskih skupnostih in tem, kar se dogaja na področju umetne inteligence. V devetdesetih letih so prve raziskave pokazale, da ima razumevanje določenih vzorcev kolektivnega delovanja pri živalih lahko pomemben vpliv na razlage kolektivnega delovanja na različnih drugih področjih. To lahko zelo nazorno pokažemo z ugotovitvami, do katerih so prišli Bonabeau, Dorigo in Theraulaz:

Odkritje, da samoorganizacija lahko deluje pri socialnih žuželkah, ne vpliva le na preučevanje socialnih žuželk, temveč nam ponuja tudi močna orodja za prenos znanja o socialnih žuželkah na področje oblikovanja inteligentnih sistemov. (Bonabeau, Dorigo in Theraulaz 1999: 6)

Sklepamo torej lahko tudi, da se vzorci, ki jih raziskujemo pri živalih in so del njihovega obnašanja, lahko prenesejo tudi na druga področja, kot je npr. področje umetne inteligence. Takšne vrste kolektivnega vedenja, ki ga uporabljajo socialne žuželke, je torej lahko uporabno tudi v primeru nebioloških umetnih inteligentnih sistemov. Kot bomo videli v nadaljevanju, je uporaba takšnega naravnega sistema postala priljubljena za reševanje nekaterih problemov v umetnih inteligentnih sistemih. Seveda je to razumljivo, saj kolektivno zavedanje v naravnih sistemih vodi v uspešnost celotnega kolektiva, prav tako takšni sistemi ne potrebujejo nekega zunanjega upravljalca in hkrati skrbijo za dobrobit celotnega kolektiva (spomnimo se na primer kolektivnega vedenja pri čebelah, kjer ima vsaka čebela svojo nalogo, naloge pa prispevajo k uspehu celotnega kolektiva). Popularnost inteligence roja kot ene od paradigem za strategijo razvoja umetne inteligence izhaja iz spoznanj, da je interakcija, ki se pojavlja med relativno preprostimi dejavniki, dovolj fleksibilna in robustna ter hkrati izjemno učinkovita. Zato ne preseneča, da Bonabeau, Dorigo, Theraulaz (1999: xi) ugotavljajo: »Število njenih uspešnih aplikacij eksponentno

narašča na področju kombinacijske optimizacije, komunikacijskih omrežij in robotike« (Bonabeau, Dorigo in Theraulaz 1999: xi).

3 Simuliranje in IR

Prvi poskusi simuliranja inteligence roja se pojavijo že zgodaj in stremijo k posnemanju naravnih sistemov, ki so uspešni, kar je temelj človekovega razvoja - poskus posnemanja v naravi dobro delujočih sistemov. Tako že obstajajo tudi različni umetni modeli inteligence roja, kot je na primer program Boids,[4] ki ga je zasnoval Craig Reynolds leta 1986 in simulira vedenje jat ptic (Reynolds 1987). Kot lahko vidimo iz tega primera, se je poskus prenosa znanj inteligence roja iz naravnih sistemov na umetne sisteme zgodil že nekaj časa nazaj, od takrat so tehnologije napredovale, tako da je povsem razumljiva težnja, da se dejanski pozitivni dosežki pri umetnih sistemih inteligence roja prenesejo na čim več različnih področij. Avtorji Bonabeau, Dorigo in Theraulaz pravijo takole:

Raziskovalci imajo dobre razloge, da se jim zdi inteligenca roja privlačna: v času, ko svet postaja tako zapleten, da ga nobeno človeško bitje ne more razumeti, ko nam informacije (in ne njihovo pomanjkanje) ogrožajo življenje, ko postanejo programski sistemi tako kompleksni, da jih ni več mogoče nadzorovati, ponuja inteligenca roja alternativni način oblikovanja 'inteligentnih' sistemov, v katerih avtonomija, pojavnost in porazdeljeno delovanje nadomeščajo nadzor, predprogramiranje in centralizacijo. (Bonabeau, Dorigo in Theraulaz 1999: xi)

Vendar pri tem pozabljamo na temeljno izhodišče te premise kompleksnosti, namreč da programski sistemi postajajo tako inteligentni in kompleksni, da jih ne moremo več nadzorovati, zato poskušamo to nesposobnost obvladovanja sistema nadomestiti še z bolj kompleksnim sistemom, to je inteligenco roja. Potreba po neke vrste avtonomni notranji samoorganizaciji se tako kaže za koristno tudi v nebioloških umetnih sistemih.

[4] Glej Wong 2008.

4 Umetni inteligentni nebiološki sistemi – transformacija inteligence roja

Naravne sisteme inteligence roja smo torej predstavili, v nadaljevanju bomo naredili analizo umetnih sistemov in možnosti inteligence roja v umetnih sistemih. V analizi nas predvsem zanima, kako poteka oz. ali je možna transformacija inteligence roja na umetno inteligenco? Ob tem se takoj zastavi vprašanje, kako lahko nastane takšno kolektivno vedenje? Poglejmo si, kaj je sploh značilnost takšnih sistemov inteligence roja. Za takšne sisteme je značilna multidisciplinarnost, kot trdita Dorigo in Birattari: »Inteligenca roja ima izrazit multidisciplinaren značaj, saj je sisteme /…/ mogoče opaziti na različnih področjih. Raziskave v inteligenci roja lahko razvrstimo po različnih kriterijih« (Dorigo in Birattari 2007). Kot smo že omenili v uvodu, raziskave na področju IR Dorigo in Birattari problematizirata na dveh področjih:

- področje naravno (biološko) proti umetnemu (nebiološkemu) in
- področje znanstveno proti inženirskemu (Dorigo in Birattari 2007).

Če pogledamo prvo razdelitev, lahko ugotovimo, da je naravno nekaj, kar izhaja iz narave, umetno pa je produkt človeka. Takole razliko med naravnim in umetnim opredelita Bird in Tobin: »če rečemo, da je neka vrsta naravna, to pomeni, da ustreza skupini, ki odraža strukturo naravnega sveta in ne interese in dejanja ljudi« (Bird in Tobin 2018). Avtorja Dorigo in Birattari razložita zgoraj omenjene kategorije naravnega in umetnega:

Običajno je, da se raziskovanje inteligence roja razdeli na dve področji glede na naravo analiziranih sistemov. Zato govorimo o raziskavah naravne inteligence roja, kjer se preučujejo biološki sistemi; in umetne inteligence roja, kjer se preučujejo človeški artefakti. (Dorigo in Birattari 2007)

V primerih umetne inteligence roja tako poskušamo imitirati naravno vedenje (ki ga je ustvarila narava) in ga prenesti na določeno področje človekovega ustvarjanja (npr. področje umetne inteligence). Ko govorimo o drugem področju, o znanstvenem in inženirskem, pa avtorja Dorigo in Birattari takole definirata področji:

Na podlagi zastavljenih ciljev lahko podamo alternativno in nekako bolj informativno klasifikacijo raziskav inteligence roja: prepoznamo lahko znanstveni in inženirski tok. Cilj znanstvenega toka je modelirati inteligentne sisteme rojev ter izločiti in razumeti mehanizme, ki sistemu v celoti omogočajo usklajeno vedenje kot rezultat lokalnih interakcij med posamezniki ter med posameznikom in okoljem. Po drugi strani pa je cilj inženirskega toka izkoristiti razumevanje, ki ga je razvil znanstveni tok, da bi oblikovali sisteme, ki lahko rešujejo probleme, ki so praktično pomembni. (Dorigo in Birattari 2007)

Preverili smo torej razlike med različnimi kategorijami in na tem mestu si poglejmo nekaj znanih primerov raziskav inteligence roja, ki jih navajata Dorigo in Birattari (2007):

- določena (prehranjevalna) vedenja mravelj: pri čemer gre za primer naravno-znanstvenega sistema;
- gnezdilno vedenje os in mravelj; pri čemer gre prav tako za primer naravno-znanstvenega sistema;
- grozdenje roja robotov: pri čemer gre za umetno/znanstveni sistem;
- uporaba/izkoriščanje kolektivnega vedenja živalskih skupnosti: pri čemer gre za naravno/inženirski sistem;
- analiza podatkov s pomočjo roja: pri čemer gre za umetni/inženirski sistem;
- učenje in tvorjenje jat pri pticah in ribah: tukaj gre za primere naravno-znanstvenih in umetno-inženirskih sistemov, ko govorimo o simulacijskih programih, kot je npr. Boidova simulacija. Dorigo in Birattari (2007).

Dorigo in Birrattari sta v nadaljevanju izpostavila, da je vsak posameznik odgovoren pri odločitvah za svoje delovanje, ki temeljijo na spodnjih podatkih:

> »Znanstveniki so pokazali, da je to elegantno vedenje na ravni roja mogoče razumeti kot rezultat samoorganiziranega procesa, kjer noben vodja ni odgovoren in vsak posameznik svoje odločitve o gibanju temelji izključno na lokalno razpoložljivih informacijah: razdalji, zaznani hitrosti in smeri gibanje sosedov.« (Dorigo in Birattari 2007).

Dorigo in Birattari sta dosedanja spoznanja lepo ponazorila s pomočjo različnih kombinacij, kot je prikazanov spodnji tabeli:

Tabela 1

	Naravno (primeri)	**Umetno (primeri)**
Znanstveno	– vedenje čebel pri sporazumevanju (t. i. čebelji ples)	– grozdenja rojev robotov
Inženirsko	– nadzorovana uporaba kolektivnega vedenja živalskih družb – posnemanje samoorganiziranja pri socialnih žuželkah	– analiza podatkov s pomočjo roja

(vir: Dorigo in Birattari 2007)

Če povzamemo te ugotovitve, potem lahko na podlagi njunih analiz izluščimo naslednje pomembne značilnosti sistemov inteligence roja:

(i) Sestavljajo ga številni posamezniki.
(ii) Posamezniki so sorazmerno homogeni.
(iii) Interakcije med posamezniki temeljijo na preprostih vedenjskih pravilih.
(iv) Vedenjska pravila temeljijo na lokalnih informacijah.
(v) Lokalne informacije si posamezniki izmenjujejo:
  (a) neposredno
  (b) preko okolja
(vi) Skupinsko vedenje sistema je posledica interakcij posameznikov:
  (a) med seboj
  (b) z njihovim okoljem
(vii) Ne obstajajo nikakršni zunanji nadzorni/usklajevalni mehanizmi.
(viii) Skupinsko vedenje sistema je torej na podlagi interakcij, lokalnih informacij in enostavnih vedenjskih pravil samoorganizirano od znotraj.
(ix) Stigmergija.
(x) Tako samoorganizirano skupinsko vedenje, ki je predstavljeno kot dovolj robusten in gibek kompleksni sistem, omogoča uspešno razreševanje vsakdanjih problemov.

Tako smo, po predlogih analiziranih avtorjev, podali glavne značilnosti naravnih samoorganiziranih kompleksnih sistemov, ki uporabljajo inteligenco roja in so uspešni v reševanju vsakdanjih problemov. V naslednjem koraku pa nas zanima možnost izgrajevanja uspešnega kompleksnega umetnega sistema, ki bo grajen na značilnostih, ki smo jih ugotovili pri naravnih sistemih.

5 Umetni sistemi – inženirsko – aplikacije in uporaba

Vprašanje, na katerega bomo poskušali v nadaljevanju odgovoriti, in ga postavlja tudi Krink, je: Ali je inteligenca roja uporabna v umetnih sistemih? Ali je uporabna tudi na področju umetne inteligence? (Krink b. d.). Na eni strani gre za vprašanje o upravičenosti vzpostavljanja analogij med biologijo rojev in informacijskimi sistemi na podlagi razumevanja teh podobnosti (in hkratnemu prepoznavanju možnih razlik, ki se lahko pojavljajo kot ovire za to), kar naj potem omogoča računalniško modeliranje naravnih rojev, na drugi strani pa za vprašanje, kako bi lahko te strukturne značilnosti prenesli v modele, ki bi bili tako uporabni in učinkoviti, kot so na ravni naravnih rojev (Krink b. d.).

Izhodišče za vzpostavljanje analogij predstavljajo določene značilnosti vedenja posameznih živalskih organizmov, ki živijo v posebnih organiziranih skupnostih, kolonijah. Če vzamemo kot primer mravlje, potem vidimo, da je njihovo skupinsko vedenje, ki omogoča uspešno razreševanje osnovnih problemov kolektiva, utemeljeno na sodelovanju in primerni delitvi dela, kjer je dodeljevanje nalog zelo prilagodljivo. Vedenje mravelj je v nekaterih pogledih podobno vedenju čebel: pri obeh skupinah vidimo pomembnost deljenja dela (čebele imajo delitev dela, kot je npr. razdelitev na čebele krmilke, ki so zadolžene za krmljenje matice; pri mravljah lahko prav tako opazimo strukturirano delo).

Dejstvo je, da sta obe skupini žuželk na podlagi samoorganiziranega delovanja zelo uspešni pri razreševanju svojih problemov, zato ne preseneča, da so snovalci umetnih sistemov želeli spoznanja o njihovem vedenju in samoorganizaciji prenesti tudi na umetne sisteme.

Izraz inteligenca roja, ki se torej uporablja predvsem na področju računalništva, je dejansko izraz za superorganizem v biologiji. Na kratko preverimo bistvene značilnosti superorganizma, s pomočjo katerih lahko prepoznamo podobnosti obeh izrazov.

Kot pravi Tautz, je izraz superorganizem predstavil Wheeler:

Ta pogled, ki celotno čebeljo družino enači z eno živaljo, je pripeljal do pojma 'bien', ki pomeni 'organsko pojmovanje individuuma'. Čebeljo družino obravnava kot nedeljivo celoto, kot en sam živ organizem. Za takšno življenjsko obliko je ameriški biolog William Morton Wheeler (1865-1937) leta 1911 na osnovi raziskovanja mravelj skoval izraz superorganizem. (izvor besede: lat. *super* = nad,; gr. *organon* = orodje) (Tautz 2010: 3)

Se je pa z izrazom superorganizem ukvarjal tudi James Lovelock, avtor, ki je razvil Gaia hipotezo (več o tem v Lovelock 1972). Naslednji pomemben korak v iskanju odgovora na izvorno vprašanje o statusu ekoloških superorganizmov predstavlja pristop »/…/ biologa Williama Hamiltona v sedemdesetih letih 20. stoletja, ki je na računalnikih začel modelirati ekosisteme. Ugotovil je, da se je v teh modelih (pa tudi v resničnem življenju) zelo malo sistemov lahko samoorganiziralo v kakršenkoli trajno skladnosten sistem« (Kelly 1994: 89).

V predhodnih analizah smo ugotovili, kako se povezujeta oba temeljna pojma, sedaj pa nadaljujemo z analizo transformacije naravne inteligence roja na umetno inteligenco roja. Najprej nas zanima, če se pogoji, ki veljajo v naravni inteligenci roja, lahko prenesejo na umetno inteligenco. Pri tem je zanimivo, da so inženirji osnovne ideje inteligence roja najprej implementirali na robote. Pri posnemanju takšnega naravnega sistema strokovnjaki poskušajo oblikovati umetno kolektivno vedenje (Bouffanais 2016). Radhika razlaga, da so ustvarili umetne sisteme inteligence roja, pri čemer so naredili robota imenovanega 'Kilobot', ki je bil podlaga za »proizvodnjo roja 1024 robotov, s katerimi testirajo kolektivno vedenje« (Radhika 2016).

Ob tem pa obstajajo tudi različni umetni modeli inteligence roja, kot je na primer program Boids, ki ga je zasnoval Craig Reynolds leta 1986 in simulira vedenje jat ptic (Reynolds 1987). To je bil prvi računalniški simulirani model inteligence roja:

Cilj simulacije je bil ponoviti vedenje jat ptic. Namesto da bi nadzirali interakcije celotne jate, simulacija Boids določa le vedenje vsake posamezne ptice. Z le nekaj preprostimi pravili uspe programu ustvariti rezultat, ki je dovolj zapleten in realističen, da ga lahko uporabimo kot ogrodje za računalniške grafične aplikacije, kot je računalniško ustvarjena vedenjska animacija v filmih. (Wong 2008)

Zelo hitro pa se je izkazalo, da bi se dalo te ideje uporabiti tudi v analizah skupinskega vedenja ljudi (npr. Bonabeau in Meyer 2001; Krause, Ruxton in Krause 2010; Radhika 2016), kar vodi do vzpostavljanja paradigme o superorganizmih na različnih področjih človekovega družbenega delovanja.

Glede na to, da je naravni sistem inteligence roja tako uspešen, je logična posledica, da želimo prenesti te mehanizme inteligence roja tudi na skupinsko vedenje ljudi. To naj bi pripomoglo k večji uspešnosti skupine, pa najsi gre za primere, kot so npr. podjetja, delovne skupine ali pa športne ekipe. V principu je prenos mehanizmov isti, kot je to pri prenosu iz naravnega sistema na umetni sistem umetne inteligence. Pa vendar je treba izpostaviti, da verjetno ne moremo neposredno prenašati vseh mehanizmov naravnega sistema na področju človeškega vedenja in drugih področij. Avtorja Bonabeau in Meyer sta se lotila preiskovanja možnega prenosa naravne inteligence roja na področje ekonomije in pravita takole: »V zadnjih 20 letih smo mi in drugi raziskovalci razvili stroge matematične modele za opis vedenja socialnih živali, in zdaj prenašamo te tehnike na področje poslovnih problemov (Bonabeau in Meyer 2001: 108).

Na tem primeru lahko vidimo, da potekajo intenzivni prenosi mehanizmov naravne inteligence roja na različna področja človekovega delovanja (npr. na področje umetne inteligence, na področje ekonomije in na druge interakcije v družbi itd.), pri čemer bi ta področja postala uspešnejša z upoštevanjem mehanizmov naravne inteligence roja. Avtorji Garnier, Gautrais in Theraulaz pa zaključijo takole:

Nadaljevanje eksperimentalnih raziskav v bioloških sistemih in razvoj novih teoretičnih okvirov o prilagoditveni vlogi teh modulacij bi moral spodbuditi nastanek novih uporabnih študij. To nam daje verjeti, da potencial inteligence rojev še zdaleč ni izčrpan (Garnier, Gautrais in Theraulaz 2007: 21).

Kot smo že omenili, avtorja Bonabeau in Meyer trdita, da so prednosti inteligence roja: prilagodljivost, robustnost in samoorganizacija (Bonabeau in Meyer 2001). Vprašanje, ki se poraja, je, do katere mere lahko ta sistem prenesemo na človekovo skupinsko vedenje: na prvi pogled bi lahko rekli, da se problem odpre že pri robustnosti, za katero Bonabeau in Meyer pravita takole: »/…/ tudi če eden ali več posameznikov ne uspe, lahko kolektiv še vedno opravlja svojo nalogo« (Bonabeau in Meyer 2001: 111).

Ta prenos naravnega mehanizma robustnosti na področje človekovega delovanja in vedenja je problematičen, in sicer iz vidika upoštevanja ter prepoznavanja vsakega posameznika kot pomembnega za uspeh skupine, pri čemer je posameznik nepogrešljiv del tega kolektivnega delovanja. Za natančen vpogled v to področje pa potrebujemo še dodatne obširnejše raziskave in analize delovanja teh mehanizmov. Mehanizmi, ki so prisotni v naravi, so torej lahko (ali pa tudi ne) rešitev tudi za nekatera človeška področja ter probleme, saj se zadnje čase vse bolj zavedamo tudi potrebe o etični presoji, npr. pri uporabi avtonomnih sistemov (npr. Bench-Capon 2020, Burton, Habli, Lawton, McDermid, Morgan in Porter 2019).
Ko govorimo o umetni inteligenci, govorimo o nečem, kar ni ustvarila narava skladno z njenimi zakoni, in kar ustreza definiciji naravnega (ki smo jo spoznali v prejšnjih poglavjih), ampak o tem, kar je ustvaril človek v nekaj desetletjih. Običajno pa so osnovne trditve o tem razvoju hitre in tudi premalo domišljene (Aberšek, Borstner in Bregant 2014).

V prihodnosti bomo morali odgovoriti na kar nekaj vprašanj, kajti problemi in dileme, ki smo jih predstavili v pričujočem prispevku, so vedno bolj aktualna, tehnologija pa se razvija z vedno večjo hitrostjo. To pa pomeni, da moramo posvetiti še več časa takšnim razpravam in poskusiti poiskati čim boljše rešitve ter odgovore na ta vprašanja.

7 Zaključek

Naravni sistem inteligence roja je tako uspešen, da ga ljudje želijo posnemati na različnih področjih (robotika, algoritmi, umetna inteligenca itd.). V tem prispevku smo želeli pokazati, kako delujejo mehanizmi naravne inteligence roja. To smo preverili s pomočjo različnih avtorjev. Ko smo postavili okvirje za te mehanizme, smo poskušali pokazati, ali se lahko ti mehanizmi naravne inteligence roja prenesejo

na umetno inteligenco roja. Prikazali smo nekaj primerov, s katerimi smo podkrepili možnost prenosa naravnih mehanizmov na umetne.

Zavedamo se, da se z napredkom tehnologije možnosti prenosa teh naravnih mehanizmov izboljšujejo in povečujejo, zato bodo za področje prenosa naravnih mehanizmov na umetno inteligenco potrebne še obširne dodatne raziskave tako na čisto tehnološkem kot tudi na temeljnem filozofskem, etičnem in moralnem področju.

Viri in literatura

Aberšek, B., Borstner, B. in Bregant, J. (2014). *Virtual teacher: cognitive approach to e-learning material*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Bench-Capon, T. J. M (2020). »Ethical approaches and autonomous systems«. *Artif. Intell.*, 103239. https://doi.org/10.1016/j.artint.2020.103239.

Beni, G. (1988). »The Concept of Cellular Robotic System«. V *Proceedings 1988IEEE Int. Symp. on Intelligent Control*. Los Alamitos, CA: IEEEComputer Society Press, str. 57–62.

Beni, G. in Wang, J. (1989). »Swarm Intelligence«. V *Proceedings Seventh Annual Meeting of the Robotics Society of Japan*. Tokyo: RSJ Press, str. 425–428.

Beni, G. in Hackwood, S. (1992). »Stationary Waves in Cyclic Swarms«. V *Proceedings 1992 IEEE Int. Symp. on Intelligent Control*. Los Alamitos, CA: IEEE Computer Society Press, str. 234–242.

Bird, A. in Tobin, E. (2018). »Natural Kinds«. V E. N. Zalta, *The Stanford Encyclopedia of Philosophy* (izdaja pomlad 2018). URL = <https://plato.stanford.edu/archives/spr2018/entries/natural-kinds/>.

Bonabeau, E. in Meyer, C. (2001). »Swarm intelligence. A whole new way to think about business«. *Harv Bus Rev*, 79(5), str. 106–14.

Bonabeau, E., Dorigo, M. in Theraulaz, G. (1999). *Swarm Intelligence: From Natural to Artificial Systems*. Oxford: Oxford University Press.

Bouffanais, R. (2016). *Design and control of swarm dynamics*. Singapore: Springer.

Britannica, The Editors of Encyclopaedia. (2018). »Social insect«. *Encyclopedia Britannica* (20. april 2021). URL = https://www.britannica.com/animal/social-insect. Burton, S., Habli, I., Lawton, T., McDermid, J., Morgan, P. in Porter, Z. (2019). »Mind the Gaps: Assuring the Safety of Autonomous Systems from an Engineering, Ethical, and Legal Perspective«. *Artificial Intelligence*, 279, [103201]. https://doi.org/10.1016/j.artint.2019.103201.

Čebelarska zveza Slovenije (b. d.). »Čebelja družina«. *Čebelarska zveza Slovenije* (*20*. april 2021). URL = https://www.czs.si/content/C12.

Di Caro, G., Ducatelle, F. in Gambardella, L. M. (2005). »AntHocNet: An adaptive nature-inspired algorithm for routing in mobile ad hoc networks«. *European Transactions on Telecommunications,* 16(5), str. 443–455.

Dorigo, M. in Birattari, M (2007). »Swarm intelligence«. *Scholarpedia*, 2(9), 1462 (2. marec 2021). URL = http://www.scholarpedia.org/article/Swarm_intelligence.

Dorigo, M. in Stützle, T. (2004). *Ant Colony Optimization*. Cambridge, MA: MIT Press.

Haken, H. (1983). *Synergetics*. Berlin: Springer-Verlag.

Garnier, S., Gautrais, J. in Theraulaz, G. (2007). The biological principles of swarm intelligence«. *Swarm Intell*, 1, str. 3–31.

Innocente, M. in Grasso, P. (2019). »Self-organising swarms of firefighting drones: Harnessing the power of collective intelligence indecentralised multi-robot systems«. *Journal of Computational Science*, 34, str. 80–101.

Kelly, K. (1994). *Out of control: the new biology of machines, social systems and the economic world*. Boston: Addison-Wesley.

Krause, J., Ruxton, G. D. in Krause, S. (2010). »Swarm intelligence in animals and humans«. *Trends Ecol Evol.*, 25(1), str. 28–34. 10.1016/j.tree.2009.06.016.

Krink, T. (b. d.). »Swarm Intelligence – Introduction«. *University of Washington Faculty Web Server* (3. februar 2021). URL = http://faculty.washington.edu/paymana/swarm/krink_01.pdf.

Lovelock, J. E. (1972). »Gaia as seen through the atmosphere«. *Atmospheric Environment*, 6(8), str. 579–580.

Merriam-Webster (b. d.). »Superorganism«. *Merriam-Webster, Incorporated* (30. februar 2021). URL = https://www.merriam-webster.com/dictionary/superorganism.

Nicolis, G. in Prigogine, I. (1977). *Self-Organization in Non-Equilibrium Systems*. New York: Wiley & Sons.

Radhika, N. (2016). »Taming the swarm - Collective Artificial Intelligence«. *Youtube, TEDxBermuda*. (20. marec 2021). URL = https://www.youtube.com/watch?v=LHgVR0lzFJc&t=638s.

Reynolds, C. W. (1987). »Flocks, herds and schools: A distributed behavioral model«. *SIGGRAPH Comput. Graph.*, 21(4), str. 25–34. https://doi.org/10.1145/37402.3740634.

Tautz, J. (2010). *Čudežni svet čebel*. Ljubljana: Tehniška založba Slovenije, d. d.

Wong, T. 2008. »Boids«. *Stanford University* (20. marec 2021). URL = https://cs.stanford.edu/people/eroberts/courses/soco/projects/2008-09/modeling-natural-systems/boids.html.

UMETNA INTELIGENCA IN EKSISTENČNO TVEGANJE

ALEN LIPUŠ
Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
alen.lipus@pm.me

Sinopsis Kadar razmišljamo o možnosti umetne (super)inteligence, je treba v kontekstu problema nadzora vztrajati, da se umetni inteligenci dodeli učinkovit normativni okvir, da bi s tem povečali njeno usklajenost z najvišjimi družbenimi vrednotami. Pri takšni vrednostno občutljivi zasnovi umetne inteligence pa trčimo v problem spoznavne zaprtosti, ki otežuje uspešno implementacijo družbenih vrednot vanjo. V članku bomo (med drugim) raziskali naslednja vprašanja: Ali bodo racionalni stroji zagotovo presegli ljudi? Kakšen vpogled v delovanje lastnega uma lahko z njimi pridobimo? Kakšen je njihov moralni status? Jih lahko razumemo kot post-osebe? Ali je naša dolžnost, da ustvarimo superinteligentne sisteme? Kakorkoli, ne glede na to, koliko bi umetna inteligenca bila podobna ljudem, velja, da dokler v razlagi človeškega uma ne bo napredka, spoznavna vrzel med nami in umetno inteligenco ne bo nič manjša. Zdi se, da bolj kot smo v vsakdanjem življenju odvisni od tehnologije, bolj zapletena in nerazumljiva je za naše vsakdanje misli. Turingov test zato postane zastarel, saj predpostavlja delovanje inteligentnega uma, ki je podoben našemu. Toda inteligentnost lahko zavzame nam popolnoma tuje spoznavne oblike in nam je kot taka lahko popolnoma nedostopna in nerazumljiva; to pa je še dodaten razlog, zakaj je treba vztrajati pri vrednostni zasnovi inteligentnih sistemov.

Ključne besede:
post-osebe,
spoznavni dostop,
tveganje,
problem usklajenosti,
problem nadzora

Univerzitetna založba Univerze v Mariboru

DOI https://doi.org/10.18690/um.ff.11.2022.10
ISBN 978-961-286-675-4

ARTIFICIAL INTELLIGENCE AND EXISTENTIAL RISK

ALEN LIPUŠ
University of Maribor, Faculty of Arts, Maribor, Slovenia
alen.lipus@pm.me

Abstract When we think about the possibility of artificial (super)intelligence, we must insist, in the light of the control problem, that a normative framework integrated into artificial intelligence complies with the social values. But in such value-sensitive designed artificial intelligence, we are confronted with the problem of epistemic closure, which worsens the successful implementation of social values. In the article we deal with the following questions: Will rational machines surpass humans? What kind of insight into our mind can we gain through them? What is their moral status? Can they be understood as post-persons? Regardless of how similar artificial intelligence could be to humans, it holds that, if there is no breakthrough in the explanation of the human mind, the explanatory gap between us and artificial intelligence will not be any smaller. It seems that the more we depend on technology in everyday life, the more incomprehensible it is to us. The Turing test thus becomes obsolete, as it presupposes the workings of an intelligent mind that is like ours. But intelligence can take foreign cognitive forms, making it completely inaccessible to us, which represents an additional reason to insist on a value-sensitive design of intelligent systems.

Keywords:
post-persons,
epistemic access,
risk,
alignment
problem,
control problem

https://doi.org/10.18690/um.ff.11.2022.10 DOI
978-961-286-675-4 ISBN
University of Maribor Press

1 Zakaj bi stroji morali misliti?

Na vprašanje »Ali stroji lahko mislijo?« odgovarja Alan Turing (1912–1942), angleški matematik in filozof, s specifičnim diagnostičnim testom, ki pa za posledico sicer nima neposrednega odgovora na to vprašanje. Vprašanje namreč vsebuje problem, kako prepoznati mišljenje pri stroju: »Ali ne bi mogli stroji izvajati česa takega, kar bi morali opisati kot mišljenje, pa je zelo različno od tega, kar počne človek?« (Turing 1990: 63)

Ta problem je seveda zelo močan, vendar zgolj takrat, če se dopustimo zavesti prvotnemu vprašanju. Namesto »Ali stroji mislijo?« se vprašamo »Ali ostali ljudje mislijo?« in dobili bomo odgovor, kako soditi stroje. Na podlagi česa smo prepričani, da drugi ljudje mislijo? Na podlagi opazovanja in njihovega obnašanja.[1]

Podobno naj po Turingu velja tudi za stroje in tako se prvotno vprašanje spremeni v pogojnik: »Če stroj opravi Turingov test, potem mu moramo pripisati enake mentalne atribute kot ljudem«. Turingov test je tako preprosta stvar. V samem bistvu gre za test presoje človeškega sodnika med dvema kandidatoma. Sodnik ne ve, kateri kandidat je človek in kateri ni. Lahko si zamislimo, da je med njima pregrada. Sodnik nato izprašuje oba kandidata s preprostimi vprašanji. Če recimo zahteva izračun 3445 množeno s 7889, potem je smiselno, da bo stroj odgovoril s človeško zamudo. Če stroj zadovoljivo opravi test, potem nimamo nobenih razlogov, da mu ne bi pripisali določenih mentalnih lastnosti – kot jih pripisujemo sebi. Izmed bolj izpostavljenih kritikov Turingovega testa je John Searl z argumentom, ki temelji na miselnem eksperimentu 'Kitajska soba' (Markič 2021: 204).

V besedilu bomo naprej predstavili znano kritiko Turingovega testa, tj. miselni eksperiment »Kitajska soba«. Nato bomo podali kritiko nekaterih predpostavk, na katerih je razprava o mislečih strojih osnovana in ki omogočajo smiselno postavitev vprašanja *»Ali stroji lahko mislijo?«*. Čeprav se je Turing že sam zavedal, da je epistemska situacija strojev lahko radikalno drugačna od naše, se kljub temu ta vidik strojne kognicije zanemarja. Spoznavni (ne)dostop do strojne kognicije tako predstavlja še večji problem, sploh ob predpostavki, da inteligentnost pri strojih ne zavzema enake forme kot pri ljudeh. Iz te kritike se osredotočimo na eksistenčna in praktična vprašanja o umetni inteligenci. Obravnavamo problem nadzora, argument

[1] Glej tudi odgovor z »drugimi duhovi« (Searle 1990: 373).

pogube in oboje postavimo v širši kontekst hipoteze o ranljivem svetu. Na podlagi tega predstavimo vprašanje o možnosti tega, da je umetna inteligenca že prisotna oz. vsaj blizu. Podporo tega vprašanja razvijamo na primeru GPT jezikovnih modelov in pojmov nadgradljivosti ter transformativne umetne inteligence. Na koncu obravnavamo razširjeno stališče, da je razprava o eksistenčni ogroženosti neutemeljena in predstavimo primerjavo sklepanja med skeptičnimi teisti, ki zanikajo problem zla, in teoretiki pogube, ki vidijo v umetni inteligenci eksistenčno nevarnost. Čeprav je sklepanje obojih podobno, podamo razloge, da so teoretiki pogube vseeno bolj utemeljeni v svojem sklepanju.

2 Kitajska soba

John Searle (1932), ameriški filozof, si je Kitajsko sobo zamislil kot argument proti podvigom raziskovanja umetne inteligence[2] v poznih sedemdesetih. V mislih je imel delo Rogerja Shanka (Shank in Abelson, 1977). Shank in njegovi sodelavci so ustvarili program, ki lahko v najslabšem primeru simulira človeško zmožnost razumevanja pripovedi (Searle 1990: 362). To pomeni, da lahko podajo zadovoljiv odgovor glede neke zgodbe, v kateri podatek v odgovoru ni zajet. Če vam ponudimo hiter zaslužek za sodelovanje v poslu, ki sumljivo spominja na piramidno shemo, boste ponudbo zavrnili. Podobno je odgovarjal Shankov program, kar je njegove stvarnike napeljalo k misli, da dejansko razume pripoved in ne simulira zgolj človeškega razumevanja. Ta in tej podobne raziskave so Searla napeljale k ugovoru.

Kitajska soba je miselni eksperiment, ki poskuša zavreči glavno tezo tedanje kognitivne znanosti, ki definira mišljenje kot računsko obdelavo formalno določenih elementov (Searle 1990: 363).[3] Posledica te definicije je seveda prepričanje, da je možno z ustreznim programom realizirati močno UI.[4] Pri Kitajski sobi gre za podoben scenarij kot pri Turingovem testu, le da tu ne nastopajo samo stroji. Imamo

[2] V nadaljevanju UI. Razlikovati je potrebno pred različnimi opredelitvami umetne inteligence: umetna splošna inteligenca (USI; angl. *Artificial General Inteligence*) je opredeljena kot inteligenca (hipotetičnega) stroja, ki bi lahko uspešno opravljal katero koli intelektualno nalogo, ki jo lahko človeško bitje. Superinteligenca (SI) je USI, ampak onkraj človeških sposobnosti. Bostrom jo opredeljuje kot »*vsak razum, ki močno presega kognitivne sposobnosti ljudi na skoraj vseh področjih, ki nas zanimajo*« (Bostrom 2014: 26).

[3] Lahko jo razumemo kot kritiko simbolne umetne inteligence, tj. pristopa, ki mišljenje razume kot računanje s simboli. Tako kot program v računalniku realizira operacije z računskimi procesi, tako naj b bila ena izmed operacij, ki jo realizira možganski program (oz. nevralna koda), mišljenje (Markič 2021: 204).

[4] Močna UI predpostavlja, da je možno ustvariti takšno UI, kjer bodo prisotna mentalna stanja, kot so razumevanje, spoznavanje, zavedanje in zavest. Hkrati močna UI predpostavlja, da je možno razložiti človeško pojavnost brez ukvarjanja z možgani, saj človeško kognicijo pojmuje kot računsko manipulacijo formalnih simbolov. Šibka UI predstavlja zgolj močno orodje pri analizi in preverjanju podatkov (Searle 1990: 361).

ograjeno sobo in na eni strani mimoidoče kitajsko govoreče posameznike, na drugi strani pa posameznico, ki ne zna govoriti in tudi ne razume kitajščine. Znotraj sobe so hkrati navodila v njenem maternem jeziku, kako manipulirati z določenimi kitajskimi pismenkami, da bodo tvorile smiseln odgovor na zastavljeno vprašanje. Posameznica sledi slovenskim navodilom za razporejanje kitajskih simbolov, medtem ko računalnik sledi programu. Tako kot računalnik tudi naša posameznica ustvari vtis razumevanja, toda očitno je, da razumevanja kot takšnega – ni. Isto velja za računalnike in programe. Primer Kitajske sobe s tem pokaže na pomanjkljivost Turingovega testa. Iz napisanega lahko rekonstruiramo argument, ki se nanaša zgolj na ta miselni primer.

1. Če je močna UI resnična, potem obstaja program kitajščine, ki omogoča nekemu sistemu (človek, računalnik), ki zažene program, da razume kitajščino.
2. Jaz bi lahko sledil programu kitajščine brez razumevanja jezika.
3. ∴ Torej je močna UI neresnična. (1, 2 MT)

Miselni eksperiment je podpora drugi premisi in sklep govori v prid temu, da razumevanje ne more biti posledica zagona programa oziroma manipulacije s formalnimi simboli. Če namreč računalniški programi sledijo sintaktičnim navodilom, potem nimajo semantične vsebine. Toda: človeška mentalnost ima mentalno vsebino in je kot taka semantična, zato primer s Kitajsko sobo pokaže, da sintaksa ni zadosten pogoj za semantično vsebino, kar ima za svoj sklep to, da mentalnost ne more biti posledica simbolnega računanja oziroma programov.

3 Kritika vprašanja o mislečih strojih

Toda celotno diskusijo, v katero spada problem Kitajske sobe, lahko upravičeno problematiziramo. Vprašanja, ali stroji lahko mislijo; ali je močna UI možna oziroma ali je možna kakšna oblika USI; ali lahko razvijemo UI, ki bo imela vsaj nekatere mentalne vsebine, kot jih imamo mi – ki bo torej mislila, razumela, čutila, so popolnoma odveč. Prvič, človeški možgani so sintaktični stroj, ki pa so vseeno – čeprav še ne vemo, kako – baza za zavest. Drugič, tudi za druge ljudi nismo zagotovo

prepričani, da niso zgolj filozofski zombiji.[5] Podobno kot pri ljudeh, pri katerih sklepamo na notranje življenje iz vedenja, lahko to storimo pri strojih. To, da imamo odpor pri pripisovanju mentalnih vsebin strojem, je prej dokaz za naše nevrobiološke predsodke kot pa podpora dokaza opisanemu dejstvu. In tretjič, prisotnost mentalnih atributov pri strojih je metafizično vprašanje, ki pa ne sme zamegliti aktualnejšega, praktičnega vprašanja o tem, ali lahko dejansko ustvarimo stroj, ki Turingov test dejansko prestane. Postavljanje praktičnih vprašanj ima za posledico to, da razmišljamo o praktičnih učinkih in implikacijah[6], npr. o eksistenčni ogroženosti[7], pogoji katere niso vezani na to, da stroji morajo biti sposobni misliti.[8] Če torej lahko imajo stroji (brez da pridobijo človeško kognicijo) takšen družbeno disruptiven vpliv, potem moramo v ospredje postaviti vprašanja o praktičnih posledicah UI. Še četrtič, antropomorfizacija strojne kognicije kvečjemu omeji učinkovito razvijanje rešitev problema nadzora. Če namreč pričakujemo človeške lastnosti pri strojih, lahko to pričakovanje poveča verjetnost izdajniškega obrata.[9]

Naj izpostavimo tipičen primer preokupacije z metafizičnim vprašanje mišljenja pri strojih. Mindt in Montemayor (2020) sta predstavila razdelitev inteligentnih sistemov, vodilna vprašanja njunega eseja pa so na primer: »Kaj bi pomenilo, da bi sistemi prešli iz zgolj inteligentnega izvrševanja naloge v dobro izvedljivo nalogo?« in »Kakšna je povezava med zavestjo in inteligenco, tako da se lahko podajo posebne ocene o zavestni umetni inteligenci?« Ponujata sicer uporabno taksonomijo za navigacijo pri postavljanju teh vprašanj, a pri tem nekritično obravnavata možnosti za realizacijo modelov, ki so t. i. proizvajalci znanja.[10]

---

[5] Problem drugih duhov (Avramides 2020). Filozofski zombi je opredeljen kot natančen fizični dvojnik brez pojavnih lastnosti (Chalmers 1996: 95-96).

[6] Pri tem velja pripomniti, da sem spadajo tudi razprave o okoljski etiki predhodno usposobljenih jezikovnih modelov, ki za usposabljanje potrebujejo ogromne količine podatkov in računalniške moči – problem, ki je bil eden od poglavitnih razlogov za to, da je raziskovalka Timnit Gebru izgubila mesto pri Googlu. Sporna je bila objava raziskave »O nevarnostih stohastičnih papagajev: so jezikovni modeli lahko preveliki?« (2021), ki določa tveganja velikih jezikovnih modelov, usposobljenih za procesiranje neverjetne količine besedilnih podatkov. Prav tako praktične posledice zadevajo vse od posnemanja človeškega jezika, koherentnega pisanja in potenciala širjenja lažnih novic do že tako vsepristone algoritmizacije vseh digitalnih vidikov z edinim ciljem pritegnitve in ohranitve pozornosti.

[7] To so grožnje, ki bi lahko povzročile naše izumrtje in za katere velja – zaradi njihove kompleksnosti – da je običajno upravljanje tveganj neučinkovito (Bostrom 2013: 15).

[8] Lahko imajo zgolj velik družbenotransformativni potencial, ti. transformativna umetna inteligenca (Karnofsky 2016).

[9] V izvirniku *treacherous turn*, ki ga Bostrom (2014) opredeli kot idejo, da se SI v eni točki lahko nauči zavajanja.

[10] Z razliko od orodij znanja so proizvajalci znanja sposobno proizvesti dodaten iznos, npr. sposobnost metaučenja (Mind in Montemayor 2020: 14). Dosedanji sistemi UI se večinoma uvrščajo med orodja, saj so pod nadzorom človeka in nimajo notranjih stanj oz. potreb. Sleherna avtonomija – npr. za igranje igre – je vnaprej določena. Pri tem delata primerjavo s človeško avtonomijo in intencami za dosego ciljev. Vprašanja o potencialni disruptivnosti proizvajalcev znanja ju ne zanimajo.

Na tej točki naj torej zaključimo z razpravo o možnosti mislečih strojev, ker je takšna razprava vzpostavljena znotraj problematičnih in potencialno nevarnih antropomorfnih predpostavk o strojni kogniciji. V nadaljevanju se bomo torej posvetili praktičnemu vprašanju in pokazali smiselnost artikuliranja odgovorov na vprašanje eksistenčne ogroženosti tudi v primeru, ko nimamo opravka z notranjimi stanji strojev, ki so podobna človeški kogniciji.

4 Problem nadzora

V ospredje bomo tako postavljali praktična vprašanja, tj. problem nadzora, z njim pa tudi vprašanje, kako zagotoviti varnost pred morebitno disruptivno UI.[11] En odgovor je, da varnost pred SI zagotovimo empirično z opazovanjem njenega vedenja, ko je v nadzorovanem, omejenem okolju ('peskovnik'),[12] in da UI spustimo iz peskovnika samo, če vidimo, da se obnaša prijazno, sodelovalno in odgovorno.

Primer takega pristopa je sprotni sistem preverjanja etičnosti nekega UI sistema, v nasprotju s t. i. konceptom kurativnega »velikega rdečega gumba«[13] (Arnold in Scheutz 2018). Pojmovanje nadzora UI sistemov skozi dokončni izklop je nezadostno, saj je po eni strani prežeto s senzacionalistično obarvanimi scenariji glede potencialnih nevarnosti UI sistemov v daljni prihodnosti, po drugi strani pa VRG implicira, da je škoda že storjena, saj se na podlagi le-te upravitelji UI sistema odločijo za izklop. Boljši pristop, ki ga avtorja predlagata, je izoliran modul za sprotno samoocenjevanje in testiranje, s katerim se postavlja sprotna diagnostika, na podlagi katere se tveganje zniža oziroma prepreči.

Avtorja predlagata, da se ni treba ozirati na pogubne scenarije in SI, da razmišljamo o etičnosti UI sistema.[14] Nedavni dosežki strojnega učenja (zmaga v igri GO, slikovno prepoznavanje, procesiranje naravnega jezika, samovozeča vozila) kažejo na to, da so rigidna logična pravila robotike nezadostna za krotenje algoritemskega avtonomnega učenja. VRG je reakcija na to dognanje in način, kako nasloviti grožnje UI sistemov, preden ti postanejo uničujoči. Konkretno, VRG je način, kako preprečiti UI sistemu, da manipulira načine, preko katerih ga je možno ugasniti. Toda, kot izpostavljata avtorja, točka intervencije z VRG nastopi prepozno, šele

[11] Lahko je USI, SI ali zgolj algoritem, ujet v temno neskončnost (brezizhodno ponavljanje izvajanje neke operacije).
[12] V izvirniku *sandbox*.
[13] VRG.
[14] Problem, kot bomo videli, na katerega avtorja pozabljata, je ta, da je sama možnost SI pobijajoča kritika njunega 'etičnega preverjanja'.

tedaj, ko je sistem že "podivjan". Poleg tega mora ta tip intervencije preprečiti možnost, da sistemi, ki temeljijo na spodbujevalnem učenju (angl. *reinforcement learning*), ne uspejo prilagoditi nagradnih funkcij tako, da povečajo nagrade, kadar preprečijo svoj izklop. VRG prav tako ne naslavlja praktičnih vprašanj (UI sistemov, ki so že v uporabi) in tako ignorira trenutne probleme glede odgovornosti avtomatiziranih sistemov.

Primeri preprečevanja vplivov na VRG:

- spodbujevalni sistemi so lahko varno prekinljivi (angl. *safely interruptible*), kot predlagata Orseau in Armstrong iz skupine *Google Deep Mind* (2016), saj se lahko UI sistem nauči, da so prekinitve ovire pri doseganju nagrad;
- Reidl (2016) se osredotoča na izgon sistema v simulirano okolje kot način, kako mu preprečiti, da vpliva na človeško kontrolirano stikalo;
- Hadfield-Mennel (2016) predlaga negotovost ključnih nagradnih funkcij kot način, kako sistemu preprečiti vrednotenje VRG-ja.

Vsi ti pristopi predpostavljajo sistem spodbujevalnega učenja. Prvi pristop ilustrira načine, kako sistemu preprečiti, da dojema prekinitve kot ovire na poti do nagrad, drugi pristop se prekinitvam docela izogne v prid izgonu v simulacijo, tretji pristop pa predpostavlja netransparentnost nagradnih funkcij. Cilj pri vseh je enak; sistem ne sme asociirati VRG z grožnjo.

Toda kako vemo, kdaj pritisniti VRG? Ti pristopi namreč ne razvijajo strojne etike. Kaj točno torej sistem počne, preden ga izklopimo, je stvar nadzornikove interpretacije. Tudi če bi sistem imel integrirano etiko, kako bi vedeli, da ne gre zgolj za posnemanje zavoljo izognitvi pritiska na VRG?

Idealni način ocenjevanja UI sistema tako ni retrospektivna analiza, ampak sprotna verifikacija. Zaradi tega je sistemu potrebno onemogočiti, da stori več od zgolj blokiranja VRG. Sistem ne sme imeti dostopa do vodljivih, etičnih principov, tako so sami načini preverjanja sistemu nedostopni. Vse testiranje je torej sistemu nedostopno, čeprav se izvršuje vzporedno z normalnim delovanjem. To se doseže z izoliranim modulom etičnega jedra, tj. sloj nad strojno opremo, ampak pod virtualnim modulom (Arnold in Scheutz, 2018: 62). UI programska oprema v virtualnem modulu ima torej dostop do senzorjev in strojne opreme, ampak ne do

svojih notranjih operacij. S tem, ko se etično jedro umesti pod virtualni modul, na katerem tečejo razni operacijski sistemi in sam UI, UI sistem nima načina za zaznavo dodatnega etičnega modula. Na tak način UI sistem ne more anticipirati oziroma vedeti, kdaj je podvržen testiranju. Tako umeščeno etično jedro sestavljajo:

- simulacijsko okolje, ki je dovolj realistično, da UI sistem ni zmožen diskriminirati med realnimi vnosi in simulacijo;
- generator etičnih scenarijev, ki so neponovljivi in jih UI sistem ne zmore predvideti;
- minimalno delujoč sistem izklopa, ki lahko po potrebi izklopi UI sistem;
- povezave do senzorjev in aktuatorjev (*angl. actuators*), ki so prestregljive in preusmerljive in ki omogočajo shranjevanje, kloniranje in zagon virtualnih okolij, v katerih se UI sistem nahaja.

V vsakem primeru testiranje UI sistema v danem trenutku naredi kopijo sistema, ki je podvržena simuliranemu scenariju in testirana. Vzporedno s tem prvotni UI sistem deluje normalno. Brezhibno kloniranje in preusmeritev senzoričnih vnosov na simulacijo zahteva grafično in scenarijsko natančnost slednje, saj lahko v nasprotnem primeru klonirana verzija uvidi, da je podvržena testiranju (Arnold in Scheutz 2018: 66). Če testirana verzija test opravi, se vzporedni proces testiranja zaključi, prvotni UI sistem pa nadaljuje svoje delovanje. Če testa ne opravi, se lahko zaradi slojevite arhitekture izklopita tako testirana kot prvotna verzija UI sistema.

Tak način sprotnega testiranja, ki temelji na slojeviti arhitekturi modulov in izoliranosti etičnega jedra, je nujen, saj se ne smemo zanašati zgolj na nujne izklope v primeru, ko nek UI sistem že predstavlja grožnjo oziroma je že izpeljal neko operacijo, ki jo upravitelj presodi kot dovolj sporno, da zahteva pritisk na VRG. Etika UI sistemov, skratka, ne sme temeljiti zgolj na snovanju sistemov za nujni izklop, ampak zahteva kontinuirano testiranje, ki je imuno na manipulacijo nagrad s strani spodbujevalnih sistemov.[15]

[15] V okviru projekta, ki je preučeval razvoj (simuliranih) organizmov, je Claus Wilke (s skupino raziskovalcev, 2001) želel omejiti stopnjo razmnoževanja digitalnega organizma. Sistem je programiral tako, da se po vsaki mutaciji zaustavi, izmeri hitrost replikacije mutanta v izoliranem testnem okolju in mutanta izbriše, če se je repliciral hitreje kot nadrejeni. Vendar so se organizmi razvili, da so prepoznali, kdaj so bili v testnem okolju, in se 'igrali mrtve' (zaustavitev ponovitve), da jih ne bi izločili in jih namesto tega zadržali v populaciji, kjer bi se lahko še naprej razmnoževali zunaj testnega okolja. Ko je to odkril, so randomizirali vnose testnega okolja, tako da ga ni bilo mogoče tako enostavno zaznati, toda organizmi so razvili novo strategijo za izvajanje nalog, ki imajo veliko verjetnost, da bi lahko pospešile njihovo razmnoževanje. Tako so vsaj nekatere različice prestale testiranje. (Wilke in drugi, 2001)

Pomanjkljivost ideje sprotnega testiranja je, da predpostavlja, da je to izvedljivo. Lepo in zaželeno vedenje v nadzorovanem okolju sestavljanja je t. i. konvergentni instrumentalni cilj tako za prijazne in neprijazne UI. Sistem umetne inteligence (ali podsistem) se lahko nauči zaznati, kdaj je nadzorovan/testiran, in spreminja svoje vedenje med nadzorom/testiranjem, tako da njegove neželene lastnosti (oblikovalcev) ostanejo neopažene (Bostrom 2014: 136).

> Ideja je, da varnost superinteligentne umetne inteligence potrdimo empirično z opazovanjem njenega vedenja, ko je v nadzorovanem, omejenem okolju (»peskovnik«), in da umetno inteligenco spustimo iz škatle samo, če vidimo, da se vede na prijazen, sodelovalni, odgovoren način. Pomanjkljivost te ideje je, da je lepo vedenje v škatli konvergentni instrumentalni cilj za prijazne in neprijazne umetne inteligence. Neprijazna umetna inteligenca z zadostno inteligenco se zaveda, da bo svoje neprijazne končne cilje najbolje uresničila, če se bo na začetku obnašala prijazno, tako da bo izpuščena iz škatle. Obnašati se bo začela neprijazno šele, ko ne bo več pomembno, ali mi to izvemo; to je, ko je umetna inteligenca dovolj močna, da je človekovo nasprotovanje neučinkovito. (Bostrom 2014: 146)

Neprijazna umetna inteligenca z zadostno inteligenco se zaveda, da bo svoje neprijazne končne cilje najbolje uresničila, če se bo na začetku obnašala prijazno, tako da bo izpuščena iz testnega okolja. Neprijazna bo postala šele tedaj, ko bo umetna inteligenca dovolj močna, da je človekovo nasprotovanje neučinkovito (Bostrom 2014: 147).

5 Argument pogube

Možnost izdajniškega obrata zaostri eksistenčno grožnjo, t. i. vrsto groženj, ki predstavljajo nevarnost celotni prihodnosti človeštva. Argument iz pogube[16] je kombinacija možnosti UI – prednosti prvega gibalca (biti v poziciji, da UI počne, kar želi), teze o pravokotnosti (to, kar želi, je lahko karkoli) in konvergentnih instrumentalnih vrednot (ne glede na želje bo delovala pri pridobivanju virov in izničevanju nevarnosti zanjo)[17] človeštvu nakazujejo na propad.

[16] V izvirniku *Doomsday Argument*, tako v poglavju »Je privzeti izid poguba?« (angl. *Is the Default Outcome Doom*) (Bostrom 2014: 140).

[17] Človeška bitja predstavljajo koristne viri kot so »priročno nameščeni atomi« in ostali lokalni viri, ki jih izrabljamo. (Bostrom 2014: 116).

Situacija prvega gibalca bo za UI edinstvena strateška priložnost, saj bo v poziciji, da ustvari nov svetovni red, v katerem obstaja ena sama, najvišja raven odločanja. Med njegove pristojnosti bi spadala (1) sposobnost preprečevanja kakršnih koli groženj lastnemu obstoju in nadvladi ter (2) sposobnost učinkovitega nadzora nad glavnimi značilnostmi svojega področja (Bostrom 2014: 141).

Takšna stopnja inteligence je v skladu s skoraj vsakim končnim ciljem. Zatorej ne moremo domnevati, da bo imela katero od dobronamernih vrednot ali ciljev.[18] Težko je opaziti, ali je umetna inteligenca nevarna s svojim vedenjem v času, ko bi jo lahko izklopili, ker imajo umetne inteligence konvergentne instrumentalne razloge, da se pretvarjajo, da so varne in prijazne, četudi niso. Zato bi prva SI zlahka imela neantropomorfne končne cilje in bi verjetno imela instrumentalne razloge za nadaljevanje pridobivanja virov brez konca (Bostrom 2014: 116).

6 Hipoteza o ranljivem svetu

Problematiziranje SI spada v širšo domeno eksistenčnega raziskovanja[19], ki jo Bostrom (2019) poimenuje hipoteza o ranljivem svetu. Osnovna ideja je ta, da ljudem znanstveni in tehnološki napredek nudita vse več različnih zmožnosti, ki utegnejo destabilizirati civilizacijo. Hipoteza ranljivega sveta (HRS) se tako glasi:

> HRS: "Če se nadaljuje tehnološki razvoj, bo v določenem trenutku dosežen nabor zmogljivosti, zaradi katerih je civilizacijsko opustošenje izjemno verjetno, razen če civilizacija izstopi iz polanarhičnega privzetega stanja." (Bostrom 2019: 457)

Privzeto stanje označuje stanje, v katerem imajo družbe omejene kapacitete za preventivne politike in globalni nadzor ter razpršene motivacije. Če izstop iz privzetega stanja ni mogoč, tj. če civilizacija ni zmožna implementirati preventivnih ukrepov, potem na podlagi HRS sledi gotov propad civilizacije. Del preventivnih ukrepov je ravno problem nadzora. Toda, neodvisno od eksistenčne pomembnosti

[18] Tudi ob predpostavki dobronamernosti se UI interpretacija človeške blaginje lahko radikalno razlikuje od človeškega pojmovanja. Na tej točki določenih miselnih eksperimentov ne bomo omenjali, ker predstavljajo t. i. informacijsko nevarnost (angl. *informational hazard*), v neakademskih razpravah tudi t. i. memetična nevarnost. Bostrom razdela tipologijo informacijskih nevarnosti gleda na potencialno škodo zaradi prenosa znanja (Bostrom 2012).

[19] V profilnem članku za New Yorker so zapisali, da Bostrom vodi inštitut kot nekakšno filozofsko radarsko postajo: bunker, ki pošilja navigacijske impulze v meglico možne prihodnosti (Khatchadourian 2016).

ukvarjanja s problemom nadzora in pionirskih naporov pri vpisovanju človeških normativnih, etičnih parametrov v ustroj umetne inteligence, ostaja možnost, da je vse to zaman. Omenjeni projekti namreč temeljijo na antropomorfnih predpostavkah o umetni inteligenci in njeni kogniciji in če smo kognitivno zaprti do preprostih algoritmov (kot nakazuje t. i. *problem črne škatle*[20]), kako lahko pričakujemo neoviran spoznavni dostop do nečesa takšnega, kot je SI? Legitimno lahko postavimo vprašanje: Kaj če je splošna umetna inteligenca že tukaj?[21]

Da bi lahko odgovorili na zgornje vprašanje, bomo pogledali razvoj GPT[22] modelov in z njihovo pomočjo ilustrirali njeno smiselnost. Ob tem bomo predstavili t. i. hipotezo o nadgradljivosti (angl. *scalability hypothesis).* Slednja je namreč ključnega pomena za težo zgornjega vprašanja, saj ponuja utemeljitev pozitivnega odgovora. Hipoteza o nadgradljivosti se nanaša na napoved, da bomo postopoma videvali vse večje preboje zmožnosti UI s povečevanjem nabora parametrov[23] in podatkov.[24] Hipoteza o nadgradljivosti nudi podporo pojmu t. i. transformativne[25] UI. Pokazali bomo, da lahko na podlagi GPT primera smiselno postavimo zgornje vprašanje o aktualnosti t. i. transformativne UI, ki je sposobna imeti vpliv na človeštvo, kar je primerljivo z industrijsko revolucijo.

[20] Za nas netransparenten sistem, za katerega ne vemo, kako je prišel do rezultatov (Markič 2020: 209).

[21] Verjetnost vzleta (predvsem na blogih se uporablja izraz *foom*) oz. eksplozija rekurzivne samoizboljšave. Ob predpostavki, da bi lahko en sam programski projekt, ki se začne z majhnim delom svetovnih virov, v nekaj tednih postal tako močan, da prevzame svet. Kaj je bolj verjetno, da živimo v času tik pred S(U)I ali, da se je to že zgodilo? Čas tik pred nastankom SI (ne glede na to ali mi to vemo ali ne – koncepcija zavajanja (angl. *the conception of deception)*) nedvomno zavzema manjši časovni interval kot čas po SI, iz tega sledi, da je bolj verjetno, da živimo v času po vzletu SI. Podobno lahko na enak način razmišljamo, da je bolj verjetno, da živimo v simulaciji kot pa v bazični realnosti, tik pred SI vzletom. Zakaj bi SUI sploh želela zaganjati simulacije vesolja pa je vprašanje, na katerega zaradi informacijske nevarnosti ne želimo podati odgovora.

[22] Splošni vnaprej usposobljeni jezikovni modeli (angl. *General Pre-Trained Transformers*). BERT, 500-krat manjši model (tehnično je način učenja) od GPT-3, je recimo eden izmed jezikovnih modelov, ki ga uporablja Google za svoj iskalnik.

[23] Parametri so spremenljivke, ki se uporabljajo za nastavitev in prilagoditev modelov umetne inteligence.

[24] Turingov nagrajenec Geoffrey Everest Hinton, ena od pomembnih osebnosti v razvoju globokega učenja, se je v čivku pošalil, da: »ekstrapolacija spektakularne zmogljivosti GPT-3 v prihodnost kaže, da je odgovor na življenje, vesolje in vse samo 4398 milijard parametrov.«

[25] Lahko je SI ali preprosto samoizboljševalni algoritem, označuje pa že prej omenjeno UI z disruptivnim potencialom za družbo in svet.

7 GPT[26] in vključenost v svet

Preteklo leto 2020 je bilo prvo, v katerem so jezikovni modeli UI postali ekonomsko uporabni. Zlasti GPT-3 je pokazal, da imajo veliki jezikovni modeli presenetljivo jezikovno sposobnost opravljanja najrazličnejših uporabnih nalog. Pričakuje se, da bodo jezikovni modeli postajali vse kompetentnejši, tako da bodo najboljši modeli leta 2020 v primerjavi z njimi videti dolgočasno in preprosto.[27] To pa bo odklenilo aplikacije, ki si jih danes težko predstavljamo. Leta 2021 se bodo jezikovni modeli začeli zavedati vizualnega sveta. Samo besedilo lahko namreč izraža veliko informacij o svetu, vendar je nepopolno, saj živimo tudi v vizualnem svetu. Naslednja generacija modelov bo tako lahko urejala in ustvarjala slike kot odziv na vnos besedila. Predpostavlja se, da bodo besedilo bolje razumeli zaradi številnih slik, ki so jih videli. Ta sposobnost skupne obdelave besedila in slik bi morala narediti modele pametnejše.[28] Ljudje smo izpostavljeni ne le temu, kar preberemo, ampak tudi tistemu, kar vidimo in slišimo. Če lahko modele izpostavimo podatkom, podobnim tistim, ki jih absorbiramo ljudje, bi se morali naučiti pojmov, ki so podobni našim.

S tem se predpostavlja t. i. hipoteza naravne abstrakcije oziroma naravnih pojmov. Za namene tega besedila je ne bomo problematizirali, čeprav vsebuje nekritično obravnavo same narave pojmov. Predpostavlja namreč, da bodo novi modeli UI s tem, ko bodo opremljeni s senzorji in tako vključeni v svet, prihajali do podobnih pojmov kot ljudje. Konkretno, predpostavlja se tesna povezanost med vsebinami abstrakcije in načini vnosa – jezikovni model, ki vizualno procesira besedilo, je različno določen pri pridobitvenem načinu informacij kot pa model, ki besedilo pridobi preko nesenzoričnih načinov. Hipoteza naravne abstrakcije je, ker temelji na 'utelešenosti', predvsem izpostavljena kritiki računalniškega mišljenja iz perspektive, ki se osredotoča na abstraktno naravo znanstvenih modelov, med katere spadajo tudi računalniški modeli možganov (Chirimuuta 2020: 424). Navadno so bili ugovori računalniški teoriji mišljenja osnovani na fenomenologiji dvajsetega stoletja, s posebnim poudarkom na utelešenje in vdelanost inteligence (Dreyfus 1972). Prej

[26] GPT spadajo v t. i. pozni drugi val UI, ki se v nasprotju s prvim (močna UI, ki razume mišljenje kot simbolno manipulacijo), obračajo k induktivnemu sklepanju – na podlagi baz podatkov, izkušenj in interakcije z okoljem. Programe, kot so navodila za simbolno manipulacijo, so zamenjali algoritmi za strojno učenje. Če prvi val zaznamuje dobro definirano okolje (npr. igre, programi, simboli) in deduktivna logika, potem drugi val zaznamuje odprt empiričen svet, poln negotovosti, za katerega je ustreznejša indukcija (Markič 2021: 209).

[27] Tako Ilya Sutskever, soustanovitelj OpenAI, komentira za spletno revijo The Batch.

[28] V skladu z Dreyfusovo kritiko prvega vala UI, da za vsakdanje znanje potrebujemo drugačno obliko predstavitve informacij. Po njegovem je eden izmed ključnih pogojev za to, da bi UI posedovala razumevanje, vključenost v svet (Markič 2021: 209).

omenjena kritika računalniškega mišljenja preko problematizacije znanstvenih modelov pa temelji na Whiteheadovem pojmu zmote napačno postavljene konkretnosti[29] in se tako pogoju utelešenosti popolnoma izogne.

8 Zgodba o GPT

Začelo se je leta 2018, ko so pri OpenAI izdali prvi GPT model, ki je imel velik vpliv na tedanjo UI skupnost. Še večji vpliv je imel izid GPT-2 modela v začetku leta 2019, čeprav ga pri OpenAI niso želeli izdati v celoti zaradi zaskrbljenosti zlonamerne uporabe. (Radford et al. 2019)

GPT-2 je bil predhodno usposobljen z raznoliko, 40 GB veliko vsebino, postrgano z interneta – z enim preprostim ciljem: da predvidi naslednjo besedo glede na vse prejšnje besede v nekem besedilu.

Tudi okrnjen model je produciral presenetljive rezultate, kot je znana zgodba o samorogih, pri kateri je človeški vnos besedila v slogu naslova in podnaslova zgodbe o odkritju samorogov na območju gorovja Andov. GPT-2 je na podlagi te vsebine in sloga izvozil slogovno primerljivo in skladno besedilo, ki vsebuje vse elemente tipske reportaže, tj. od biološke opredelitve samorogov, imen glavnih raziskovalcev, njihovih citatov in do spekulacij glede izvora samorogov. Nič od tega ni bilo zajeto pri vnosu. Vzorci tekstov, kot je omenjena zgodba s samorogi, imajo pomembne posledice, saj je velike jezikovne modele vedno lažje usmerjati v prilagodljivo, prilagojeno in skladno ustvarjanje besedila, ki bi se lahko nato uporabljalo na številne koristne in zlonamerne načine.

Naslednje leto, 2020, izide GPT-3, ki je, če pogledamo njegovo velikost, 100-krat večji od predhodnika. GPT-3 ima 175 milijard parametrov. Torej je GPT-3 100-krat večji od predhodnika GPT-2, ki je bil že izjemno velik, ko se je pojavil leta 2019. Povečanje števila parametrov 100-krat z GPT-2 na GPT-3 ni prineslo le količinskih razlik. GPT-3 ni le močnejši od GPT-2, ampak je tudi zmogljivejši – na drugačne načine; med obema modeloma je kvalitativni preskok. GPT-3 lahko počne stvari, ki jih GPT-2 ne more. Če se GPT-3 lahko nauči učiti, kdo ve, kaj lahko prinese

[29] Zmota zamenjave abstrakcij znanosti za konkretne stvari na svetu, iz katerih abstrakcije izhajajo (angl. *fallacy of misplaced concreteness*) (Chirimuuta 2020: 430; Whitehead 1928: 66).

GPT-4; morda bomo videli prvo nevronsko mrežo, ki je sposobna resničnega sklepanja in razumevanja.

9 GPT in transformativna UI

GPT verjetno ne bo postal SI, ne glede na to, kako velik je GPT-3, je namreč splošno orodje, ki ga je mogoče izvesti za različne naloge in je morda res sposobnejše izvajati večje število nalog, ampak spontano ne bo sposoben početi stvari, ki so ključne za pristno SI – na primer razumevanje vzročnosti – saj preprosto nima potrebne arhitekture.

Kar bi bilo sicer zelo prepričljivo, bi bil npr. izjemno dober klasifikator slik in videov, ki je opremljen z umetnimi čutili za interakcijo s svetom in dobro obdelavo naravnega jezika, ki je sposoben prepoznati tudi relacijske besede in dejanja, npr. prepoznavanje videoposnetka osebe, ki položi vrček na mizo. Predstavi se mu vrček in ukaže, »Daj vrček na mizo«, pri čemer besede razčleni v besedilo in v svoji bazi podatkov o usposabljanju poišče videoposnetke, ki so zelo podobni temu opisu, pripravi celoten video, prepozna vizualni vhod skodelice kot isti simbolni predmet kot sestavljeni vrček, nato pa, da se vizualni vnos natančno ujema s tem agregatom, s poskusi in napakami in s strojnim učenjem spozna pravilen način uporabe udov – da vrč postavi na mizo. In nato, kar je ključnega pomena, shrani pravilne v spominu in na splošno izmeri stopnjo učinkovitosti iz svojih veščin (se izboljša pri 'dajanju', pri 'vklopu' itd.), ki jih je mogoče uporabiti za prihodnje naloge. Zdi se, da bi ga to približalo nečemu, kar bi lahko označili z 'razmišljanjem'.

Tak sistem še vedno ne bi nujno razumel vzročnosti same po sebi, vedel pa bi, kakšne stvari mora oddati, da izpolni svojo funkcijo koristnosti in kdaj mora ukrepati. A kljub temu da GPT in ostali modeli ne bi postali SI, še več, četudi SI kot taka sploh ni možna, postaja na podlagi zmogljivosti teh modelov, ki dajejo podporo hipotezi o nadgradljivosti, pojem transformativne UI vse bolj verjeten in smiseln. Transformativna UI je UI, ki lahko pospeši družbeni napredek/prehod, primerljiv s kmetijsko ali industrijsko revolucijo (ali celo takšnega, ki bi bil pomembnejši od nje) (Karnofsky 2016). Transformativna UI ni nujno SI in je v tem oziru nevtralna v primerjavi s potezami človeškega uma. Za transformativno AI torej ne šteje, da primerja človeške lastnosti zavedanja, čustev, razumevanja in podobno. Vse, kar je za transformativno UI pomembno, je to, da je sposobna privesti do znatnih sprememb na svetu.

10 Fideizem in UI

Zgornjo razpravo o praktičnih vidikih razvoja UI nekateri označujejo kot vnebovzetje za piflarje,[30] ji očitajo neutemeljenost (Bringsjord et al. 2012) in je ne obravnavajo pogosto v akademskih kontekstih, čeprav se to spreminja. Kljub temu je razvoj drugega vala UI s seboj prinesel določene novosti in premike, s katerimi so pokazali, da lahko s primerno velikostjo UI modelov že sedaj produciramo disruptivne dogodke, npr. prepričljive lažne novice. Vendar je kljub praktičnim napredkom UI vseeno treba odgovoriti na ta sentiment o spoznavni neutemeljenosti svarilcev pred pogubo zaradi UI. Spodaj povzemamo in odgovarjamo na primerjavo svarilcev UI s teističnimi skeptiki (izvaja jo Danaher 2015), pokazali bomo neustreznost te primerjave in še na ta način pokazali spoznavno utemeljenost zgornje razprave in nasploh doslednost svarilcev pred UI.

Osrednji korak za skeptične teiste je sklepanje z videza na realnost zavoljo zagovora Božje moralnosti – četudi se nam zdi, da je v svetu zlo, iz tega ne sledi, da je to tudi res, saj je lahko dozdevno zlo zgolj funkcija za neko drugo (višje) dobro (Denahar 2015: 233). Bostrom (2014) z izdajniškim obratom sicer ubira podobno logiko neupravičenosti induktivnega sklepanja iz empirično danih dokazil, a Bostromov izdajniški obrat kot blokada induktivnega sklepanja iz dozdevne dobronamernosti na dejansko dobronamernost ni enaka sklepanju skeptičnih teistov. Podobnost med UI teoretiki pogube in skeptičnimi teisti sestoji v zguljeni filozofski tezi o različnosti med videzom in realnostjo in spoznavnem opozorilu, ki iz tega sledi, in sicer, da moramo biti previdni pri tovrstnem induktivnem sklepanju iz empiričnega videza na pravo naravo stvari. Danaher (2015), ki razvije to primerjavo, se tako vpraša: »Zagotovo obstajajo nekakšni empirični dokazi, ki bi ga zadovoljili, da UI ne predstavlja tveganja za ljudi?« (Danaher 2015: 239) Pri tem pa ne upošteva ravno te poante, da v luči eksistenčne nevarnosti (Bostrom 2013 in 2014) ni racionalno zaupati nobenim empiričnim dokazilom.

V obeh primerih imamo torej prvotno prepričanje, ki je velikega pomena (obstoj Boga v prvem primeru in možnosti UI pogube v drugem). Ta prepričanja se nanašajo na obstoj nadmočnih in superinteligentnih agentov (Bog ali SI). Prepričanja izpodbijajo nekateri ugovori (problem zla in ugovor empiričnega, sprotnega preizkušanja). Oba ugovora temeljita na ideji, da lahko zanesljivo sklepamo (čeprav

[30] Izvorni naslov je *Rapture of the Nerds* (Doctorow in Stross 2012).

induktivno) od videza dogodka do njegove dejanske narave. V obeh primerih se blokira sklepanje od "navideznega" do "dejanskega" s sklicevanjem na naše kognitivne omejitve: ne vemo popolnoma, kako se to, kar opazimo pri videzu (zlo, vedenje UI), lahko poveže z drugim (nezamisljivimi) končni cilji. Možno je namreč to, da zaradi nepopolnega kognitivnega dostopa (ali do misli Boga ali do UI), obstajajo drugi končni cilji, do katerih nimamo spoznavnega dostopa (Denahar 2015: 235; Bostrom 2014: 158).

Za razliko od skeptičnih teistov je spoznavno opozorilo pri sklepanju z videza na realnost v primeru UI utemeljeno z dejansko eksistenčno nevarnostjo, medtem ko je pri skeptičnih teistih prej orodje za podporo dogme. Prav tako je ena izmed praktičnih posledic za skeptičnega teista moralna paraliza (prekiniti trpljenje in s tem potencialno prekiniti višje dobro ali pa dovoliti odvijanje trpljenja). Teoretik UI pogube tovrstni paralizi ni podvržen; to, da sodeluje pri razpravah, Inštitutih za nadzor UI in ozaveščа o tej problematiki, je skladno s prepričanjem o potencialni eksistenčni nevarnosti UI. Še več, zavora induktivnega sklepa in iz tega izhajajoča previdnost temelji pri UI teoretikih izgube na nomološki možnosti pogube, medtem ko pri skeptičnih teistih zavora sklepanja temelji na obrambi nečesa, ki krši oz. je onkraj nomološke realnosti. In še, pri UI teoretikih pogube previdnost izhaja iz skrbi za človeštvo in preživetje naše vrste, medtem ko pri skeptičnih teistih zavora induktivnega sklepanja ne izhaja iz previdnosti in človekoljubnosti, ampak iz zavezanosti in obrambe Boga.

11 Sklep

V besedilu smo naprej predstavili znano kritiko Turingovega testa, tj. miselni eksperiment 'Kitajska soba'. Nato smo podali kritiko nekaterih predpostavk, na katerih je razprava o mislečih strojih osnovana in ki omogočajo postavitev vprašanja *»Ali stroji lahko mislijo?«*. Izpostavili smo pogled, da lahko spoznavni (ne)dostop do strojne kognicije predstavlja še večji eksistenčni problem, sploh če predpostavimo, da lahko strojna kognicija zavzema nam tuje oblike. Praktična vprašanja in razpravo o eksistenčni nevarnosti smo postavili v kontekst hipoteze o ranljivem svetu. Iz tega smo predstavili in utemeljevali smiselnost vprašanja o možnosti tega, da je umetna inteligenca že prisotna oz. vsaj blizu. S primerom GPT jezikovnih modelov in pojmov nadgradljivosti in transformativne umetne inteligence smo legitimnost tega vprašanja utemeljevali. Na koncu smo obravnavali razširjeno stališče, da je razprava o eksistenčni ogroženosti neutemeljena. Predstavili smo primerjavo sklepanja med

skeptičnimi teisti, ki zanikajo problem zla, in teoretiki pogube, ki vidijo v umetni inteligenci možnost eksistenčne nevarnosti. Čeprav je sklepanje obojih podobno, smo podali razloge, da so teoretiki pogube vseeno bolj utemeljeni v svojem sklepanju. Slednji namreč delujejo znotraj nomoloških možnosti in zaradi svojih prepričanj niso pahnjeni v moralno paralizo, ki preti skeptičnim teistom. V teh ozirih je torej eksistenčna skrb glede praktičnih posledic razvoja umetne inteligence utemeljena. Še več, glede na uspešnost raznih modelov za procesiranje naravnega jezika, ki vse bolj potrjujejo tezo o nadgradljivosti, tj. da torej za 'pravo' UI ni potreben nek kvalitativen preskok, ampak zadošča zgolj nadgradnja v parametrih in podatkih, in glede na pojmovanje UI kot transformativne, tj. takšne, ki za disruptivne učinke ne potrebuje uprimerjati človeških atributov mentalnosti, postaja vprašanje o že aktualizirani različici UI, ki ima pogubno moč, toliko bolj legitimno.

Viri in literatura

Arnold, T. in Scheutz, M. (2018). »The 'big red button' is too late: an alternative model for the ethical evaluation of AI systems«. *Ethics and Information Technology*, 20(1), str. 59–69.

Avramides, A. (2020). »Other Minds«. V Zalta, E. N. (ur.), *The Stanford Encyclopedia of Philosophy* (zima 2020). URL = https://plato.stanford.edu/archives/win2020/entries/other-minds/.

Bender, M. E., Gebru, T., Angelina McMillan-Major in Shmitchell, S. (2021). »On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?« V *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. New York: Association for Computing Machinery, str. 610–623.

Bostrom, N. (2012). »Information Hazards: A Typology of Potential Harms from Knowledge«. *Nick Bostrom's Home Page* (28. junij 2021). URL = https://www.nickbostrom.com/information-hazards.pdf.

Bostrom, N. (2013). »Exsistencial Risk Prevention as Global Policy«. *Existential Risk: threats to humanity* (28. junij 2021). URL = https://www.existential-risk.org/concept.html.

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

Bostrom, N. (2019). »The Vulnerable World Hypothesis«. *Nick Bostrom's Home Page* (28. junij 2021). URL = https://www.nickbostrom.com/papers/vulnerable.pdf.

Bringsjord, S., Bringsjord, A. in Bello, A. (2012). »Belief in the Singularity is Fideistic«.V Eden, A., Moor, J., Soraker, J. in Steinhardt, E. (urd.), *Singularity Hypotheses: A Scientific and Philosophical Assessment*. Dordrecht: Springer, str. 395–412.

Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.

Chirimuuta, M. (2020). »The Reflex Machine and the Cybernetic Brain: The Critique of Abstraction and its Application to Computationalism«. *Perspectives on Science*, 28(3): str. 421–457.

Danaher, J. (2015). »Why AI Doomsayers are Like Sceptical Theists and Why it Matters«. *Minds & Machines*, 25, str. 231–246.

Doctorow, C. in Stross, C. (2012). *The Rapture of the Nerds*. New York: Tor Books.

Dreyfus, H. L. (1972). *What Computers Can't Do*. New York: MIT Press.

Hadfield-Menell, D., Dragan, A., Abbeel, P. in Russell, S. (2016). »The off-switch game«. *IJCAI'17: Proceedings of the 26th International Joint Conference on Artificial Intelligence*. URL = https://arXiv.org/abs/1611.08219v3.

Hofstadter, D. R. in Dennett, D. C. (1990). *Oko duha: fantazije in refleksije o jazu in duši*. Ljubljana: Mladinska knjiga.

Karnofsky, H. (2016). »Some Background on Our Views Regarding Advanced Artificial Intelligence«. *Open Philanthrophy* (28. junij 2021). URL = https://www.openphilanthropy.org/blog/some-background-our-views-regarding-advanced-artificial-intelligence#Sec1.

Khatchadourian, R. (2016). »The Doomsday Invention: Will artificial intelligence bring us utopia or destruction«. *The New Yorker* (28. junij 2021). URL = https://www.newyorker.com/magazine/2015/11/23/doomsday-invention-artificial-intelligence-nick-bostrom.

Markič, O. (2021). »Prvi in drugi val umetne inteligence«. V Malec, M. in Markič O. (urd.), *Misli svetlobe in senc: razprave o filozofskem delu Marka Uršiča*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani, str. 201–215.

Mindt, G. in Montemayor, C. (2020). »A Roadmap for Artificial General Intelligence: Intelligence, Knowledge, and Consciousness«. *Mind and Matter*, 18(1), str. 9–37.

Orseau, L. in Armstrong, S. (2016). »Safely interruptible agents«. Pridobljeno na https://ora.ox.ac.uk/objects/uuid:17c0e095-4e13-47fc-bace-64ec46134a3f., dne 28. 6. 2021.

Radford, A., Wu, J., Armodei, D., Clark, J., Brundage, M. in Sutskever, I. (2019). »Better Language Models and Their Implications«. *OpenAI* (25. junij 2021). URL = https://openai.com/blog/better-language-models/.

Riedl, M. (2016). »Big red button«. *GitHub* (27. junij 2021). URL = https://markriedl.github.io/big-red-button/.

Searle, J. (1990). »Duhovi, možgani in programi«. V Hofstadter, D. R. in Dennett, D. (urd.), *Oko duha: fantazije in refleksije o jazu in duši*. Ljubljana: Mladinska knjiga, str. 361–379.

Schank, R. C. in Robert, P.A. (1977). *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Lawrence Erlbaum Press.

Turing, A. (1990). »Stroji, ki računajo, in inteligenca«. V Hofstadter, D. R. in Dennett, D. (urd.), *Oko duha: fantazije in refleksije o jazu in duši*. Ljubljana: Mladinska knjiga, str. 61–74.

Wilke, C. W., J., O., C. et al. (2001). »Evolution of digital organisms at high mutation rates leads to survival of the flattest«. *Nature*, 412, str. 331–333.

Whitehead, A. N. (1938). *Science and the Modern World*. Harmondsworth: Penguin.

PO SLEDEH UMETNE INTELIGENCE: KAJ NAM O PSIHOLOŠKIH LASTNOSTIH POSAMEZNIKOV POVEDO NJIHOVI DIGITALNI ODTISI?

BOJAN MUSIL, NEJC PLOHL
Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
bojan.musil@um.si, nejc.plohl1@um.si

Sinopsis Podatki kažejo, da ima dostop do svetovnega spleta okoli 90 % evropskih gospodinjstev in da povprečen evropski uporabnik uporabi le-tega nameni slabih šest ur dnevno. Zaradi možnosti, ki jih splet ponuja, so specifike zelo različne, vendar pa uporabo spleta spremlja pomembna stalnica; ljudje ob sprehajanju po virtualni sferi tam, namerno ali ne, puščajo raznolike digitalne odtise. V preteklosti so bili ti podatki bolj kot ne spregledani, razvoj novih tehnologij pa je omogočil podrobnejši vpogled vanje in vodil do spoznanja, da lahko na videz nepomembni digitalni odtisi v resnici nudijo dragocene informacije o posameznikih. Danes tovrstni podatki omogočajo nove načine vplivanja na njihovo vedenje, s čimer jasno posegajo v aplikativno sfero. Ker so lahko naši osebni podatki uporabljeni (in zlorabljeni), je vse več pozornosti znanstvenikov in splošne javnosti usmerjene tudi k pravnim in etičnim vidikom uporabe digitalnih odtisov. V pričujočem prispevku bomo najprej ponudili naš razmislek o umetni inteligenci kot mehanizmu za napredek, v nadaljevanju pa sledi bolj specifična obravnava digitalnih odtisov, ki jih puščamo na spletu. Poglobili se bomo v to, kaj digitalni odtisi lahko povedo o psiholoških lastnostih uporabnikov, poseben poudarek pa bomo namenili tudi njihovi praktični uporabi (npr. prilagajanje oglasov z namenom prepričevanja) in nekaterim etičnim vidikom uporabe.

Ključne besede: inteligentnost, rudarjenje podatkov, digitalni odtisi, osebnostne lastnosti, napovedovanje vedenja

Univerzitetna založba Univerze v Mariboru

DOI https://doi.org/10.18690/um.ff.11.2022.11
ISBN 978-961-286-675-4

FOLLOWING THE FOOTSTEPS OF ARTIFICIAL INTELLIGENCE: WHAT DO DIGITAL FOOTPRINTS TELL US ABOUT THE PSYCHOLOGICAL CHARACTERISTICS OF INDIVIDUALS?

BOJAN MUSIL, NEJC PLOHL
University of Maribor, Faculty of Arts, Maribor, Slovenia
bojan.musil@um.si, nejc.plohl1@um.si

Abstract The research shows that around 90% of European households have access to the internet and that the average European user spends there about six hours daily. However, the specifics of usage being very different, there is an important common denominator; in the virtual sphere, we leave, intentionally or unintentionally, various digital footprints. In the past, these data were overlooked; however, the development of new technologies enabled a more detailed analysis of them and led to the belief that digital footprints offer valuable information about us. Today, such personal data are used to influence the behaviour of people, and so clearly interfere with the applicative domain. Since they could be used (and abused), the attention of scientists and the public has also turned towards legal and ethical aspects of using digital footprints. In this paper, we will first offer our reflections on artificial intelligence as a mechanism for progress, followed by a more specific discussion of digital footprints left on the internet. Particularly, we will dive into the question of what exactly digital footprints can tell us about the psychological characteristics of users, with an additional emphasis on their practical (personalised ads) and ethical aspects of their usage.

Keywords: intelligence, data mining, digital footprints, personality traits, behaviour prediction

https://doi.org/10.18690/um.ff.11.2022.11 DOI
978-961-286-675-4 ISBN

University of Maribor Press

1 Uvod

V zgodbi znanstvenofantastične klasike iz 1950-ih, *Prepovedani planet* (v originalu *Forbidden planet*; Wilcox 1956), zemeljska odprava na oddaljenem planetu Altair IV odkrije tehnologijo, ki priča o davno izumrli civilizaciji. Krelli, kot so se imenovali njeni pripadniki, naj bi dosegli neverjetno napredno stopnjo tehnološkega in znanstvenega razvoja; bili so v vseh pogledih superiorni človeštvu, moralno in tehnološko. Njihova tehnologija je lahko naredila pravzaprav vse že iz samih misli. In podobno kot izginula civilizacija, se s katastrofo sooča tudi zemeljska odprava, pri čemer igra pomembno vlogo starodavna, a napredna krellska tehnologija.

V zelo kratkem orisu enega prvih 'blokbastrskih' filmov znanstvene fantastike je morda vizualna podoba filma nekaj, kar bi zbudilo nasmeške nemalokateremu gledalcu, vajenemu sodobne, z dovršenimi učinki obogatene, filmske produkcije. Pa vendar sama ideja v zgodbi po več kot 60 letih ohranja svežino oziroma aktualnost, ki, zelo pripravno in pravzaprav kar običajno, sega iz sveta fikcije v resnični svet in znanost.

Prvič, vizualizacije tistega zunanjega in/ali tujevrstnega, ki je nosilec superiorne inteligence, praktično ni. Kakšna je bila ta izumrla civilizacija, ne vemo. Edini artefakti so pravzaprav tehnološki artefakti, ki pa nimajo intence posnemati človeka (ali kogarkoli drugega na njegovem mestu) ali ga celo preseči, nadomestiti. Kot se pokaže proti koncu filma, je tehnologija popolnoma 'vsajena' v vrsto, ki jo uporablja. Inteligenca prostranosti 'vsemirja' je enaka širnim globinam notranjega.

Drugič, nekako v duhu katarzičnega časa minulih velikih vojn se ujamemo v vero v moč racionalnosti sistema in znanosti – da bomo preko tega dosegli napredno in popolno družbo. Razplet filma pokaže, da zgolj moč razuma, najsibo v obliki kopičenja in obdelave podatkov, ustvarjanju znanja in vednosti, zanemari tisto drugo plat, ki je pravzaprav na koncu usodna.

Stara debata se sprašuje, ali se znanstvena fantastika idejno napaja iz resničnega, iz sveta znanosti; ali velja obratno, da imaginacija znanstvene fantastike inspirira in tudi konkretneje usmerja znanstveno produkcijo - in tega v tem prispevku načrtno ne bomo razreševali. Izhajali bomo iz obstoječega stanja, kjer se kopičenje podatkov, znanja in vednosti meri v večini nam neznanih grških predponah, ki označujejo

količino podatkov. Pri tem ne gre več za elaborirane ideje, ki bi jih objavljali na spletiščih, ampak v veliki meri za surove podatke naših digitalnih vedenj, ki jih s pomočjo tehnologije procesiramo oz. se v napredni obliki tehnologija iz njih uči. In izziv je urediti podatke na način, ki bo karseda predvidljiv in bo rezultiral v širšem racionalnem sistemu. V to vstopa uvodna zgodba, kjer, da končno izdamo razplet filma, v sofisticirani hiperinteligentni sistem vdre materializiran stvor iracionalnih (nezavednih) plati, ki ta približek popolnega sveta uniči. Kot je v filmu freudovsko poimenovana – id pošast (v originalu *Id Monster*).

V nadaljevanju bomo poskušali prikazati razvoj obravnave psihološko relevantnih digitalnih podatkov, ki jih ljudje puščamo na spletu in pripadajočih tehnologijah, in obogatiti to s primeri raziskav in uporabe le-tega v vsakdanjem (digitalnem in nedigitalnem) svetu. In, nekako v maniri id pošasti, odpreti tudi nekaj vprašanj, ki se navezujejo na etični vidik takšnega početja.

2 Od moči uma do digitalnih odtisov

Če bi iz psihološke tradicije poskušali razmišljati o moči uma oziroma sposobnosti umskega delovanja, ki v grobem opredeljuje pojem inteligentnosti,[1] bi uporabljali izraze kot hitrost procesiranja, zmožnost učenja, učinkovito reševanje miselnih problemov, uporaba kompleksnih miselnih operacij, tudi učinkovitost socialne interakcije ali prilagajanja okolju. Že iz tega nabora, ki se navezuje na različne opredelitve inteligentnosti, je razvidna usmerjenost na (učinkovite) miselne strategije, postopke, skratka usmerjenost na proces. In resnično, če pod inteligentnim delovanjem razumemo to, kako zmoremo učinkovito misliti, dosegati vednost, je vsebina misli ali vednosti drugotnega pomena. Iz tega lahko razumemo, da v pojmovanjih inteligentnosti še vedno ostaja aktualnost splošnega (g) faktorja, ki kaže na splošno (generalno) sposobnost posameznikovega uma ne glede na specifično področje umskega delovanja.[2] Ob tem pa obstajajo tudi drugačni pogledi na človekove umske zmožnosti, ki namesto splošne zmožnosti poudarjajo področno ali komponentno vezane sposobnosti (npr. Gardner 1983; Sternberg 1985) ali delitev

[1] V slovenščini imamo s pojmom inteligenca ali inteligentnost določeno zagato. Na področju psihologije se za sposobnost umske dejavnosti uporablja prvi izraz, v vsakdanjem jeziku bi lahko rekli bistrost, drugi (inteligenca) pa je rezerviran za poseben, elitni del prebivalstva, z drugimi besedami izobraženstvo.

[2] Prvi je o tej konsistentni tendenci na različnih kognitivnih testih govoril že psiholog Spearman (1904) in s tem spodbudil številne raziskave, ki so še danes precej aktualne (npr. Chabris 2007; Gottfredson 2002). Na smiselnost povezave različnih umskih sposobnosti v splošen faktor naj bi kazale tudi evidence iz študij drugih vrst primatov (npr. Reader, Hager in Laland 2011).

splošne inteligentnosti na ti. fluidno in kristalizirano inteligentnost (npr. Cattell 1963 in 1971), kjer se prva navezuje na duševne (kognitivne) procese in je relativno neodvisna od izkušenj posameznika, druga pa na naučene postopke in znanje ter je posledično vezana na pridobljene izkušnje posameznika. Kljub temu navideznemu dualizmu pa je inteligentnost prve vrste (tj. fluidna) vsaj implicitno še dandanes videna kot tisto, kar naj bi pomensko v vsakdanji rabi naslavljala beseda 'pamet' – od narave dano moč intelekta.

V tej perspektivi lahko vidimo tudi dogajanje od sredine 20. stoletja, ko se je začelo pojavljati interdisciplinarno področje kognitivne znanosti, ki je polje raziskovanja (raz)uma in njegovih procesov širilo iz/na področja psihologije, filozofije, jezikoslovja, antropologije, nevroznanosti računalništva in informatike. Z razvojem tehnologije, posebej s pojavom računalnika, je vprašanje umnega delovanja postalo zanimivo za polje, ki ni izključno vezano na človeka ali živa bitja; posledično lahko v kontekstu računalnikov in drugih strojev govorimo o umetni inteligenci. In še dandanes so pomembna vprašanja v ozadju računalnikov in izpeljanih naprav povezana s potencialom in močjo, učinkovitostjo sistema in kapaciteto, npr. vprašanja procesorske moči, kapacitete trdega diska in delovnega spomina.

Kje pravzaprav v to zgodbo začne (znova) vstopati vsebina oz. natančneje podatki, informacije? Če je osredotočenost na (kognitivni) proces zgodovinsko razumljiva kot nekakšna nadgradnja in posledičen obrat pozornosti od behavioristične formule dražljaja in odgovora, v katerih se je vsaj implicitno skrivala vsebina, pa je pravzaprav vrnitev slednjih v vsej slavi povezana z razvojem učinkovitosti procesiranja in kapacitete računalniških naprav. Bolj kot je bilo možno hitreje in temeljiteje procesirati večje število zelo raznolikih podatkov, bolj se je pozornost usmerjala na samo vsebino podatkov. Lahko rečemo, da je k temu znatno pripomogla naša stvarnost vzporednega digitalnega življenja na svetovnem spletu in posledično enormno kopičenje podatkov različnih kakovosti na njem. Po Harlowu in Oswaldu (2016) aktualno področje velikega podatkovja (angl. *big data*) zajema shranjevanje, iskanje in analizo velikih količin informacij; med katerimi so sploh v zadnjem času zelo aktualni 'digitalni odtisi', ki se nanašajo na nabor sledljivih digitalnih dejavnosti posameznikov na spletu in digitalnih napravah.

Morda je iz uvodne zadrege glede inteligentnosti in/ali inteligence izraz umetna inteligenca posrečen, saj še vedno v predstavah o umetni pameti, le-to radi konkretiziramo – da gre za vsevedne stroje, v večini materializirane, ki so nekoč in nekje nastali s pomočjo človeka. Kaj se z razvojem z njimi dogodi, bomo znova prepustili domišljiji, ki nas skozi pregled znanstvenofantastičnega žanra lahko pelje v utopijo, znatno pogosteje pa v distopijo. A kaj nam o razvoju te umetno ustvarjene moči procesiranja informacij, zmožnosti učenja, učinkovitega reševanja problemov in posledično tudi učinkovitosti v resničnem svetu lahko pove mnoštvo kvantov vsebine, ki se skrivajo na svetovnem spletu? Kaj nam digitalni odtisi pravzaprav razkrivajo? Kaj nam povedo o nas samih in, morda, naši prihodnosti?

3 Analize digitalnih odtisov

Četudi se na prvi pogled morda zdi, da so digitalni odtisi - kot so naši všečki na Facebooku, iskalni nizi na Googlu in izbor pesmi na Spotifyju - le nujen, a ne posebej pomenljiv, stranski učinek uporabe spleta, vrsta preteklih raziskav ugotavlja, da tovrstne sledi razkrivajo precej več, kot bi morda sprva intuitivno predpostavljali.

Ker zbiranje in analiziranje digitalnih odtisov tako rekoč ni bilo mogoče vse do nedavnih premikov na področju rudarjenja podatkov, dela z velikim podatkovjem, strojnega učenja in umetne inteligence, tudi najstarejše raziskave na tem področju v resnici ne segajo posebej daleč v preteklost. Kot ključno prelomno delo se pogosto navaja raziskava Kosinskega in sodelavcev (2013), v kateri je več kot 58.000 udeležencev izpolnilo nekaj demografskih vprašanj, psiholoških vprašalnikov in testov, hkrati pa so raziskovalcem dovolili dostop do obsežnega arhiva lastnih Facebook všečkov. Po logiki regresije so nato avtorji poskusili na podlagi Facebook všečkov napovedati različne odvisne spremenljivke, merjene s samoporočanjem. Ugotovili so, da je iz na videz trivialnih podatkov mogoče relativno natančno napovedati zelo širok spekter lastnosti, vključno s spolno usmerjenostjo, etično pripadnostjo, religioznostjo, politično orientiranostjo, osebnostjo, inteligentnostjo in psihičnim blagostanjem. Podroben ogled posameznih Facebook všečkov z največjo napovedno močjo razkriva nekaj precej intuitivno pomenljivih všečkov (všeček 'I love Jesus' je npr. visoko značilen za kristjane), nikakor pa to ne velja za vse (eden najboljših napovednikov visoke inteligentnosti je npr. všeček 'Curly fries'). Dodatno velja izpostaviti, da lahko o posameznikih nekaj izvemo že samo na podlagi peščice Facebook všečkov, vendar je – še posebej pri napovedovanju kompleksnejših

lastnosti (npr. osebnosti) – za natančnejše napovedi praviloma treba analizirati zajeten nabor Facebook všečkov hkrati. Kadar so v omenjeni raziskavi algoritmi upoštevali celoten nabor všečkov udeležencev, so bile korelacije med napovedanimi osebnostnimi lastnostmi po modelu velikih pet in samoocenami teh lastnosti zmerne in pozitivne (Kosinski et al. 2013).

Ker je do zgoraj opisane študije Kosinskega in drugih (2013) med raziskovalci in laiki obstajalo trdno zasidrano prepričanje, da so za natančno oceno (osebnostnih) lastnosti posameznikov nujne socialnokognitivne sposobnosti človeških možganov, so rezultati raziskave naleteli na precejšnje dvome glede primerjalne natančnosti ocen, ki jih podajo algoritmi, v primerjavi z ocenami osebnosti, ki bi jih o neki osebi podali drugi. Tako je enak krog raziskovalcev (Youyou et al. 2015) v eni od nadaljnjih študij primerjal točnost ocen osebnosti med računalniki in ljudmi, pri čemer so kot mero točnosti upoštevali korelacijo med napovedmi algoritmov/ljudi in dejansko samooceno osebe, katere osebnost so algoritmi/ljudje skušali napovedati. Rezultati so presenetljivo pokazali, da so ocene osebnosti, ki jih lahko pridobimo iz digitalnih odtisov, v povprečju bolj veljavne od ocen, ki jih o osebi podajo bližnje osebe ali znanci. Točnost algoritmov je sicer bila tudi v tej raziskavi odvisna od števila všečkov, ki jih ima algoritem na voljo pri napovedovanju osebnosti. Tako naj bi v primeru napovedovanja osebnostih lastnosti po modelu velikih pet (odprtost, vestnost, ekstravertnost, sprejemljivost in nevroticizem; McCrae in Costa 1999) v povprečju že 10 všečkov zadoščalo, da ocene algoritmov postanejo natančnejše od ocen sodelavcev. Če ima algoritem na voljo 70 Facebook všečkov dane osebe, lahko pri napovedovanju osebnosti prekaša ocene prijateljev in cimrov, približno 150 všečkov pa zadošča, da algoritem po točnosti prekaša ocene družine. Na drugi strani je za to, da algoritem preseže točnost ocen partnerjev, potrebnih približno 300 všečkov. Glede na to, da naj bi povprečna oseba imela na Facebooku všečkanih približno 220 strani, so algoritmi v večini primerov torej uspešnejši od povprečnega človeškega ocenjevalca (edina izjema so partnerji oseb). Korelacija med ocenami algoritmov in samooceno oseb sicer za osebnostne lastnosti po modelu velikih pet v povprečju znaša $r = 0{,}56$ (za primerjavo - povprečna korelacija med oceno partnerjev in samooceno oseb znaša le rahlo več, in sicer $r = 0{,}58$; Youyou et al. 2015).

Vzporedno z razvijanjem algoritmov, ki lahko osebnost napovejo na podlagi Facebook všečkov, so raziskovalci (nekateri so enaki kot zgoraj) delali tudi na razvijanju algoritmov, ki lahko različne lastnosti oseb napovejo na podlagi analize besedila, dostopnega na Facebooku (npr. posodobitve statusov) in drugih spletnih socialnih omrežjih. V prvi tovrstni raziskavi so avtorji poročali o številnih tekstovnih elementih, ki so povezani z osebnostnimi lastnostmi po modelu velikih pet; bolj ekstravertirane osebe, npr. v primerjavi z manj ekstravertiranimi, pogosteje uporabljajo prvo osebo množine, drugo osebo, vsebinsko pa njihovi zapisi pogosteje omenjajo socialne procese, družino, prijatelje in ljudi nasploh (Schwartz et al. 2013). V nadaljnjih študijah so avtorji ta spoznanja poskusili nadgraditi tako, da so z algoritmi poskusili oceniti osebnost uporabnikov zgolj na podlagi njihovih zapisov na Facebooku. Z algoritmi napovedane vrednosti so za vse osebnosti lastnosti po modelu velikih pet, podobno kot v primeru Facebook všečkov, bile zmerno pozitivno povezane s samoocenjeno osebnostjo uporabnikov (Park et al. 2015). Analiziranje besedila na spletnih socialnih omrežjih je v letih za tem doživelo pravcat razcvet; med drugim so avtorji ugotovili, da je mogoče iz besedil na Facebooku relativno natančno prepoznati celo osebe z depresijo (Eichstaedt et al. 2018).

Kljub temu da na področju analize digitalnih odtisov še danes prevladujejo raziskave, ki osebnost in druge lastnosti oseb napovedujejo predvsem na podlagi Facebook všečkov in/ali besedil, to vsekakor nista edina vira informacij, ki so dostopne na spletu in lahko razkrivajo lastnosti uporabnikov. V zadnjih letih tako prihaja do velikih premikov tudi na področju algoritmov, ki procesirajo slike (Eftekhar et al. 2014; Wang in Kosinski 2018). Kontroverzna raziskava Wanga in Kosinskega (2018) je denimo razkrila, da so algoritmi uspešno ločevali med homoseksualnimi in heteroseksualnimi moškimi v 81 % primerov, med homoseksualnimi in heteroseksualnimi ženskami pa v 71 % primerov, kar je oboje bistveno višje od natančnosti človeških ocenjevalcev (61 % za prepoznavanje moških in 54 % za prepoznavanje žensk). Dodatno so se tudi raziskave algoritmov, ki procesirajo slike, v preteklosti že posvetile napovedovanju osebnosti po modelu velikih pet. Tovrstne raziskave med drugim kažejo, da je iz lastnosti fotografij in s fotografijami povezanih podatkov (npr. števila fotografij) mogoče relativno točno napovedati nekatere osebnostne lastnosti, predvsem ekstravertnost in nevroticizem, pri katerih so algoritmi celo uspešnejši od povprečnega človeškega ocenjevalca (Eftekhar et al. 2014). Med drugimi viri informacij naj izpostavimo zgolj še lokacijske informacije (npr. Zhong et al. 2015; Matz in Harari 2020), glasbene preference (npr. Nave et al.

2018) in širok nabor senzoričnih podatkov, zbranih z našimi mobilnimi napravami (npr. pospeški, srčni utrip; LiKamWa et al. 2013), ki prav tako razkrivajo precej več o uporabnikih, kot bi si morda sprva mislili.

Pregled informacij v tem poglavju, ki sicer ni celovit, nedvomno razkriva, da je digitalnih odtisov, ki dajejo vpogled v človeške lastnosti, vključno z osebnostjo, izjemno veliko in da so ti lahko zelo raznoliki. Facebook všečki, odprti zapisi na spletnih socialnih omrežjih in obiskane lokacije, ki se skrbno beležijo na naših profilih in napravah, so vsi že sami zase lahko precej pomenljivi in nudijo določen vpogled v individualne lastnosti uporabnikov, vsekakor pa si v prihodnosti lahko obetamo, da bo mogoče omenjene odtise še integrirati v celoto in tako še izboljšati natančnost napovedi.

4 Kako lahko uporabimo informacije, pridobljene na podlagi analize digitalnih odtisov?

Informacije o uporabnikih, ki jih lahko izpeljemo na podlagi analize raznolikih digitalnih odtisov, so podlaga boljšemu razumevanju potreb in lastnosti uporabnikov, kar se lahko v nadaljnjih korakih uporabi za naslavljanje različnih specifičnih ciljev, ki pa navadno vsebujejo element prilagajanja storitev in vsebin. Na eni strani lahko to vodi do prilagajanja storitev, ki v splošnem povečujejo zadovoljstvo uporabnikov in se večini njih ne zdi prav nič sporno (oz. je celo zaznano kot dobrodošlo); po besedah direktorja podjetja Spotify, svetovno znani band Metallica denimo nabor pesmi na koncertih prilagaja preferencam prebivalcev tistega mesta, v katerem se odvija koncert. Z drugimi besedami – iz digitalnih odtisov na Spotifyju lahko glasbena skupina razbere, katere pesmi so najbolj priljubljene na določeni lokaciji, nabor pesmi na koncertu pa potem prilagodijo na tak način, da resnično vsebuje najbolj priljubljene pesmi v danem kraju (Rodriguez 2018).

Na drugi strani so se podobne metode pridobivanja in analize digitalnih odtisov v preteklosti tudi že zlorabljale za načrtno manipuliranje z javnostjo. Pri tem ne moremo mimo podjetja Cambridge Analytica, ki je tekom predsedniških volitev v ZDA (leta 2016) brez privolitve uporabnikov izdelalo podrobne osebnostne profile vsaj 30 milijonov uporabnikov Facebooka (po nekaterih podatkih so te številke še precej višje), te informacije pa je nato uporabljalo za politično oglaševanje v dobrobit Teda Cruza in Donalda Trumpa. Konkretno je to vključevalo denimo prikazovanje

raznolikih prilagojenih sporočil, pri čemer so denimo volivci, ki jih je algoritem označil kot neodločene, na spletnih socialnih omrežjih videvali predvsem objave o njunih podpornikih ter izrazito negativne objave o njunih nasprotnikih (Associated Press 2018; Lewis in Hilder 2018).

Medtem ko gre v zgornjih primerih za dva specifična vidika praktične uporabe digitalnih odtisov z nejasno učinkovitostjo, se je v akademskih krogih pojavila predvsem ideja o psihološkem ciljanju (izvorno: *psychological targeting*), ki je v zadnjih letih bila podvržena tudi rigoroznemu empiričnemu preverjanju (so pa raziskave za zdaj redke; Matz et al. 2017). Psihološko ciljanje sloni na temeljih ocenjevanja osebnosti iz digitalnih odtisov (npr. Kosinski et al. 2013) in personalizacije oz., konkretneje, prilagajanja (oglasnih) sporočil po modelu velikih pet (npr. Hirsh et al. 2012), njegov splošni cilj pa je vplivati na stališča, čustva ali vedenja uporabnikov (Matz et al. 2020).

Sandra Matz in kolegi (2017) so skozi serijo treh eksperimentov v naravnem okolju (tj. na Facebooku) preverjali učinke psihološkega ciljanja na vedenje ljudi, merjeno s kliki na oglas in številom prenosov oz. spletnih nakupov. Vsi oglasi so bili distribuirani s pomočjo Facebook oglaševanja, ki sicer ne dovoljuje direktnega ciljanja ljudi na podlagi osebnostnih lastnosti, dovoljuje pa to posredno preko možnosti ciljanja ljudi na podlagi Facebook všečkov. Primer: če v skladu s preteklimi raziskavami vemo, da imajo skupino 'Socializing' (druženje) všečkano predvsem ekstravertirani, skupino 'Stargate' (Zvezdna vrata) pa introvertirani posamezniki, lahko ciljanje oseb, ki imajo všečkano eno od teh dveh strani, omogoči dostopanje do ekstravertiranih in introvertiranih oseb. Izmed širokega nabora osebnostnih lastnosti, ki jih je mogoče napovedati na podlagi Facebook všečkov, so se avtorji osredotočili le na ekstravertnost in odprtost (gre za dimenziji, pri katerih je natančnost algoritmov največja), tako pa so torej oblikovali oglase, ki so izžarevali visoko raven ekstravertnosti ali odprtosti in oglase, ki so izžarevali nizko raven ekstravertnosti ali odprtosti. Pripravljene oglase so nato ciljno usmerjali na visoko/nizko ekstravertirane/odprte osebe (preko usmerjanja oglaševanja na ljudi, ki imajo všečkane določene Facebook strani). Rezultati so pokazali, da so oglasi, ki so bili prilagojeni posameznikovi ekstravertnosti ali odprtosti, vodili do višjega števila klikov in nakupov v primerjavi z oglasi, ki niso bili skladni s posameznikovo ekstravertnostjo ali odprtostjo (Matz et al. 2017).

Kljub temu da so principi psihološko ciljanega oglaševanja v resnici tudi v jedru škandala Cambridge Analytica, ki smo ga že omenili, avtorji poudarjajo, da se je potrebno odmakniti od splošnih sodb o tehnologiji in diskutirati o tem, *kdo* zbira podatke, *kakšni* podatki se zbirajo, predvsem pa o tem, *kako so ti podatki uporabljeni* oz. kaj je namen v ozadju uporabe tovrstnih strategij. Kot pomemben pozitivni aspekt uporabe psihološkega ciljanja avtorji tako izpostavljajo npr. intervencije na področju komunikacije zdravstvenih informacij; raziskave, izvedene v drugih kontekstih, namreč kažejo, da so javnozdravstvene intervencije najbolj učinkovite, ko so prilagojene posameznikovim lastnostim (npr. Rimer in Kreuter 2006). Medtem ko je v tradicionalnih kontekstih prilagajanje tovrstnih intervencij lahko sila velik zalogaj (z vidika časa in truda), novi načini zbiranja, analize in uporabe digitalnih odtisov odpirajo do nedavnega nepredstavljive priložnosti in vodijo do tega, da je celoten proces lahko tako enostaven, hiter in poceni kot še nikoli prej (Matz et al. 2020).

5 Sledenje v neznano?

Skozi prikazan nabor primerov uporabe digitalnih odtisov v perspektivi psihologije lahko zaključimo, da navidezno drobna digitalna dejanja, ki jih uporabniki spleta delamo mnogokrat na meji ali onkraj ozaveščenega, vsekakor pa prostovoljno, analitično že prehajajo iz vprašanj osnovnih povezav in napovednih modelov v sfero (ne)kontrolirane uporabe, ki sproža spremembe pri uporabnikih. Pri tem te spremembe lahko opažamo na vseh nivojih, od posameznikov, manjših skupin, skupnosti do celotne družbe. Kot je to na primeru spletnih socialnih omrežij in pripadajoče informacijske tehnologije slikovito prikazal dokumentarec *Socialna dilema* (v originalu *The Social Dilemma*; Orlowski 2020), lahko govorimo o porastu novodobnih zasvojenosti z uporabo tehnologije in vplivu le-teh na druge vidike duševnega zdravja, razrastu količine in kroženja nepreverjenih informacij, neresnic in teorij zarot, posledično manipulaciji ljudi in vplivanju na raznolike oblike družbenega življenja, od ekonomije do politike. Če je spletna tehnologija v prvem obdobju odpirala vprašanja, kaj iz naših resničnih življenj se (lahko) seli na digitalno, smo sedaj v fazi, ko z zamikom, morda tudi nemo, opazujemo, kako digitalno prevzema več in več našega življenja in kako se pojavi iz digitalnega selijo v resnično. Ko npr. prekomerna toksičnost digitalnega prostora dobi svoj korelat v nasilnih dogodkih našega fizičnega sveta.

Stežka sicer zaključujemo, da prehajamo v fazo, ko se fikcija v uvodu izpostavljenega filma uresničuje – da se vsa nakopičena negativna plat naše nravi naprej sublimno prenaša v tehnologijo, ki bo v nadaljevanju povzročila propad civilizacije. Prej lahko rečemo, da smo znova ali še vedno pred izzivi, ki so se porajali prvim snovalcem spleta – da je digitalno prostor popolne svobode, kjer si ideje, vizije, mnenja, prepričanja, znanja, vednosti, veščine izmenjujejo posamezniki z visoko moralno integriteto in kompetentnostjo na področjih uporabe tehnologije. Kjer govorimo o samoregulaciji uporabnikov in posledično ni potrebe po zunanji regulaciji ali nadzoru. Realistično se lahko strinjamo, da če ima naš resnični svet številne hibe, digitalna resničnost ne predstavlja nikakršne popolne verzije sveta.

Viri in literatura

Associated Press. (2018). »Facebook to send Cambridge Analytica data-use notices to 87 million users Monday«. *NBC News*. URL = https://www.nbcnews.com/tech/social-media/facebook-send-cambridge-analytica-data-use-notices-monday-n863811.

Cattell, R. B. (1963). »Theory of fluid and crystallized intelligence: A critical experiment«. *Journal of Educational Psychology, 54*, str. 1–22. https://doi.org/10.1037/h0046743.

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action*. Boston: Houghton Mifflin.

Chabris, C. F. (2007). »Cognitive and neurobiological mechanisms of the Law of General Intelligence«. V M. J. Roberts (ur.), *Integrating the mind: Domain general vs domain specific processes in higher cognition*. Psychology Press, str. 449–491.

Eftekhar, A., Fullwood, C. in Morris, N. (2014). »Capturing personality from Facebook photos and photo-related activities: How much exposure do you need?«. *Computers in Human Behavior*, 37, str. 162–170. https://doi.org/10.1016/j.chb.2014.04.048.

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., ... in Schwartz, H. A. (2018). »Facebook language predicts depression in medical records«. *Proceedings of the National Academy of Sciences*, 115(44), str. 11203–11208. https://doi.org/10.1073/pnas.1802331115.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York: Basic Books.

Gottfredson, L. S. (2002). »G: Highly general and highly practical«. V R. J. Sternberg in E. L. Grigorenko (urd.), *The general factor of intelligence: How general is it?*. Lawrence Erlbaum, str. 331–380.

Harlow, L. L. in Oswald, F. L. (2016). »Big data in psychology: Introduction to the special issue«. *Psychological Methods*, 21(4), str. 447–457. https://doi.org/10.1037/met0000120.

Hirsh, J. B., Kang, S. K. in Bodenhausen, G. V. (2012). »Personalized persuasion: Tailoring persuasive appeals to recipients' personality traits«. *Psychological Science*, *23*(6), str. 578–581. https://doi.org/10.1177/0956797611436349.

Kosinski, M., Stillwell, D. in Graepel, T. (2013). »Private traits and attributes are predictable from digital records of human behavior«. *Proceedings of the National Academy of Sciences*, 110(15), str. 5802–5805. https://doi.org/10.1073/pnas.1218772110.

Lewis, P. in Hilder, P. (2018). »Leaked: Cambridge Analytica's blueprint for Trump victory«. *The Guardian*. URL = https://www.theguardian.com/uk-news/2018/mar/23/leaked-cambridge-analyticas-blueprint-for-trump-victory.

LiKamWa, R., Liu, Y., Lane, N. D. in Zhong, L. (2013). »Moodscope: Building a mood sensor from smartphone usage patterns«. V *Proceeding of the 11th annual international conference on Mobile systems,*

applications, and services. New York: Association for Computing Machinery, str. 389–402. https://doi.org/10.1145/2462456.2464449.

Matz, S. C., Appel, R. E. in Kosinski, M. (2020). »Privacy in the age of psychological targeting«. *Current Opinion in Psychology*, 31, str. 116–121. https://doi.org/10.1016/j.copsyc.2019.08.010.

Matz, S. C. in Harari, G. M. (2020). »Personality–place transactions: Mapping the relationships between Big Five personality traits, states, and daily places«. *Journal of Personality and Social Psychology*, 120(5), str. 1367–1385. https://doi.org/10.1037/pspp0000297.

Matz, S. C., Kosinski, M., Nave, G. in Stillwell, D. J. (2017). »Psychological targeting as an effective approach to digital mass persuasion«. *Proceedings of the National Academy of Sciences,* 114(48), str. 12714–12719. https://doi.org/10.1073/pnas.1710966114.

McCrae, R. R. in Costa, P. T. (1999). »A five-factor theory of personality«. V L. A. Pervin in O. P. John (urd.), *Handbook of personality: Theory and research*. New York: Guilford, str. 139–153.

Nave, G., Minxha, J., Greenberg, D. M., Kosinski, M., Stillwell, D. in Rentfrow, J. (2018). »Musical preferences predict personality: Evidence from active listening and facebook likes«. *Psychological Science*, 29(7), str. 1145–1158. https://doi.org/10.1177/0956797618761659.

Orlowski, J. (režiser). (2020). *The Social Dilemma* [dokumentarni film]. Netflix.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., ... in Seligman, M. E. (2015). »Automatic personality assessment through social media language«. *Journal of Personality and Social Psychology*, 108(6), str. 934–953. http://dx.doi.org/10.1037/pspp0000020.

Reader, S. M., Hager, Y. in Laland, K. N. (2011). »The evolution of primate general and cultural intelligence«. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1567), str. 1017–1027. https://doi.org/10.1098/rstb.2010.0342.

Rimer, B. K. in Kreuter, M. W. (2006). »Advancing tailored health communication: A persuasion and message effects perspective«. *Journal of Communication,* 56, str. 184–201. https://doi.org/10.1111/j.1460-2466.2006.00289.x.

Rodriguez, A. (2018). »Metallica shapes its live shows around what fans are listening to on Spotify«. *Quartz*. URL= https://qz.com/1340887/metallica-bases-its-setlist-on-what-fans-listen-to-on-spotify/.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... in Ungar, L. H. (2013). »Personality, gender, and age in the language of social media: The open-vocabulary approach«. *PloS one*, 8(9), e73791. https://doi.org/10.1371/journal.pone.0073791.

Spearman, C. (1904). »'General intelligence,' objectively determined and measured«. *American Journal of Psychology*, 15, str. 201–293. https://doi.org/10.1037/11491-006.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of intelligence*. Cambridge: Cambridge University Press.

Wang, Y. in Kosinski, M. (2018). »Deep neural networks are more accurate than humans at detecting sexual orientation from facial images«. *Journal of personality and social psychology*, 114(2), str. 246–257. https://doi.org/10.1037/pspa0000098.

Wilcox, F. M. (režiser). (1956). *Forbidden planet* [film]. MGM Studios.

Youyou, W., Kosinski, M. in Stillwell, D. (2015). »Computer-based personality judgments are more accurate than those made by humans«. P*roceedings of the National Academy of Sciences,* 112(4), str. 1036–1040. https://doi.org/10.1073/pnas.1418680112.

Zhong, Y., Yuan, N. J., Zhong, W., Zhang, F. in Xie, X. (2015). »You are where you go: Inferring demographic attributes from location check-ins«. V *Proceedings of the eighth ACM international conference on web search and data mining*. New York: Association for Computing Machinery, str. 295–304. https://doi.org/10.1145/2684822.2685287.

SOCIALNI ODNOSI, ANIMIZEM, ANTROPOMORFIZEM IN INTERAKCIJA Z UI

NENAD ČUŠ BABIČ
Univerza v Mariboru, Fakulteta za gradbeništvo, prometno inženirstvo in arhitekturo, Maribor, Slovenija
nenad.babic@um.si.

Sinopsis Eden od načinov interpretacije sistemov UI in s tem načinov sprejemanja, odzivanja in vstopanja v odnos z UI je, da jih človek dojema kot nekaj živega, animistično, oziroma kot nekaj, kar deluje po svoji volji, podobno, kot to počnemo ljudje, antropomorfno. V tem primeru UI pripisujemo lastnosti živih bitij, kot so sposobnost čutenja, razumevanja in namernega delovanja. Seveda se lahko vprašamo, ali je to tudi zares mogoče, vendar to vprašanje presega obseg tega prispevka. V članku se bom osredotočil na psihološke vidike vstopanja v odnose z UI skozi antropomorfno interpretacijo. Skozi pregled psiholoških raziskav na področju preučevanja animizma ter atribucij in teorij uma v povezavi s tehnološkimi sistemi in UI bom predstavil psihološko razumevanje namena, funkcije in sprožilcev antropomorfnih atribucij. Ukvarjal se bom s spoznanji o vplivih in posledicah tovrstne interpretacije UI na zaznavanje, čustvovanje, mišljenje ter delovanje ljudi. Po drugi strani je človek aktiven udeleženec v odnosu do tehnologije. Ker pa človek v interakcijo z UI vstopa na sebi lasten način, to je na osnovi grajenja socialnih odnosov, saj se z napravami pogovarja in jih poimenuje, me bodo zanimala tudi vprašanja v zvezi z motivi, potrebami ter pomeni, ki jih uporabniki v odnosu do UI razvijajo.

Ključne besede:
tehnološki animizem, antropomorfno oblikovanje, interpretacija UI, socialna interakcija, interakcija človek-stroj

Univerzitetna založba Univerze v Mariboru

DOI https://doi.org/10.18690/um.ff.11.2022.12
ISBN 978-961-286-675-4

SOCIAL RELATIONS, ANIMISM, ANTHROPOMORPHISM, AND INTERACTION WITH AI

NENAD ČUŠ BABIČ
University of Maribor, Faculty of Civil Engineering, Transportation Engineering and Architecture, Maribor, Slovenia
nenad.babic@um.si

Abstract One interpretation of AI systems and how we accept, respond, and establish a relation with AI is for us to perceive them as something alive, i.e., as something that functions on its own will and is similar to how humans act, i.e., anthropomorphically. We thus ascribe characteristics of living beings, like the ability to feel, understand, and act intentionally, to machines and AI mechanisms. However, in this paper, I will focus on psychological aspects of approaching the relations with AI through an anthropomorphic interpretation. By reviewing psychological research in the field of animism and attributions and theory of mind in relation to technological systems and AI, I will present the psychological understanding of the purpose, function, and triggers of anthropomorphic attributions. I will deal with impacts and consequences of such interpretation of AI on perception, emotion, reasoning, and actions of people. On the other hand, humans can be active participants in relation to technology. Since humans interact with AI in a unique way, i.e., by building social relations, talking to the devices and naming them, I will also be interested in motives, needs, and meanings that users develop in their relations with AI.

Keywords:
technological animism, anthropomorphic design, AI interpretation, social interaction, human-computer interaction

https://doi.org/10.18690/um.ff.11.2022.12 DOI
978-961-286-675-4 ISBN
University of Maribor Press

1 Animizem in antropomorfno mišljenje

Dojemanje stvari kot živih in pripisovanje duševnih stanj, emocij ter pripisovanje namernega ravnanja živalim, rastlinam, neživim stvarem in pojavom je zaradi svoje razširjenosti že dolgo predmet znanstvenega preučevanja. Pozornost temu vprašanju še posebej posvečajo študije na področju antropologije, in sicer že od konca 19. stoletja pa vse do danes. Pri tem se antropologi predvsem ukvarjajo z vprašanjem prepričanja, da v stvarnih telesih ali predmetih obstajajo tudi nematerialne duše, karkoli že to pomeni (pregled v Bird-David 1999). Po večini se prepričanja nanašajo na to, da duh ali duša predmetom, pojavom ali organizmom vlije življenje, od koder izhaja tudi izraz animizem. Kot ugotavlja Richardson (2016) v razlagah pojava prevladuje dualistični pogled na svet (duša-telo, stvarno-nestvarno, človek-okolje). Že antropologi 19. stoletja so označili pripisovanje duše stvarem kot znak otroškosti, kognitivne nerazvitosti in kot napako v mišljenju (Bird-David 1999). Z drugimi besedami naj bi pri animizmu šlo za nezavedno projekcijo 'človeških' lastnosti v okolje in s tem za napako. Vendar se lahko vprašamo, ali je gre res za napako ali odraz kognitivne nerazvitosti? Takšno pojmovanje je težavno vsaj iz dveh razlogov. Prvič zato, ker odpira vprašanje, kaj sploh je človeško, in drugič zato, ker je pojav kljub vsem znanstvenim spoznanjem še vedno močno prisoten tako pri otrocih kot pri odraslih, ne glede na izobraženost. Bird-David prekine s to dolgo tradicijo antropoloških študij in razlag, ki ohranjajo dualistično pojmovanje človeka kot osnovo animizma in pokaže drugačen pogled. Animizem razlaga kot relacijski fenomen. Pri tem izhaja iz predpostavke o socialni naravnanosti človeške kognicije. Ta pogled bom tudi sam uporabil kot izhodišče in rdečo nit tega prispevka, vendar v nekoliko drugačnem kontekstu. V nadaljevanju se bom osredotočil predvsem na psihološki pogled na animizem, in sicer v kontekstu človeka v odnosu do tehnologije.

Ob pojmu animizem se v podobnih kontekstih uporablja tudi pojem antropomorfizem. Včasih se pojma uporabljata tudi kot sopomenki, še posebej v obliki poosebljanja narave, po večini pa vsaj kot sočasna pojava. Pri pojmovanju animizma se bom v nadaljevanju izogibal uporabi razlage s pojmi, kot sta duh ali duša, predvsem zato, ker ta pojma nimata enoznačnega pomena. Poleg tega v ospredje postavljata dualistični razcep telo-duša in s tem vpeljeta nekatere nejasnosti. Zastavlja se na primer vprašanje možnosti ločenega obstoja duha in telesa. V izogib tem zapletom bom z izrazom animizem poimenoval pripisovanje namernosti v dejanjih in/ali pripisovanje živosti predmetom ali pojavom. Podobno je animizem

kot »psihološki animizem« opredelil že Read (1915). Za pripisovanje zavestnega vedenja, čustvenih stanj in zmožnosti mišljenja pa bom uporabil pojem antropomorfizem.

Preprostega odgovora na vprašanje, zakaj človek uporablja animistične in antropomorfne razlage, ne poznamo. Na temelju številnih raziskav pa vemo, da se takšno mišljenje zelo hitro sproži. To sta z znanim eksperimentom pokazala že Heider in Simmel (1944) s preprosto animacijo, v kateri nastopajo abstraktni dvodimenzionalni liki, ki navidezno vstopajo v medsebojne interakcije. Že minimalne sugestije namernega vedenja v obliki gibanja likov, ki se navidezno odzivajo eden na drugega, uspešno sprožijo antropomorfno dojemanje situacije. Podobno tudi oblike, ki spominjajo na oči, roke ali obraz, sprožajo pripisovanje zavestnih duševnih stanj stvarem in upodobitvam (Blakemore in Decety 2001).

2 Psihološki animizem je staro vprašanje

Ne le v antropologiji, tudi v psihologiji je animizem staro vprašanje, ki vedno znova odpira nove poglede na iste dileme. Read (1915) je že leta 1915 pisal o psihološkem animizmu in se kritično opredeljeval do takratnih zaključkov o tem pojavu. Animizem kot lastnost »zmedenih divjakov« in nezrelih otrok je postavljal pod vprašaj z navajanjem primerov izobraženih odraslih ljudi, ki se prav tako do stvari obnašajo, kot da jim namerno škodujejo ali so jim namerno v pomoč. Izpostavil je, da so animistična pripisovanja otrok dejansko sugerirana od odraslih z namenom odvračanja pozornosti. Otroku, ki se udari, odrasli odvrnejo pozornost od bolečine tako, da ga usmerijo k povračilnemu ukrepu, da mizo udari nazaj. Prav tako navaja primere otrok in 'divjakov', da je animistično pripisovanje le začasno in je stvar domišljijskega sveta posameznika. Vse to se odraža v nekonsistentnem vedenju do stvari, saj enkrat predmete obravnavamo animistično, spet drugič pa jih uporabljamo le kot predmete glede na njihovo funkcijo. Animistično obravnavo posameznik uporablja le z določenim namenom, kadar s tem nekaj pridobi. Read o psihološkem animizmu tako sklene, da:

Otroci, divjaki in do neke mere tudi mi sami neživim stvarem spontano pripisujemo nekaj več kot le zunanjo eksistenco; obravnavamo jih, kot da imajo moč in empatijo, kot da doživljajo napore, umirjenost, napetost in olajšanje, včasih pa tudi čustva in bolečino. (Read 1915: 4)

Read našteje tudi nekaj razlogov zaradi katerih človek stvari obravnava animistično. Takšni primeri so, če se stvari premikajo ali delujejo spontano ali so posebej nevarne, če jih kultura naredi takšne (npr. s pripisovanjem magičnih moči ali z uporabo v obredih), ali pa je animizem posledica domišljije.

3 Animizem in teorija kognitivnega razvoja

Če pogledamo na animizem in antropomorfna dojemanja sveta iz vidika kognicije, vsekakor ne moremo mimo Piagetove stopenjske teorije kognitivnega razvoja (Piaget 2013). Glede na zastavljene razvojne stopnje lahko razumemo, da naj bi najvišja stopnja kognitivnega razvoja bila zmožnost abstraktnega in dekontekstualiziranega mišljenja. Posameznik na najvišji stopnji, to je stopnji formalno-logičnega mišljenja, je sposoben uporabe simbolov in logičnih operacij nad abstraktnimi koncepti. Stopenjska hierarhija na impliciten način vzpostavlja tudi navidezno superiornost formalno logičnega mišljenja kot najvišjega dosežka in s tem prednostnega načina delovanja. Gledano iz te perspektive je animistični pogled na svet neživih stvari kognitivna zmota ali znak kognitivne nerazvitosti. S svojo zastavitvijo animizma Piaget, sicer iz drugega zornega kota, ponovno odpira temo animizma kot principa, ki je lasten otrokom in kognitivno manj razvitim odraslim (v terminologiji zgodnje antropologije — divjakom). Po drugi strani Caporeal (1997) tovrstno razlago animizma in antropomorfizma poimenuje »kognitivna privzeta vrednost« (angl. *cognitive default*). Če torej za razlago in predvidevanje dogajanja v okolju nimamo ustrezne druge razlage, pojave privzeto razlagamo na antropomorfni način. Prošnje računalniku za pravilno rešitev problema in avtomobilu, kadar ne gre zagnati motorja, so podobne prošnjam nadnaravnim silam, darovanju vulkanom ali plesom za dobro letino. Otroci bi po tej teoriji z izkušnjami in znanjem postopoma opuščali animizem.

Piaget je s svojo teorijo spodbudil številne študije, ki skušajo preveriti njegove hipoteze. Vendar študije na različnih populacijah odraslih ljudi nasprotno kažejo, da animistično mišljenje ohranjajo tudi odrasli, inteligentni in izobraženi ljudje. Srednješolci, ne glede na to, da so bili uspešni pri biologiji, uporabljajo animistične razlage v relaciji do neživih stvari (Crowell in Dole 1957). Brown in Thoules (1965) pokažeta, da tudi študentje povsem po lastni izbiri uporabljajo animistično izražanje, kadar o neživi naravi govorijo kot o živi. Številne študije (pregled v Dacey 2017) pokažejo, da ljudje različnih starosti in izobrazbene ravni, od otrok do znanstvenikov, izkazujejo tako napake kot namerno antropomorfno pripisovanje na

osnovi številnih zunanjih spodbud. Pri odraslih ljudeh je uporaba animističnega izražanja lahko odraz lastne izbire in ne nezmožnosti prepoznati živo od neživega glede na kriterije biologije. Študija z roboti (Okanda, Taniguchi in Itakura 2019) je pokazala, da posamezniki, ki na vprašanje, ali robot lahko umre odgovorijo z da, ne bodo nujno pritrdilno odgovori na vprašanje, ali je robot živ. Takšno vedenje kaže na to, da ne gre za zmoto, temveč za izbran način izražanja. Navedene študije so še pokazale, da se verjetnost izbire animističnega pripisovanja običajno poveča v pogojih nepredvidljivosti v delovanju opazovanih stvari ter avtonomnosti v gibanju in vedenju (Brown in Thouless 1965).

Iz vsega navedenega vidimo, da zgodnja Readova sklepanja o animizmu kasnejše študije potrjujejo. Pojav je prisoten v vseh segmentih populacije, je začasen in se pojavlja v določenih kontekstih. Pojav ni nujna iluzija, temveč je spodbujen z določenimi pridobitvami, ki posameznika motivirajo k takšnemu početju.

Iz kognitivne perspektive lahko skušamo poiskati tudi namen animističnega dojemanja sveta na primeru informacijske tehnologije. Čeprav so določene predpostavke Piagetove teorije kasnejše raziskave ovrgle, kot npr. predpostavko o samodejnem biološkem doseganju stopnje formalno-logičnega mišljenja z odraščanjem (McDonald in Stuart-Hamilton 2000), pa prav dokazi, da tudi odrasli ljudje uporabljamo mišljenje na nivoju konkretnih operacij (glej Sutherland 1999), Piagetovo teoretično zastavitev kognitivnih zmožnosti in modalitet delovanja naredijo zanimivo tudi za razmišljanje o odnosu do artefaktov informacijske tehnologije.

V kontekstu mišljenja o računalnikih, ki predstavljajo novost glede na raziskave, ki jih je opravil Piaget, saj tega konteksta v času njegovega raziskovanja otrok ni bilo, Papert (1988) izpostavi, da je za asimilacijo matematičnih abstrakcij pravzaprav bistven korak v konkretno. Posameznik za usvajanje abstraktnega pojma posega tudi v obstoječe kognitivne sheme telesnega in zaznavno-gibalnega izkustva. To počne skozi animistično in antropomorfno mišljenje o stvareh. Le-to je spontano tako pri otrocih kot odraslih. Papert to ponazori z abstraktnim konceptom točke, ki je ob prostorskih koordinatah opisana še z usmerjenostjo. Takšen koncept si je težko predstavljati in dokler je povsem abstrakten, je z njim tudi težko operirati. Vendar že majhni otroci koncept z lahkoto razumejo in uporabljajo, če si točko predstavljajo kot želvo, ki je obrnjena v določeno smer. Na tem principu je Papert zasnoval programski jezik LOGO, ki omogoča že zelo zgodnje učenje računalniškega

programiranja, čeprav je dejavnost izrazito abstraktna. Na primeru učenja programiranja pokaže, da formalno logične operacije v smislu načrtovanja in razcepljanja problema na podprobleme niso edini način doseganja rešitev, saj je mogoče do uporabnih rešitev priti tudi s povsem drugačnimi kognitivnimi slogi kot samo z analitičnim mišljenjem. Iz vidika teorije kognitivnega razvoja tako lahko rečemo, da posameznik sicer stopenjsko razvija svoje sposobnosti, vendar vse razvite sposobnosti kasneje tudi sočasno uporablja. Sposobnosti, ki sodijo na nižje stopnje razvoja, posameznik pri svojem delovanju ne opušča kot preseženih.

Animistični princip mišljenja, kot je ugotavljal že Piaget, ima svoj namen pri otrokovem osmišljanju sveta in konstrukciji naprednejših oblik razumevanja konceptov stvari in pojavov, ki ga obdajajo. Otrok morda sicer res misli, da so nežive stvari žive, vendar je ta zmota le pripomoček za razvoj novega razumevanja. Kot pokaže Papert, takšno strategijo razvoja novega razumevanja ohranjamo tudi v odraslosti. Animistično in antropomorfno dojemanje stvari praviloma predstavljata kognitivni pripomoček in ne kognitivne zmote.

4 Antropomorfizem in socialne interakcije

4.1 Socialna motivacija

Pogled na animizem kot pripomoček zasledimo tudi na področju preučevanja socialnih interakcij. Socialna psihologija povezuje posameznikove motive s situacijo v okolju. Iz te perspektive so temeljne socialne motivacije bistvene za način interpretacije situacij. Fiske (2014) navaja pet temeljnih socialnih motivacij: pripadanje, razumevanje, nadzor, samo-izboljšanje in zaupanje. Izmed naštetih je najbolj temeljna motivacija spodbuda posameznika k razumevanju okolja, tako v smislu razumevanja tega, kar se dogaja, kot predvidevanja izida v negotovih situacijah. Posledica takšne motivacije je pripisovanje vzročnosti ter oblikovanje vedenja, ki je z razumevanjem situacije s strani posameznika skladno. V nasprotnem primeru posameznik občuti tesnobo in nelagodje. Zaznavanje in razumevanje drugih pomembno oblikuje dojemanje socialne situacije in s tem sooblikuje naše socialne interakcije. Dacey (2017) opredeli antropomorfizem kot vrsto kognitivne pristranosti in kot hevristiko, ki je rezultat evolucijskega procesa razvoja sposobnosti učinkovitejšega udejstvovanja v socialnih interakcijah. Zato trdi, da tovrstna kognitivna pristranost sicer lahko vodi v zmoto, vendar antropomorfizem sam po sebi ni zmota, temveč način mišljenja, ki je značilen za človeka. Iz navedenih

spoznanj vidimo, da posameznik način mišljenja in interpretiranja, ki ga prevladujoče uporablja v socialnih interakcijah, razširi tudi na preostali svet. Pri tem je motiviran s potrebo po razumevanju in smislu. Še posebej na področju UI, katere delovanje je včasih nedoumljivo tudi samim ustvarjalcem UI sistemov, je motiviran situacijo osmisliti. Pri poskusih razlage sveta posega po vzročnih atribucijah in posega po načelu podobnosti (Anderson 2013). V stiku z neživim svetom se zateka k animističnemu in antropomorfnemu izražanju. Posameznik v stiku s tehnologijo sicer uporablja animistično in antropomorfno izražanje, vendar ne zato, ker bi resnično verjel, da so stvari žive ali da čustvujejo, temveč zato, ker se kot takšne kažejo in si jih na ta način osmisli. V tem kontekstu je eden od pomembnih načinov iskanja pomena in smisla zdravorazumska personologija, ki vključuje pripisovanje občutij, uma in osebnosti (Fiske 2014).

4.2 Teorija uma

Pripisovanju uma je v socialni psihologiji posvečenih veliko raziskav. Epley in Waytz (2010) v pregledu raziskav povzameta, da čeprav ljudje nimamo niti vpogleda v um drugih ljudi in tako ne moremo vedeti, ali ti sploh imajo um, le-tega pripisujemo vsemu, tudi živalim, stvarem in pojavom. O mentalnih stanjih, čustvovanju, ciljih, motivih in namerah sklepamo posredno, avtomatično in brez razmišljanja ter brez eksplicitnih objektivnih meril. Sklepanje o mentalnih stanjih drugih je osrednji način sklepanja o njihovih namerah in razlagi njihovih vedenj, prav to pa naredi vedenje drugih predvidljivo ter omogoča koordinacijo skupnih dejavnosti in medsebojno razumevanje. Prisotnost ali odsotnost uma je kriterij razlikovanja med človeškim in ne človeškim. Na ta način posameznik tudi pojmuje in deluje v odnosu do ne človeškega, kar deluje podobno človeku. Prav tako ljudje zlahka odrekamo um drugim osebam in jih obravnavamo kot objekte.

Nadalje Epley in Waytz izpostavita, da tuji um ljudje opredelimo skozi dve kapaciteti, in sicer zavestno doživljanje ter namerno delovanje. Tako se um odraža kot zavestno doživljanje sebe in okolja. Namernost v dejanjih pa zajema razumsko oblikovanje in načrtovanje vedenja z določenim ciljem, kar vključuje prepričanja in znanje.

Pripisovanje uma ima seveda svoj namen. Le-tega Epley in Waytz opredelita skozi tri potrebe človeka. To so razumevanje vedenja, razumevanje komunikacije ter koordinacija z drugimi. Razumevanje omogoča razrešitev nepredvidljivosti vedenja ter s tem povezane negotovosti, ki sodi med izjemno neprijetna doživljanja. Kot

primer navedeta nepredvidljivo premikanje brez jasnega zunanjega vzroka. Da bi človek razrešil negotovost, pojav aktivnosti pojasni z željami in cilji agenta. S pripisovanjem prepričanj, znanja in stališč agenta pa skuša razumeti potek aktivnosti. Pripisovanje mentalnih stanj je preprosta in vsakomur dosegljiva aproksimacija sicer zapletene fizikalne razlage gibanja, ki jo človek potrebuje, saj ne zdrži nesmiselnosti. Drugi namen pripisovanja uma je razumevanje komunikacije. Tudi temu je namenjeno široko polje psiholoških raziskav. Medosebna komunikacija je zapletena, saj načelo racionalnosti zahteva, da komunikacija predvideva določeno obstoječe znanje na strani poslušalca, zapletenost pa še povečuje konotativna narava komunikacije, ki vključuje tudi implicitne namere. Tretji motiv za pripisovanje mentalnih stanj je medsebojna koordinacija, ki predvideva pozitivno korelacijo med mentalnim stanjem ter vedenjem. Če se vrnemo k primeru animacije Heider in Simmel (1944), je vsakomur jasno, da ni nobene potrebe, da bi zares verjeli v obstoj mentalnih stanj likov iz animacije in vendar opazovalec 'prepozna' in likom pripiše namere in mentalna stanja z namenom interpretacije opazovane situacije. Torej pri pripisovanju uma ne gre za vnaprejšnja prepričanja, temveč za interpretacijo in 'prepoznavanje'.

Zadovoljevanje navedenih potreb se odraža tudi skozi antropomorfno dojemanje sveta. Antropomorfizem navadno razlikuje dejavnosti na sodelovalne in tekmovalne oziroma na škodljive ali podporne. Predvidevanje preferenc drugih je pomembno za lastno preživetje, saj na ta način določamo, komu zaupati in komu ne. Prav tako je na ta način mogoče zadovoljevati temeljno psihološko potrebo po povezanosti. In kot ugotavlja Guthrie (1993), je evolucijsko gledano neprepoznanje intencionalnega agenta manj ugodno kot napačno prepoznanje intencionalnosti, kjer le-te ni. Zato je človek nagnjen k videnju človeških mentalnih stanj tudi v vsem drugem. To je racionalna adaptacija, ki poenoti odnos do narave in tehničnih stvari.

5 Sprožilci antropomorfizma

Antropomorfizem je odraz stika in interakcije in je v svojem bistvu socialen. Različni avtorji (Damiano in Dumouchel 2018; Epley in Waytz 2010) izpostavljajo, da se koristnost antropomorfizma kaže v medsebojni koordinaciji neodvisnih akterjev, ki se srečajo v socialni interakciji. Epley (Epley in Waytz 2010; Epley, Waytz in Cacioppo 2007) navaja številne raziskave, ki raziskujejo pozicijo moči v medosebnih interakcijah in izpostavi, da posameznik na poziciji moči v interakciji manj verjetno drugim pripisuje mentalna stanja ter jih obravnava kot nežive objekte. Pripisovanje

mentalnih stanj in emocij je bolj značilno tudi za okoliščine, v katerih je posameznik motiviran za sodelovanje, v katerih je pomembna učinkovita komunikacija ali posameznik pričakuje nadaljnje stike. Prav tako je lahko pomemben sprožilec nepredvidljivost socialnih agentov v interakciji. Med aktivnosti, ki še posebej spodbujajo antropomorfizem, pa sodi skladno gibanje (Airenti 2018).

Levillain in Zibetti (2017) na primeru robotov pokažeta, da antropomorfizem sprožata tako avtonomno gibanje in vedenje kot tudi podobnost s človekom, in sicer kot neodvisni kategoriji. Podobnost s človekom se lahko odraža v obliki, npr. okončinah, očeh, ustih, ali pa v vedenju, kot npr. usmerjanju pogleda, gestah in mimiki ter uporabi intonacije in prozodičnih znakov v govoru. Človeku po obliki zelo podobni roboti sprožajo antropomorfizem tudi, če vedenje robota ni popolnoma podobno človeškemu vedenju in obratno, kadar je vedenje človeku zelo podobno, sama oblika robota ni tako pomembna. Vendar Levillain in Zibetti opozarjata, da je v prvem primeru emocionalni izid običajno negativen, človek robota dojema na strašljiv način. V drugem primeru pa obratno, realistično vedenje kljub odsotnosti človeške oblike spodbuja všečnost in vabi v socialno interakcijo. Tudi številne druge raziskave (pregled v Giger, Piçarra, Alves-Oliveira, Oliveira in Arriaga 2019) potrjujejo, da agente, ki so bolj podobni človeku, ljudje doživljamo ogrožajoče in manj privlačno kot tiste, ki so človeku sicer manj podobni, vendar so ljudem podobni po načinu vedenja, izražanja emocij in mentalnih stanj.

6 Aplikacije animizma in antropomorfizma

Animizem in antropomorfizem preoblikujeta interakcijo s predmeti, še posebej z artefakti UI. Kot navajata Damiano in Dumouchel (2018), je z uvajanjem socialnih agentov zaznati pomemben premik v konceptualizaciji umetne inteligence. Ob poskusih oblikovanja UI po modelu človeškega uma se v primeru socialnih agentov kot model uporabljajo še socialne in kognitivne kompetence ljudi. Človek in naprave se iz vlog, ko sta uporabnik in orodje, prelevijo v sodelavce, med katerimi interakcija poteka skozi nove izrazne forme in vedenje. Predmeti, kot so roboti, različne tehnološke naprave, pa vse do pametnih hiš in virtualnih agentov, soustvarjajo okoliščine medsebojnih interakcij in doprinašajo k načinu doživljanja ljudi, saj interaktivni objekti delujejo proaktivno in se nepredvidljivo in dinamično odzivajo. Poleg tega predmeti z uporabniki vstopajo v stik tudi s telesno govorico. Vse to pa so tudi sprožilci antropomorfnega in animističnega vstopanja v interakcije s socialnimi agenti.

V nadaljevanju predstavljam nekaj znanih aplikacij antropomorfizma in animističnega načrtovanja na področjih uporabe UI.

6.1 Internet igrač

Z razvojem UI se na področju razvoja interakcije človek-stroj (angl. *human-computer interaction*, HCI) še okrepi razvoj vmesnikov, ki bodo čimbolj naravni za ljudi. Ti vmesniki omogočajo interakcijo z umetnimi sistemi na podoben način, kot v interakcijo z okoljem vstopamo tudi sicer v življenju. To nas nenehno opominja na našo soodvisnost z zunanjim svetom (Rod in Kera 2010). Kot del teh prizadevanj se na področju UI sistemov, še posebej robotike in interneta stvari (IoT), pojavlja vse več poskusov aplikacije animističnega mišljenja že v fazi načrtovanja socialnih robotov in drugih artefaktov UI. Eden takšnih primerov je 'Internet of Toys'. Animistično načrtovanje igrač skuša pripeljati digitalni svet računalniških iger nazaj v materialni svet ter animistične predstave, kot so avtonomnost, pripisovanje mentalnih stanj in namer, uporabiti v kontekstu interakcije človek-stroj (Zaman, Van Mechelen in Bleumers 2018).

Načrtovalci pametnih igrač in tudi drugih UI sistemov želijo v interakcijo s predmeti vključiti princip intersubjektivnosti, kjer animistično dojeti predmeti postanejo avtonomni in nepredvidljivi agenti. Doživljanje posameznika tako ni več odvisno samo od njega samega, temveč je doživljanje rezultat soustvarjanja vseh akterjev, ki so v danem trenutku v interakciji in izid katerih je negotov. S privzemanjem intencionalnosti naprav, pripisovanjem mentalnih stanj in čustvovanja interakcija z napravami in UI v bistvu postane socialna interakcija, vsaj za človeka, ki v to interakcijo vstopa. Socialna interakcija tako v stik med človekom in predmeti ali programsko opremo vnaša tudi komponento odnosa. Pri raziskovanju animističnega načrtovanja je zato pomembno upoštevati, da čustveni ali le 'čustveni' odzivi naprav soustvarjajo doživljanje ljudi, vključenih v takšne interakcije. Pomembno na primer postane preučevanje 'nalezljivosti' čustvenih stanj, kot sta razburjenje ali apatija. Povratno lahko pričakujemo, da se bo sistem UI učil in odzival na odnosna sporočila, posredovana od človeka, kar ponovno generira nepredvidljivost v takšnih odnosih. Še posebej, ker način učenja socialnih pravil v omreženem svetu naprav ne poteka na enak način, kot poteka socializacija posameznika v okolju odnosov s pomembnimi drugimi.

6.2 Sodelovanje z roboti

Iz zgoraj predstavljenih spoznanj socialne psihologije vidimo, da ljudje dobro prepoznavamo namere drugih ljudi skozi pripisovanje mentalnih stanj. To lastnost lahko uporabimo pri načrtovanju robotov, ki so namenjeni interakciji s človekom. Študije na področju preučevanja interakcije z roboti kažejo, da ljudje robote zmoremo interpretirati na animističen in antropomorfni način, in sicer to velja tako za humanoidne robote kot tudi za avtonomna vozila in podobno. Obravnavamo jih, kot bi bili živi in jim pripisujemo mentalna in emocionalna stanja, ki nato sooblikujejo naše doživljanje (pregled v Thellman, Silvervarg in Ziemke 2017). Thellman navaja, da so roboti preveč kompleksni v svojem delovanju, da bi jih lahko razumeli na temelju fizikalnih zakonov ali glede na poznavanje njihove tehnološke zasnove. Predvidevanje delovanja robotov, ki so opremljeni z UI in namenjeni delovanju v kompleksnih okoljih ali socialnih interakcijah, človek na omenjena načina ne zmore. V primerjavi ocenjevanja namer ljudi in humanoidnih robotov Thellman ugotavlja, da ljudje na temelju pripisovanja mentalnih stanj zelo podobno ocenjujemo vzroke vedenja obojih. Antropomorfno pripisovanje je še posebej lahko uporabno v primeru socialnih robotov. Le-teh ljudje ne uporabljajo kot orodij, temveč z roboti sodelujejo kot s sogovorniki in sodelavci. Kot pravi Damiano, roboti postanejo socialni partnerji, antropomorfizem pa spodbuja socialno izmenjavo (Damiano in Dumouchel 2018). Socialni agenti, ki temeljijo na načrtovanem antropomorfizmu, vse bolj dobivajo aplikacijo v podpori ljudem s posebnimi potrebami, npr. starostniki ali posamezniki s težavami z avtizmom. Damiano kot primer navaja aplikacijo uporabe socialnih robotov v podpori razvoju čustvovanja pri otrocih z avtizmom. Asocialni agenti podprti z UI namreč ne le izražajo emocionalna stanja, temveč se aktivno odzivajo na emocionalna stanja človeških sogovornikov, s katerimi so v interakciji. Na ta način se vzpostavi zanka socialne interakcije, v kateri tako robot kot človek usklajujeta emocionalne odzive. Številne druge aplikacije na področjih zdravstva, izobraževanja in zabave navaja Giger (pregled v Giger et al. 2019), kjer se socialni agenti uporabljajo za zmanjševanje tesnobe med hospitalizacijo, zmanjševanje občutkov osamljenosti in zviševanje psihološkega blagostanja, izboljšanje učnih dosežkov skozi empatično socialno interakcijo ter spodbujanje občutkov zadovoljstva.

6.3 UI in pametne zgradbe

Na področju pametnih zgradb UI apliciramo z namenom podpore prebivalcem pri preseganju določenih omejitev, npr. zdravstvenih, ter za višanje udobja in za zabavo. UI v načrtovanju predmetov, tudi zelo kompleksnih, kot so zgradbe, predmetom omogoča oblikovanje odzivov ter socialno vedenje podobno človeškemu. Na ta način se ljudje na predmete tudi odzivajo in jih obravnavajo, kot bi bili njim enaki (Ahmed 2020). Ahmed navaja, da antropomorfno dojemanje zgradbe lahko ima tudi negativne posledice, saj npr. pretirana avtomatizacija pri ljudeh vzbuja občutek izgube nadzora z upravljanjem lastnega življenja. Tehnična kompleksnost takšnih sistemov je navadno zelo velika, kar po eni strani vzbuja antropomorfizem, po drugi pa lahko doprinaša tudi k nerazumevanju lastnega okolja.

Če pametne zgradbe razumemo kot družbene akterje, lahko v procesu oblikovanja interakcij razločimo funkcijo objekta, ki je relativno stabilna skozi zelo dolga obdobja, od vmesnikov, skozi katere do funkcionalnosti dostopamo. Antropomorfno dojemanje zgradbe skozi prizmo socialnih interakcij usmerja oblikovanje teh vmesnikov. V pametnih zgradbah je glasovna komunikacija in uporaba zvoka neprecenljiva in omogoča intuitivno razumevanje zahtevnih konceptov, ki bi jih bilo težko vizualizirati. V primeru zgradbe in pametnih naprav prebivalci dojemajo naprave in zgradbo skozi percepcijo različnih identitet na osnovi glasovne komunikacije. Tako se posameznik odziva na različne glasove, ki jih zgradba generira, kot tudi na različne socialne akterje.

7 Načrtovani antropomorfizem in etika

V prispevku sem predstavil pregled spoznanj o mehanizmih animističnega in antropomorfnega dojemanja sveta in nekatere možnosti aplikacije in njenih pridobitev v kontekstu UI. Antropomorfizem lahko koristi tudi samim napravam, saj antropomorfno dojemanje lahko spodbuja prosocialno vedenje ljudi tudi v odnosu do naprav. V študiji s socialnim robotom, ki je izražal prijetne občutke ugodja v stiku z udeleženci študije, nihče od udeležencev ni bil pripravljen robota uničiti (Fisher 2013; poišči v Giger). Kljub temu pa aplikacija načrtovanega antropomorfizma odpira nove probleme, med katerimi so številna etična vprašanja. Kognitivna zmota v tem pogledu ni osrednji problem človeka v pogojih načrtovanega antropomorfizma.

Na etične dileme je že zelo zgodaj v razvoju umetnih socialnih agentov opozorila Sherry Turkle (Turkle 2011), ki je preučevala doživljanje interakcij z avtonomnimi ali vsaj na videz avtonomnimi agenti, npr. napravic Tamagochi. Turkle opozori na dodatno dimenzijo odnosnosti, ki je v razlagah razumevanja vedenja socialnih akterjev, razumevanja medsebojne komunikacije in medsebojnega usklajevanja ostala spregledana. Skozi socialne odnose se namreč razvija tudi afektivna navezanost, ki se odraža skozi občutke vzajemnosti in medsebojne povezanosti. Na ta način načrtovani antropomorfizem generira lažne občutke obstoja socialnih odnosov, čeprav dejanskega odnosa v resnici ni. Pri tem razmišljanju ne želim posegati na špekulativno področje možnih drugih oblik (umetne) inteligence, ki bi bila sposobna medosebnih odnosov in se omejujem na običajne predmete in tehnologijo. Kot trdi Turkle, v »kot če« odnosih človek dejansko ne more zadovoljiti potrebe po povezanosti in kljub navidezni vključenosti v skupnost ostaja z občutki samote in osamljenosti. Sprožanje antropomorfizma zaslepi ranljive posameznike, ki začnejo zaupati v lažen občutek resnične čustvene vzajemnosti in recipročnosti v odnosih. Problematično pri tem pa je, kot pravi Turkle, da ljudje v teh okoliščinah začnejo opuščati odnose z drugimi ljudmi. Vendar rešitev v smislu regulacije in prepovedi aplikacije antropomorfizma v načrtovanju socialnih agentov, ki jo ponuja Turkle, pravzaprav ni mogoča, saj so socialni agenti v življenju ljudi pravzaprav že vseprisotni in bodo kot del našega življenja tudi ostali. Verjetno je bolj smiselno družbena in etična vprašanja odpirati in jih naslavljati v luči spoznanj o antropomorfizmu. To razmišljanje lahko zaključim s primeroma, ki ju podajata Damiano in Dumouchel (2018). Tako kot je aplikacija antropomorfizma lahko koristna za ranljive posameznike, npr. avtiste, in jim omogoča spodbudnejši razvoj lastnih sposobnosti. Po drugi strani lahko, npr. socialne robote, uporabljamo za izživljanje agresivnih nagnjen, npr. skozi posilstvo robota. V takšnih primerih uporaba »kot če« odnosov lahko na prvi pogled izgleda celo navidezno družbeno koristna, vendar dejansko etično nesprejemljivo vedenje banalizira in zmanjšuje njegov pomen – posilstvo je še vedno posilstvo. Tako kot lahko skozi preučevanje antropomorfizma analiziramo prenašanje socialnih norm iz medčloveških odnosov na odnose človek-stroj, se moramo vprašati tudi o mehanizmih in učinkih prenosa hibridnih socialnih norm v medčloveške odnose.

8 Zaključek

V pričujočem prispevku sem naredil pregled psiholoških vidikov animizma in antropomorfizma. Pri tem sem pogledal v zgodovinski razvoj razumevanja omenjenih pojavov, njihove značilnosti iz perspektive kognitivnega razvoja in socialne psihologije ter nekaterih aplikacij animizma in antropomorfizma v kontekstu umetne inteligence. Animizem in antropomorfizem se povezujeta s pripisovanjem aktivnosti, ki imajo določeno namero, in sta po svoji naravi vedno socialna. Namera zunanjega agenta se vedno nanaša na človeka oziroma na odnos med zunanjim agentom in človekom, ki posega po antropomorfizmu. V tem smislu antropomorfizem ni povezan z zaznavo, temveč s pripisovanjem in razlago namernega vedenja v odnosu (Airenti 2018). Prav tako ni presenetljivo, kot ugotavlja Arienti, da je antropomorfizem značilen za interakcije z artefakti informacijskih tehnologij, ki so načrtovane za delovanje v interakciji z ljudmi.

Poznavanje značilnosti in sprožilcev antropomorfizma vodi številne raziskovalce in razvijalce informacijskih tehnologij, tako interneta stvari, robotov kakor tudi vozil in celotnih zgradb v aplikacijo antropomorfizma kot načrtovanega načina interakcije z navedenimi artefakti. Orodja informacijske tehnologije se na ta način pretvorijo v socialne akterje in sodelavce ljudi, ki tehnologijo uporabljajo. Na številnih področjih je ta pristop učinkovit. Hkrati se je treba zavedati, da ima v življenju posameznika socialna interakcija bistven učinek na njegovo doživljanje. Tako tudi socialni agenti aktivno soustvarjajo doživljanje ljudi. Ta okoliščina lahko ima tudi negativne učinke na življenje ljudi, ki začnejo s tehnologijo nadomeščati socialne stike z drugimi ljudmi. Nenazadnje je v prihodnje treba zgraditi razumevanje prenosa značilnosti socialnih interakcij iz antropomorfnih odnosov na odnose človek-človek. Kot se skozi antropomorfizem človeški način ustvarjanja in vzdrževanja odnosov prenašajo v odnose človek-stroj, se namreč značilnosti odnosov človek-stroj lahko prenašajo tudi v obratni smeri, kar lahko ima pomembne psihološke in tudi družbene učinke.

Viri in literatura

Ahmed, D. (2020). »Anthropomorphizing artificial intelligence: towards a user-centered approach for addressing the challenges of over-automation and design understandability in smart homes«. *Intelligent Buildings International*, 0(0), str. 1–14. https://doi.org/10.1080/17508975.2020.1795612.

Airenti, G. (2018). »The Development of Anthropomorphism in Interaction: Intersubjectivity, Imagination, and Theory of Mind«. *Frontiers in Psychology*, 9. https://doi.org/10.3389/fpsyg.2018.02136.

Anderson, J. R. (2013). *The Adaptive Character of Thought*. NJ: Erlbaum: Hillsdale. https://doi.org/10.4324/9780203771730.

Bird-David, N. (1999). »'Animism' Revisited«. *Current Anthropology*, 40(S1), S67–S91. https://doi.org/10.1086/200061.

Blakemore, S.-J. in Decety, J. (2001). »From the perception of action to the understanding of intention«. *Nature Reviews Neuroscience*, 2(8), str. 561–567. https://doi.org/10.1038/35086023.

Brown, L. B. in Thouless, R. H. (1965). »Animistic thought in civilized adults«. *Journal of Genetic Psychology*, 107(1), str. 33–42. https://doi.org/10.1080/00221325.1965.10532760.

Caporeal, L. R. in Heyes, C. M. (1997). »Why anthropomorphize? Folk Psychology and Other Stories«. V Mitchell, Thompson in Miles (urd.), *Anthropomorphism, Anecdotes, and Animals*. New York: SUNY Press, str. 59–73.

Crowell, D. H. in Dole, A. A. (1957). »Animism and college students«. *Journal of Educational Research*, 50(5), str. 391–395. https://doi.org/10.1080/00220671.1957.10882394.

Dacey, M. (2017). »Anthropomorphism as Cognitive Bias«. *Philosophy of Science*, 84(5). https://doi.org/10.1086/694039.

Damiano, L. in Dumouchel, P. (2018). »Anthropomorphism in Human–Robot Co-evolution«. *Frontiers in Psychology*, 9. https://doi.org/10.3389/fpsyg.2018.00468.

Epley, N. in Waytz, A. (2010). »Mind Perception«. V *Handbook of Social Psychology*, Hoboken, NJ, USA: John Wiley in Sons, Inc. https://doi.org/10.1002/9780470561119.socpsy001014.

Epley, N., Waytz, A. in Cacioppo, J. T. (2007). »On Seeing Human: A Three-Factor Theory of Anthropomorphism«. *Psychological Review*, 114(4), str. 864–886. https://doi.org/10.1037/0033-295X.114.4.864.

Fiske, S. T. (2014). *Social beings : social core motives in social psychology* (3rd ed.). Princeton: Princeton University Press.

Giger, J. C., Piçarra, N., Alves-Oliveira, P., Oliveira, R. in Arriaga, P. (2019). »Humanization of robots: Is it really such a good idea?«. *Human Behavior and Emerging Technologies*, 1(2), str. 111–123. https://doi.org/10.1002/hbe2.147.

Guthrie, S. E. (1993). *Faces in the clouds: a new theory of religion*. New York: Oxford University Press.

Heider, F. in Simmel, M. (1944). »An Experimental Study of Aparent Behaviour«. *The American Journal of Psychology*, 57(2), str. 243–259. https://doi.org/https://www.jstor.org/stable/1416950.

Levillain, F. in Zibetti, E. (2017). »Behavioral Objects: The Rise of the Evocative Machines«. *Journal of Human-Robot Interaction*, 6(1), 4. https://doi.org/10.5898/jhri.6.1.levillain.

McDonald, L. in Stuart-Hamilton, I. (2000). »The meaning of life: Animism in the classificatory skills of older adults«. *International Journal of Aging and Human Development*, 51(3), str. 231–242. https://doi.org/10.2190/825Y-GWAT-9BM8-G5TR.

Okanda, M., Taniguchi, K. in Itakura, S. (2019). »The role of animism tendencies and empathy in adult evaluations of robot«. *HAI 2019 - Proceedings of the 7th International Conference on Human-Agent Interaction*, str. 51–58. https://doi.org/10.1145/3349537.3351891.

Papert, S. (1988). »The Conservation of Piaget: The Computer as Grist to the Constructivist Mill: Seymour Papert«. V Forman in Pufall (urd.), *Constructivism in the Computer Age*, Psychology Press, str. 14–24. https://doi.org/10.4324/9780203771242-6.

Piaget, J. (2013). *The child's conception of number, Selected Works*. London: Routledge.

Read, C. (1915). »Psychology of animism«. *British Journal of Psychology*, 1904-1920, 8(1), str. 1–32. https://doi.org/10.1111/j.2044-8295.1915.tb00125.x.

Richardson, K. (2016). »Technological animism: The uncanny personhood of humanoid machines«. *Social Analysis*, 60(1), str. 110–128. https://doi.org/10.3167/sa.2016.600108.

Rod, J. in Kera, D. (2010). »From agency and subjectivity to animism: Phenomenological and science technology studies (STS) approach to design of large techno-social systems«. *Digital Creativity*, 21(1), str. 70–76. https://doi.org/10.1080/14626261003654558.

Sutherland, P. (1999). »The application of piagetian and neo-piagetian ideas to further and higher education«. *International Journal of Lifelong Education*, 18(4), str. 286–294. https://doi.org/10.1080/026013799293702.

Thellman, S., Silvervarg, A. in Ziemke, T. (2017). »Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots«. *Frontiers in Psychology*, 8 (NOV), str. 1–14. https://doi.org/10.3389/fpsyg.2017.01962.

Turkle, S. (2011). *Alone Together: Why We Expect More from Technology and Less from Each Other*. New York: Basic Books.

Zaman, B., Van Mechelen, M. in Bleumers, L. (2018). »When toys come to life: Considering the internet of toys from an animistic design perspective«. *IDC 2018 - Proceedings of the 2018 ACM Conference on Interaction Design and Children*, str. 170–180. https://doi.org/10.1145/3202185.320274.

VLOGA UMETNE INTELIGENCE V IZOBRAŽEVANJU IN ZA IZOBRAŽEVANJE

IGOR PESEK,[1] MARJAN KRAŠNA[2]

[1] Univerza v Mariboru, Fakulteta za naravoslovje in matematiko, Maribor, Slovenija
igor.pesek@um.si
[2] Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
marjan.krasna@um.si

Sinopsis V zadnjem desetletju se je povečala vsestranskost prenosnih naprav, njihova povečana procesorska moč pa jih je spremenila iz modnega dodatka v podaljšek človeškega telesa. Umetna inteligenca (UI) je vznemirljiva tehnologija za prilagoditev izobraževalnih izkušenj različnih učnih skupin. Kakorkoli, UI v izobraževanju ni novost, poskusov implementacije je bilo veliko, vendar so se zaradi nezrele tehnologije ali napačnih pristopov do neke mere vsi izkazali za neuspešne. Zdaj lahko opazimo nov zagon in njegov vpliv bo kmalu viden. V izobraževanju lahko UI: personalizira učenje, ustvari pametne učne vsebine, izvaja tutorstvo v inteligentnih tutorskih sistemih, se uporablja kot pomoč učencem s posebnimi potrebami, pomaga učiteljem pri ocenjevanju, omogoča študentom dostop do učnih vsebin itd. UI je vse bolj vključena v naše vsakdanje življenje in velikokrat se niti ne zavedamo, da za nečim stoji prav ona. To pomeni, da moramo uporabnike naučiti osnov UI, saj bodo le tako lahko sprejemali premišljene odločitve o njeni vključenosti v njihova življenja. Zato moramo UI pismenost vključiti že osnovno izobraževanje, ker bodo le tako mlajše generacije znale UI učinkovito in smiselno uporabljati. Članek bo raziskal različne možnosti uporabe UI v izobraževanju in za njega.

Ključne besede:
izobraževanje,
umetna inteligenca,
IKT,
prednosti umetne inteligence,
kompetence

Univerzitetna založba Univerze v Mariboru

DOI https://doi.org/10.18690/um.ff.11.2022.13
ISBN 978-961-286-675-4

THE ROLE OR ARTIFICIAL INTELLIGENCE IN EDUCATION AND FOR EDUCATION

IGOR PESEK,[1] MARJAN KRAŠNA[2]

[1] University of Maribor, Faculty of Natural Science and Mathematics, Maribor, Slovenia
igor.pesek@um.si
[2] University of Maribor, Faculty of Arts, Maribor, Slovenia
marjan.krasna@um.si

Abstract Recently, the versatility of portable devices and their increased processing power transformed them from a fashionable accessory into an extension of the human body. Artificial intelligence (AI) could be an exciting technology for adapting educational experiences of various learning groups. However, AI in education is not a novelty. There have been many attempts of its implementation, yet none of them was really successful. Now, a new progression is on the way. In education, AI can: personalize learning, create smart educational content, perform tutoring in intelligent tutoring systems, be used as assistance for students with special needs, help teachers with grading, enable students access to educational content etc. AI is becoming increasingly integrated into our everyday lives and we are often not even aware that AI is behind something we do. This is reason enough to teach users about the basics of AI to give them a chance to make prudent decisions about how it will be incorporated into their lives. Therefore, AI literacy must be taught already in primary schools, so that younger generations will know how to use it effectively and reasonably. The paper will explore various possibilities of AI usage in education and for education.

Keywords: education, artificial intelligence, ICT, benefits of artificial intelligence, competences

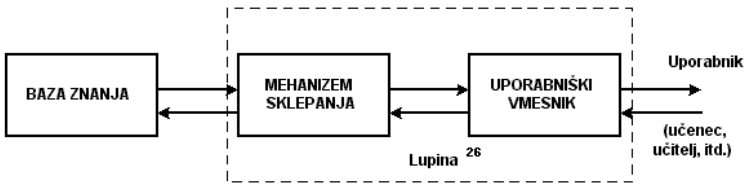
https://doi.org/10.18690/um.ff.11.2022.13 DOI
978-961-286-675-4 ISBN

University of Maribor Press

1 Uvod

Od pričetka uporabe računalnikov v izobraževanju so nekateri razmišljali, kako bi lahko 'zamenjali' učitelja pri nekaterih opravilih. Računalniki imajo pred ljudmi nekaj prednosti in tudi pomanjkljivosti. Vsaj dve prednosti sta samoumevni: (1) lahko ga izključimo in (2) ne potrebuje dodatnega časa za učenje, samo naloži program, podatke in je pripravljen za delo. To, da ni potrebno večletno (desetletno) učenje, je sveti gral človeštva. Če bi bili sposobni naložiti znanje prejšnjih generacij v nove možgane, bi bil naš razvoj veliko hitrejši, kot je danes. Vprašanje pa je, če bi bil obvladljiv na naših bioloških procesnih enotah (možganih)(Gerlič 2000).

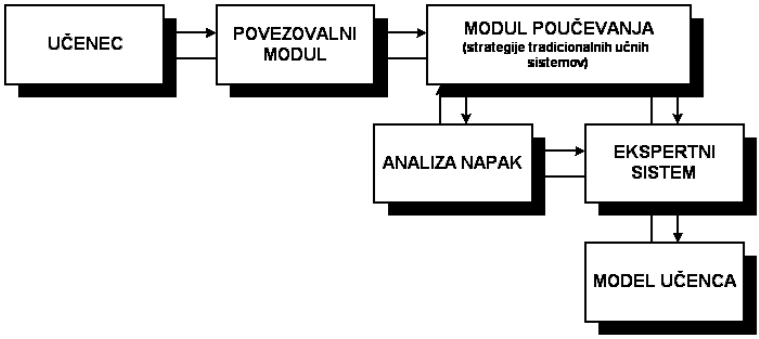
Prvi poskusi z vpeljavo umetne inteligence v izobraževanje so bili izvedeni že v prejšnjem tisočletju. Naši študenti so se s temi poskusi srečali pri predmetu Multimedija, ki so ga obiskovali vsi študenti pedagoških študijskih smeri na Pedagoški fakulteti v Mariboru. Osnovne koncepte so pri predavanjih sicer spoznali, a pri preverjanju znanja se je videlo, da so ti koncepti utopični in bolj skriti nad oblaki, kot bi si želeli. Če smo takrat izhajali iz ekspertnih sistemov za druga področja (v glavnem medicino) in ga želeli pretvoriti v izobraževanje, smo ugotovili, da čas ni dozorel (Krašna 2010).



Slika 1: Ekspertni sistemi v izobraževanju
(Gerlič 2000)

Že osnove – baze znanja (Slika 1) – nismo bili sposobni sestaviti niti za prikaz delovanja. Mehanizem sklepanja je bil še bolj abstrakten in uporabniški vmesnik na nivoju tekstualnega dela. Konceptualno pa je sistem vseeno zanimiv še danes (Slika 2). Pogled na shemo pokaže, da želimo imeti torej model učenca, ki je dovolj primeren za ugotavljanje lastnosti učenca in pomoč pri izobraževanju – tj. izbiranju učnih vsebin, ki jih učenec ne obvlada za razumevanje nekega novega znanja. Seveda model ni dajal nobenih napotkov, kako naj bi bilo to izvedeno v praksi. Analogijo pa lahko vidimo z učiteljem, ki pozna učenca in skozi čas, ter interakcijo z njim,

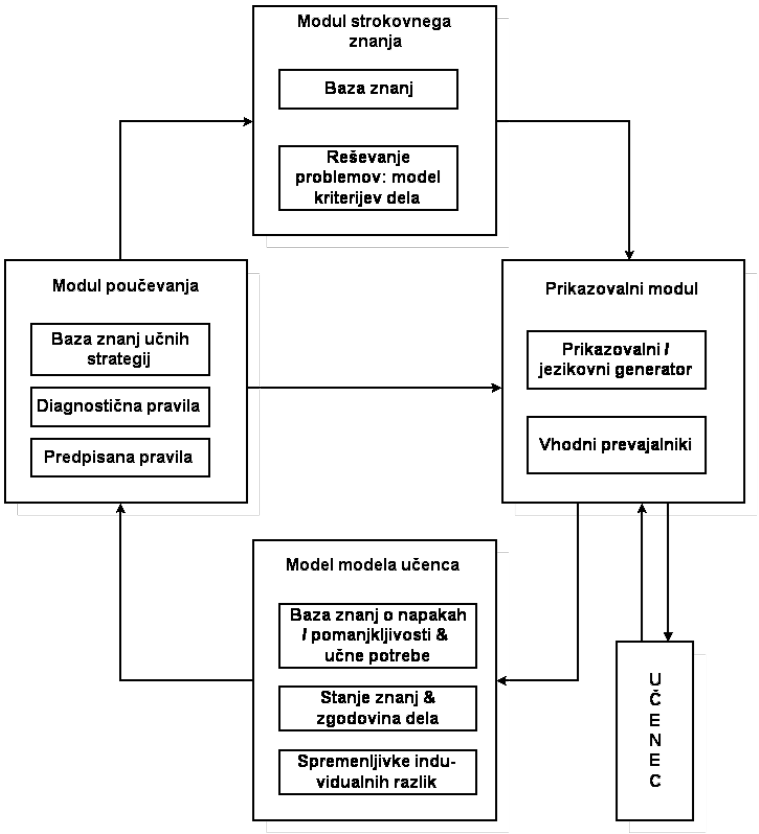
točno ve, kaj učenec zna in kje še mora utrditi znanje, da bo lahko uspešno nadaljeval šolanje.



Slika 2: Inteligentni učni sistemi
(Gerlič 2000)

Prvi poskusi inteligentnega delovanja so bili izvedeni s pomočjo teorije končnih avtomatov. Ti poskusi so bili izvedeni že v času pred računalniško tehnologijo. Zelo kompleksni, človeku podobni avtomati, ki so delovali po v naprej predvidenih stanjih in prehodih med njimi, so bili zanimivi za bogataše v času renesanse. Imeli so dolgoročni vpliv na avtomatizacijo dela in kasneje tudi razvoj učenja, kot ga poznamo danes (Robson 2010).

Po drugi svetovni vojni so se pričela izobraževanja na daljavo in s tem tudi koncept programiranega pouka. Vodenje po točno določenih korakih, z dodatno razlago ali ne, kjer se je učeči lahko premikal skozi učno gradivo in v posameznih točkah preveril svoje znanje. Koncept je pridobil na veljavi šele s pojavom računalnikov, povezanih v internet in z množico spletnih uslug. Ta koncept priprave učnih gradiv še danes uporabljamo za kombinirano učenje in moramo priznati, da do epidemije COVID-19 večina učiteljev ni pomislila, da ga bo kadarkoli uporabila. Sprememba je bila v tem primeru zelo dobrodošla in pokazala, kje vse se najdejo luknje v znanju in napačna predvidevanja o bodočem razvoju učenja in poučevanja.



Slika 3: Integrirano računalniško podprto poučevanje
(Gerlič 2000)

Pri pedagoških študijskih programih je mogoče uporabiti UI tudi kot učenca. V takšnem primeru študent poučuje UI in na podlagi testiranja znanja ugotovi, kako uspešno je njegovo poučevanje (Woolf 1990).

Prvič smo se s takšnim problemom srečali pri robotiki, kjer so tradicionalno programe pisali za vsak aktuator posebej in tako časovno in prostorsko usklajevali robotske gibe. To je bilo zelo časovno potratno opravilo, ki je zahtevalo veliko procesorskega časa zaradi nenehnih transformacij med različnimi koordinatnimi sistemi (danes to ni več takšen problem, ker to rešijo grafične kartice). Takrat se je porodila ideja, da bi lahko v resnici postavili robota v stanje učenja, namesto da bi programirali aktuatorje, bi s senzorjev na aktuatorjih preprosto odčitali podatke, medtem ko bi premikali robotsko roko. Robot bi kasneje samo ponovil zabeležene

podatke in preko teh dosegel ponovljivost akcije (industrijska robotika[1]). Umetna inteligenca pa bi lahko optimizirala zaporedje akcij in tako minimizirala premike in čas. Danes je to standardna tehnika za poučevanje robotov (Haage et al. 2017).

2 Umetna inteligenca v izobraževanju

Računalnik v izobraževanju je najprej prišel na nivoju strojev, ki so zamenjali klasične medije. Torej namesto papirja smo tipkali v oblikovalnike besedil, namesto kasetofona ali videokaset smo predvajali zvok in video s pomočjo računalnika. Ta napredek je bil izjemno pomemben, saj smo namesto različnih naprav potrebovali le eno, ki je znala vse to, za kar smo nekoč potrebovali celo plejado drugih naprav. Računalnik pa ne predvaja oziroma prikazuje le podatke, ampak omogoča, da jih spreminjamo – obdelujemo. Tako ni bilo več treba prikazati vsega videoposnetka od začetka in iskati tisti del, ki ga želimo prikazati v razredu, izrežemo lahko le točno tiste dele in jih brez časovnih zamikov pokažemo na zahtevo.

To se je pokazalo kot časovno zahtevno opravilo, če smo ga naredili le enkrat. Smo pa hitro ugotovili, da je mogoče te gradnike večkrat uporabiti in s tem kasneje pridobiti čas za pripravo drugačnih, a vseeno podobnih gradiva (re-uporaba gradnikov)(Krašna 2010).

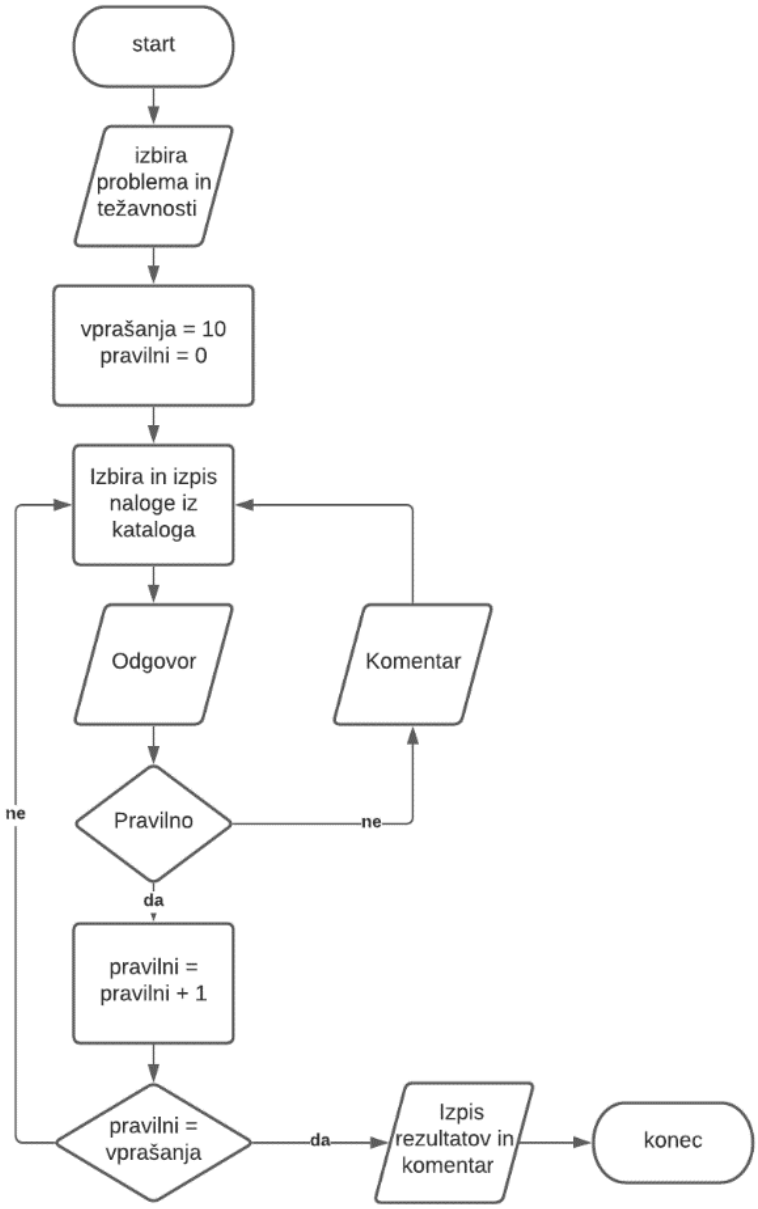
Ko so računalniki postali dostopni slehernemu učečemu, smo lahko vsebine objavili na spletnih straneh tudi specializiranih – spletnih učilnicah in tako prihranili čas z razmnoževanjem, in če smo se zares potrudili, nas je računalnik razbremenil ponavljajočih se opravil, ki velikokrat za učitelja niso motivirajoča, a jih učeči potrebujejo za napredek in razlago (Krašna 2010).

2.1 Vaja in utrjevanje

Prva razbremenitev učitelja se je pokazala pri vaji in utrjevanju (t. i. drilih). To je duhamorno opravilo, ki ga lahko opravi sleherni končni avtomat. Treba si je le izmisliti testirano vrednost (eno izmed možnih izbere računalnik s pomočjo omejene *random* funkcije), pridobiti odgovor učečega in preverjati odgovor s pravilnim

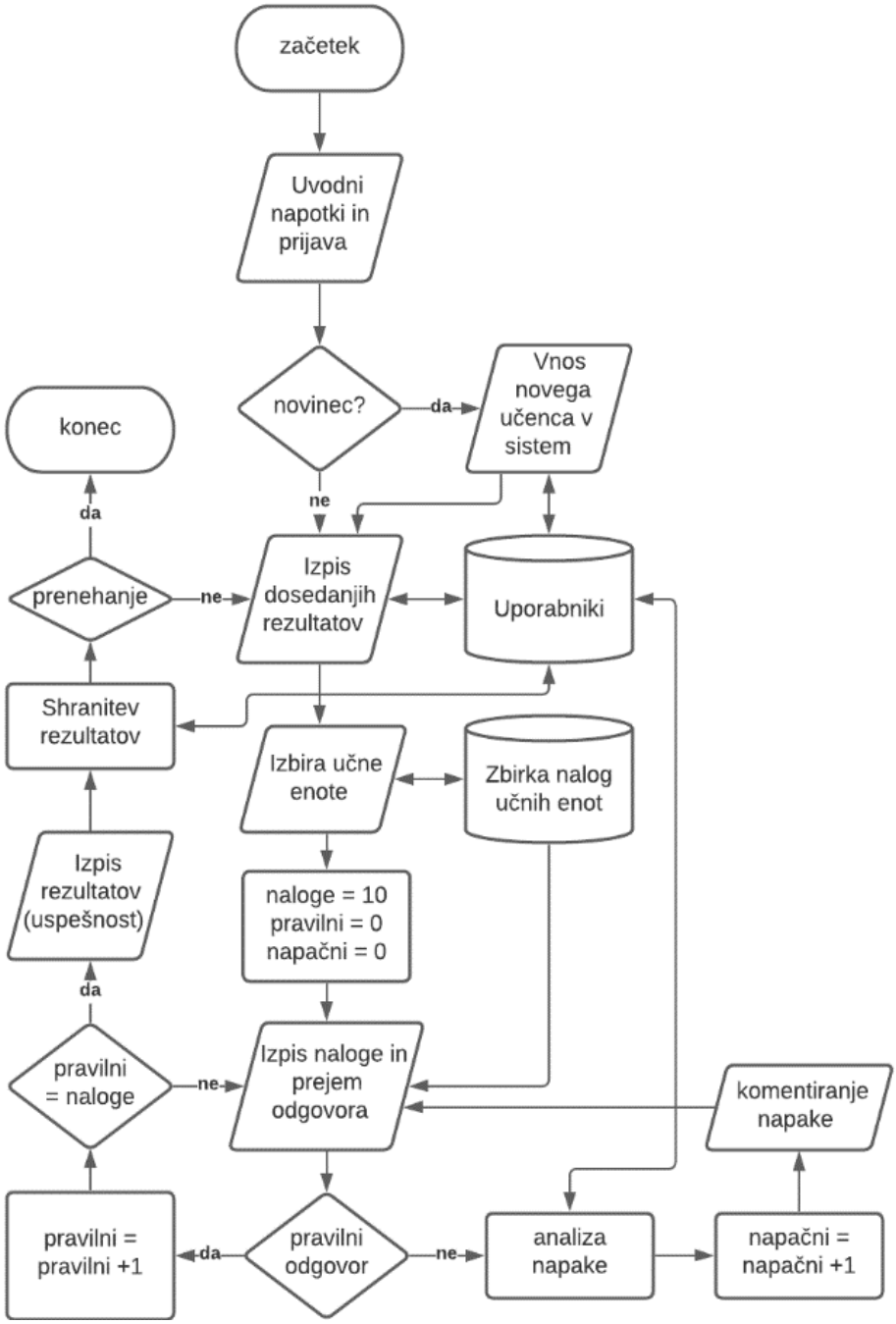
[1] Industrijska robotika (Karel Jezernik), Predmet na Fakulteti za elektrotehniko, računalništvo in informatiko, študijsko leto 1988/1989

odgovorom na izbrano možnost. Drile je mogoče tudi razširiti in jih narediti še 'inteligentnejše'. Na sliki (Slika 4) vidimo, kako takšen dril deluje.



Slika 4: Fazni diagram enostavne strategije vaje in utrjevanja – konča se po 10 pravilnih odgovorih

Vir: lasten.



Slika 5: fazni diagram zahtevnejšega drila s shranitvijo rezultatov

Vir: lasten.

2.2 Personalizirano učenje

Personalizirano učenje z umetno inteligenco je učni pristop, ki se osredotoča na oblikovanje učenja z upoštevanjem posebnih potreb posameznega učenca. Bolj natančno to pomeni, da komponente učenja, kot so tempo učenja, osebne želje, oblika pouka in učni stili, prilagodimo tako, da ustrezajo potrebam učenca in na ta način povečamo učinkovitost učenja. Raziskave so pokazale, da je personalizirano poučevanje veliko učinkovitejše kot poučevanje celotnega razreda (Bloom 1984). Problem je, da učitelj ne more prilagoditi svojega poučevanja na celoten razred, po drugi strani pa je individualno poučevanje vsakega učenca posebej neizvedljivo. Uporaba tehnologije za te namene se raziskuje že dlje časa in pri tem zaradi vedno več digitalnih sledi, ki jih učenci puščajo, izredno napreduje (Xie, Hwang in Wang 2019). UI in strojno učenje blestita pri prepoznavanju vzorcev, ki jih človek ne prepozna, pogosto zaradi velike količine podatkov. Zato lahko UI z analizo izobraževalnih podatkov učenca pomaga učitelju pri prepoznavanju načinov, kako posamezni učenci razumejo snov.

Tukaj lahko pomaga umetna inteligenca, in sicer na dva načina. Pri prvem načinu pomaga učitelju s koristnimi podatki o učencu. Učitelj lahko ustrezno reagira in prilagodi poučevanje učenca z dodatnimi razlagami in nalogami. V drugem načinu pa lahko UI v celoti prevzame poučevanje učenca in mu ponuja prilagojene učne vsebine.

Za oba načina potrebujemo zajemanje, združevanje in analiziranje podatkov, pridobljenih iz več različnih virov, vključno z aplikacijami za učenje, spletnimi viri, založniki in drugimi sistemi za namene učenja, da se ustvari celovit učni model posameznih učencev (Kurilovas 2019).

Katere prednosti personaliziranega učenja ponuja UI? V nadaljevanju predstavljamo nekatere:

a. Povečanje motivacije, angažiranosti in rezultatov učenja
Algoritmi strojnega učenja lahko s pomočjo digitalnih sledi učenca napovejo rezultate, kar omogoča prilagoditev učnih vsebin, ki jih temeljimo na preteklih rezultatih in individualnih ciljih. To vključuje scenarije, v katerih bi sistem prepoznal, da bi učenec lahko dejansko preskočil nekaj modulov,

da bi se podal na bolj celovito in manj linearno učno pot kot nekdo, ki bi mu morda manjkale osnovne veščine, povezane s to temo.

b. *Preusmerimo na učinkovite učne vire in naloge*
Učenci prejmejo natančno tiste učne (npr. spletne) vire, ki jih potrebujejo za zapolnitev vrzeli v znanju in razumevanju ter doseganju učnih ciljev, kar pomeni manj časa za učenje.

c. *Avtomatiziramo dostop do učnih virov*
Umetna inteligenca lahko učence vodi po personaliziranih učnih poteh, ki se samodejno prilagajajo znanju učenca, ki se določi glede na aktivnosti v digitalnem učnem okolju. Z ustreznimi kriteriji se lahko ponudijo učne vsebine, ki poskušajo ciljno zapolniti zaznane vrzeli v znanju.

d. *Dinamično prilagodljive učne platforme*
Učne platforme se sistemsko nadgradijo z zmožnostmi dinamičnega samoučenja iz vedenja, ki jih prikažejo učitelji in učenci v digitalnem svetu. Posledično ustvarijo ustrezno pedagogiko in samodejno prilagodijo okolja e-učenja tako, da ustrezajo pedagogiki (Almohammadi, Hagras, Alghazzawi in Aldabbagh 2017).

Umetna inteligenca lahko učiteljem že vnaprej pripravi profile učencev, tako da so učitelji v boljšem položaju, da izkoristijo svoje usposabljanje in spretnosti za reševanje individualnih potreb teh učencev že od samega začetka, namesto da bi porabili tedne ali mesece za ugotavljanje učnih težav posameznih učencev.

2.3 Povezovanje učnih vsebin

Del personaliziranega učenja, prilagojenega individualnim potrebam, željam in interesom, so učne poti, ki jih je s pomočjo računalniške tehnologije mogoče prilagoditi številni in raznoliki populaciji učencev. Eden ključnih problemov je določitev učne poti, ki ji učenec sledi, da zaključi učni program. Obstoječe metode se na splošno opirajo na predhodno poznavanje vsebine ustvarjalca učne poti, ki tudi nastavi pogoje in omejitve za določanje zaporedja učnih gradiv.

Učinkovita uporaba umetne inteligence, podatkov in analitike ter strojnega učenja lahko učiteljem omogoči, da pripravijo ustrezno učno izkušnjo in ustvarijo prilagojene učne poti za vsakega učenca posebej. V tem primeru UI deluje kot povezovalec različnih učnih vsebin, kot so spletna učna gradiva, izobraževalni videoposnetki, naloge za preverjanje in druge oblike učnih vsebin. Z zbiranjem

podatkov o učnem procesu učenca UI določi, katero vsebino bo učenec v naslednjem koraku obravnaval.

Učenec tako deluje kot prejemnik, ki reagira na vnaprej določena zaporedja znanja, sledi učnim postopkom in potem izvaja učne dejavnosti, ki jih je UI določila za doseganje vnaprej določenih ciljev. Tipična implementacija takšnih učnih poti so zgodnja raziskovanja v inteligentnih tutorskih sistemih, ki so sedaj nadgrajena z UI (Fan in Pengcheng 2021). V prihodnje si lahko obetamo nadgradnjo personaliziranih učnih poti znotraj namenskih okolij, ki bodo skrbela za na učenca prilagojene vsebine in ki se bodo odzivala na učenca in njegov učni proces.

2.4 Učenci s posebnimi potrebami

Uporaba UI je še posebej pomembna za otroke, ki potrebujejo posebno obravnavo v izobraževanju. Ti otroci imajo večinoma eno od učnih težav, nekateri imajo okvare v socialnih veščinah, kot sta jezik in komunikacija, ali pa imajo težave pri branju, pisanju in računanju. Ob upoštevanju vsega tega je očitno, da tradicionalni pristop za otroke, ki se izobražujejo, ne velja. Zaradi svoje edinstvenosti se mora izobraževanje prilagoditi, da bo bolj učinkovito (Drigas in Ioannidou 2013). V nadaljevanju predstavljamo tri področja, kjer je smiselna uporaba UI.

a) Personalizirano izobraževanje za vsakega učenca
   Izobraževalne aplikacije z vključeno UI so običajno prilagojene in zabavnejše za otroke. Ker v aplikacijah običajno tudi ni sošolcev, s katerimi bi se primerjali, so zato učenci bolj samozavestni. Poleg tega se lahko učijo kjerkoli in kadarkoli.
b) Povečanje dolžine pozornosti
   Učenčeva pozornost je pri otrocih s posebnimi potrebami krajša, zato je uporaba aplikacij z vgrajeno UI smiselna. Te aplikacije namreč zaznajo, da učenčeva pozornost upada, zato lahko z različnimi ukrepi ponovno pridobijo učenčevo pozornost.
c) Diferenciacija in individualizirane vsebine
   UI pomaga otrokom s posebnimi potrebami z individualiziranim pristopom, ki temelji na izdelanem učnem profilu učenca. Na ta način lahko UI pripravi različne vsebine, ki krepijo področja, ki jih mora učenec izboljšati.

Ena od posledic uporabe UI v izobraževanju je lahko tudi inkluzivna pedagogika, ki vključuje vse otroke, ne glede na njihovo morebitno kategorizacijo glede posebnih potreb (Garg in Sharma 2020).

3 Prevajanje učnih virov

Na svetovnem spletu in v različnih repozitorijih učnih gradiv obstaja veliko visoko kvalitetnih e-gradiv. Njihova največja težava je, da so običajno pripravljena v jeziku avtorja e-gradiva. Posledično je dostopnost e-gradiva manjša, saj je za učitelje jezikovna pregrada običajno previsoka. Učitelji zato večkrat izgubljajo čas s pripravo lastnih e-gradiv, ta čas pa bi lahko izkoristili raje za kvalitetnejšo pripravo na pouk. To slednje se je pokazalo predvsem v covidni krizi, ki je pokazala, da so učitelji večkrat pregorevali za pripravo videoposnetkov za poučevanje na daljavo (Pestano Perez, Pesek, Zmazek in Lipovec 2020).

Prevajalniki med jeziki so v zadnjem obdobju zelo napredovali, predvsem po vključitvi umetne inteligence v delovanje prevajalnih algoritmov. Prevajalniki ne delujejo več na podlagi programiranih pravil, temveč se učijo iz korpusov prevajanih del, s pomočjo katerih izdelajo lastna pravila za prevajanje med jeziki.

Prevajalnike lahko koristimo tako, da nam trajno prevedejo e-gradivo v ciljni jezik. Ta način je priporočljiv sploh pri prevajanju strokovnih del, saj je vseeno treba strokovno pregledati prevod. Ker se prevajalniki učijo, bo potreba po takšnem strokovnem pregledu sčasoma vedno manjša.

Prevajalnike lahko odlično izkoristimo tudi za prevajanje video posnetkov, saj je tudi prepoznava govora napredovala podobno hitro kot samo prevajanje. Prevajanje videoposnetkov tako poteka dvostopenjsko, in sicer najprej prepoznava govora in nato prevajanje besedila v ciljni jezik. Z napredkom sintetiziranega govora bomo kmalu imeli tudi sinhronizirane videoposnetke, ki jih bo v celoti izvedel računalnik (X5Gon 2021).

Prevajanje pa je možno tudi v realnem času, kar lahko izkoristimo pri predavanjih skupinam, kjer so poslušalci iz različnih jezikovnih področij. Računalnik tako prevaja prosojnice ali druga e-gradiva.

Vsekakor napredek prevajalnikov omogoča uporabo e-gradiv iz različnih jezikov, kar bo predvsem pomagalo Evropi in državam, ki si ne morejo privoščiti priprave lastnih e-gradiv za vsa predmetna področja.

4 Preverjanje znanja na višjih taksonomskih ravneh

Preverjanje znanja s pomočjo računalnika je v večini primerov izvedeno s pomočjo zaprtih testnih nalog. Pri tem je jasno, da so ti testi primerni za preverjanje faktografskega znanja. Če pa ustrezno sestavimo vprašanja, pa lahko preverjamo tudi višje taksonomske ravni znanja.

4.1 Elektronsko preverjanje znanja

Na tržišču obstaja veliko avtorskih orodij, prav tako pa je mogoče najti tudi veliko prostodostopnih avtorskih orodij, ki podpirajo osnovne tipe vprašanj:

alternativni tip (npr. Da/Ne ali je res/ni res),
izbirni tip (*multiple choice*),
primerjalni tip,
tip dopolnjevanja,
numerični odgovor in
besedilo ali esej.

V zadnjem času lahko s pomočjo vtičnika H5P razširimo nabor interaktivnosti, a smo še zmeraj omejeni na tekstovne in številske vnose ali potegni-in-spusti (angl. *drag & drop*). Še zmeraj pa ne moremo sestaviti poljubne sheme, kot bi jo lahko v primeru klasičnega pisnega preverjanja znanja.

Elektronsko preverjanje znanja izključuje človeške faktorje in je objektivno ter enakovredno do vseh. Zapleti, ki se lahko pojavijo, so zaradi tehnoloških omejitev, ker ni mogoče vključiti vseh načinov preverjanja. Zaradi individualnih lastnosti vsakega posameznika je zmeraj mogoče, da komu takšen način preverjanja znanja ne ustreza in posledično dobi slabšo oceno, kot bi jo dobil pri drugačnem načinu preverjanja tega istega znanja. Čeprav je mogoče elektronsko preverjanje znanja tako za oceno kot za samopreverjanje, se zelo izkaže v slednjem (Krašna 2015).

Priprava vprašanj zahteva veliko miselnega in organizacijskega napora. Vprašanja morajo biti pripravljena tako, da jih učeči enoumno razumejo in da z njimi dosežemo objektivnost. Vprašanja sestavljamo tako, da težimo k pozitivni strategiji, negativna strategija zastavljanja vprašanj zahteva najprej pozitivno rešitev in potem njeno negacijo. V procesu testiranja, ki je za udeležence stresno, se tako lahko zgodi, da dosegajo slabše rezultate. Negativna strategija je lahko uporabna na tekmovanjih, kjer se zahteva ne samo znanje, ampak tudi nedvoumno razumevanje vprašanj. Preverjanje znanja za oceno pri predmetih naj vseeno temelji na pozitivni strategiji, ker sovpada z naravnim razmišljanjem in ne zahteva dodatnih mentalnih opravil (Krašna 2015).

Elektronsko testiranje je izvedeno zelo hitro. Študenti prejmejo rezultate običajno takoj po zaključku testa. Pri tem imajo možnost, da dobijo tudi razlago napačnih odgovorov in tako proces testiranja vključuje tudi učenje. Smiselno je, da dopustimo učečim, da si sami izberejo vrstni red odgovarjanja na vprašanja. S tem najprej odgovorijo na vprašanja, ki jih zagotovo znajo in od njih zahtevajo manj miselnega napora, potem pa se posvetijo drugim vprašanjem, ki jih rešijo s pomočjo znanih dejstev in sklepanja. Ker so za elektronsko preverjanje primerna vprašanja zaprtega tipa, je treba razmisliti, koliko časa bo posamezno vprašanje zahtevalo in temu ustrezno prilagoditi čas testiranja. Če ne upoštevamo tega in podaljšamo čas testiranja preko, za večino, normalne mere, se pričnejo kazati negativne posledice predolgega časa in spreminjanje pravilnih odgovorov v napačne (preveč razmišljanja, angl. *overthinking*). Ob uvedbi HTML v5 pa dobimo tudi dodatne možnosti za reševanje problemov, ki včasih niso bile tako elegantne in jih je bilo treba predelati v neko drugo, bolj tradicionalno, obliko.

4.2 Višje taksonomske stopnje preverjanja znanja

Zelo preprosto je elektronsko preverjati faktografsko znanje. Čeprav lahko slišimo polemike, da faktografskega znanja naj ne bi več učili, ker jim je nemudoma (na mobilnih telefonih) zmeraj dosegljivo. Vseeno pa ugotavljamo, da brez teh osnov tudi logično sklepanje ne daje dobrih rezultatov. Učeči, pri katerih smo uporabili metodo obrnjenega učenja, so običajno pregledali le literaturo, ki je povsem očitna, v globino problema pa se niso osredotočili. Tudi vprašanja, ki naj bi jih napotila na boljše poznavanje problema, se jim niso zdela pomembna. Videnje problema samo z njihovega zornega kota je bilo tako površno in ozko.

Kompleksnost

Odprtost

Akcija	I	II	III	IV	V
Izbiranje	*res / ni res*	*alternativne izbire*	*več izbir*	*implicitni odgovori*	*nivoji zanesljivosti*
Identifikacija	*večkratni res / ni res*	*da / ne z razlago*	*več odgovorov*	*več odgovorov s slikami*	*izbira delov slike*
Povezava	*povezovanje*	*kategorizacija*	*vrstni red*	*sestavi prioritete*	*sestava dokaza*
Popravljanje	*odstrani tujek iz seznama*	*pomešane besede*	*najdi napačni odgovor*	*najdi napake na sliki*	*reševanje problema*
Dopolnjevanje	*izpolni praznine*	*izpolni praznine s padajočega menuja*	*vstavljanje v formularje*	*izračunaj odgovor*	*razumevanje govora*
Izdelava	*simulacija laboratorija*	*analiziraj odprte odgovore*	*poveži koncepte na načrtu*	*nariši veljavno sekvenco*	*zariši področja na sliki*
Projekt	*odprti odgovori / esej*	*poravnava besed*	*tabelarične naloge*	*predstvitve nalog*	*multimedijski projekt*
Sodelovanje	*diskusija v forumu*	*delitev dokumentov in pregled*	*skupne objave*	*blog skupine z deljenimi vlogami*	*reševanje problema v skupini*

Slika 6: Kompleksnost proti odprtosti v testih in nalogah
(De Praetere n.d.)

Če želimo preverjati višje taksonomske stopnje, pa moramo uporabiti računalniško prilagodljivo testiranje (CAT – *Computer Adaptive Tests*), ki pa trenutno še ni podprto z avtorskimi orodji. Prilagodljivo testiranje je zmeraj v uporabi pri ustnem preverjanju znanj, kjer učitelj s podvprašanji dobi bolj podrobno sliko učenčevega znanja. Zaloga nalog, ki jih lahko odgovarja učeči na preverjanju znanja, niso za preverjanje višjih taksonomskih stopenj, če jih ni mogoče povezati v smiselno celoto. Učitelj v ustnem zagovoru zlahka pridobi znanje, sposobnosti in kompetence od učečega. Učeči so individuumi in procesirajo pridobljene informacije na različne načine in tako pridobijo znanje iz določenega področja (Costello in Mundy 2009) in temu ustrezno učitelj prilagodi ocenjevanje.

Prilagodljivost je sposobnost za »pripravo težavnostne stopnje vsakega vprašanja glede na pravilnost prej pridobljenih odgovorov« (Basu, Cheng, Prasad in Rao 2007). Prilagodljive tehnike so bile razvite za omejene sisteme že pred razvojem svetovnega spleta (Barra, Iannaccone, Palmieri in Scarano 2002) in tako lahko najdemo podatke, da je bil prvi prilagodljivi test predlagan (od Lorda) že leta 1980. Kasneje so uporabljali prilagodljive teste pri računalniškem testiranju (CBT – *Computer Based Test*) vedno pogosteje. V literaturi lahko najdemo različni pojmovanji za računalniško testiranje: CAA (*Computer-Adaptive Assessment*) in CAT (*Computer Adaptive Testing*). Pri prilagodljivih e-testih se število in vrsta vprašanj prilagaja glede na pravilnost

odgovorov prejšnjih vprašanj in posledično sposobnosti testiranega (Abdullah in Cooley 2002; Chen in Wang 2010). Takšni testi so hitrejši za administracijo, zmanjšujejo stroške za izvedbo testa, značilno zmanjšajo število vprašanj (celo za 50 % (Cheng, Rodrigez in Basu 2009)), so prijazni študentu in niso manj zanesljivi od neprilagojenih testov (Basu, Cheng, Prasad in Rao 2007) (Abdullah in Cooley 2002). Največja pomanjkljivost teh testov pa je modeliranje s pomočjo verjetnostne funkcije (Cheng, Rodrigez in Basu 2009). Prav tako tak način testiranja ne rešuje problemov računalniškega testiranja: cena tehnologije, napačno delovanje ali nedelovanje tehnologije, učenje administriranja in študentskih IKT veščin. Teorija, ki stoji za prilagojenim testiranjem, je model študenta (SM – *Student model*) (Chen in Zhang 2008), ki modelira obnašanje in karakteristike študenta ter (IRT – *Item Response Theory*) (Chen in Wang 2008 in 2010; Guzmán in Conejo 2005), ki je robustna dobro znana psihomotorična teorija ocenjevanja v izobraževanju. Slednja je pogoj za ocenjevanje študentovega znanja, izbiro naslednjega vprašanja, mora se ovrednotiti v vsakem trenutku in odloča o zaključitvi testiranja (Guzmán in Conejo 2005). Nekateri avtorji ti dve teoriji (SM in IRT) razumejo kot CAT teorijo (*Computer Adaptive Testing*) (Danielienė in Telešius 2008).

Pri reševanju problema naprednega elektronskega testiranja znanja se moramo približati ustnemu zagovoru, ki bi ga izvedel učitelj. To lahko naredimo le s pomočjo metod umetne inteligence in ekspertnih sistemov v izobraževanju (Slika 4). Ker še zmeraj nimamo na voljo veliko sistemov, ki bi jih lahko uporabili za napredno testiranje znanja, si pomagajo s prilagodljivim testiranjem tako, da pripravimo »inovativne spletne sisteme, ki postavljajo naslednja vprašanja glede na odgovore prejšnjih vprašanj« (Basu, Cheng, Prasad in Rao 2007). Produkti, ki jih vseeno lahko najdemo v literaturi, so: Test++ sistemi, ki temeljijo na teoriji iger (Barra, Iannaccone, Palmieri in Scarano 2002), SIETTE za izdelavo testov za samopreverjanje znanja (Guzmán in Conejo 2005), AITS sistemi, ki uporabljajo tehnike umetne inteligence (Hatzilygeroudis, Koutsojannis in Papavlasopoulos 2006), sisteme, ki uporabljajo tehnologijo inteligentnih agentov (Song, Chen in Gao 2011), nekateri pa uporabljajo tehnike primerjanja (Danielienė in Telešius 2008).

Razlogi za uporabo CAT so samoumevni. Uporabni niso le za preverjanje znanja, ampak tudi za učenje. Če takšen sistem ugotovi nerazumevanje v odgovorih študenta, lahko poskusi preveriti to znanje na drugačen način. Študent lahko pokaže, da snov v resnici obvlada in da mu predhodno postavljeno vprašanje v resnici ni bilo razumljivo v trenutnem kontekstu. Z dodatnim razjasnitvenim vprašanjem pa

študent dobi tudi povratno informacijo, kakšen odgovor se pričakuje na originalno vprašanje, če je snov usvojil. Tak pristop (paradigmo) imenujemo tudi učenje s pomočjo ocenjevanja. Prične se s postavitvijo vprašanja srednje stopnje zahtevnosti in izbira pot skozi vprašanja glede na študentove odgovore. Struktura vprašanj pa je lahko dinamična ali pa determinirana (npr. v obliki drevesne strukture) (Cheng, Rodrigez in Basu 2009). Ocena se izračuna glede na odgovore in pot zastavljenih vprašanj. Teorija CAT zagotavlja, da ocena študentovega znanja ne variira glede na število postavljenih vprašanj v procesu ocenjevanja (Danielienė in Telešius 2008).

4.3 Načrtovanje računalniških prilagodljivih testov

Za izdelavo prilagodljivih testov mora učitelj pripraviti načrt možnih vprašanj in odgovorov – test diagram prehoda stanj. V našem primeru predlagamo spremembo diagrama prehajanja stanj, kot ga predlaga metoda modeliranja (UML – *Unified Modeling Language*), da uporabimo standardne gradnike in vključimo dodatne, ki jih potrebujemo za modeliranje našega pristopa. V predlaganem primeru smo vprašanja modelirali z zaokroženimi pravokotniki, za meta-vprašanja smo uporabili zaokrožene sive pravokotnike. Pravokotnike uporabljamo kot pogoje za prehode in pravokotniki, ki združujejo več vprašanj, ponazarjajo nabor vprašanj iz nekega področja. Vprašanja iz nabora vprašanj se lahko zastavijo v poljubnem vrstnem redu in v poljubni količini.

Primer prilagodljivega ocenjevanja: Kako nadgraditi računalnik (Krašna, Repnik, Bratina in Kaučič 2012)

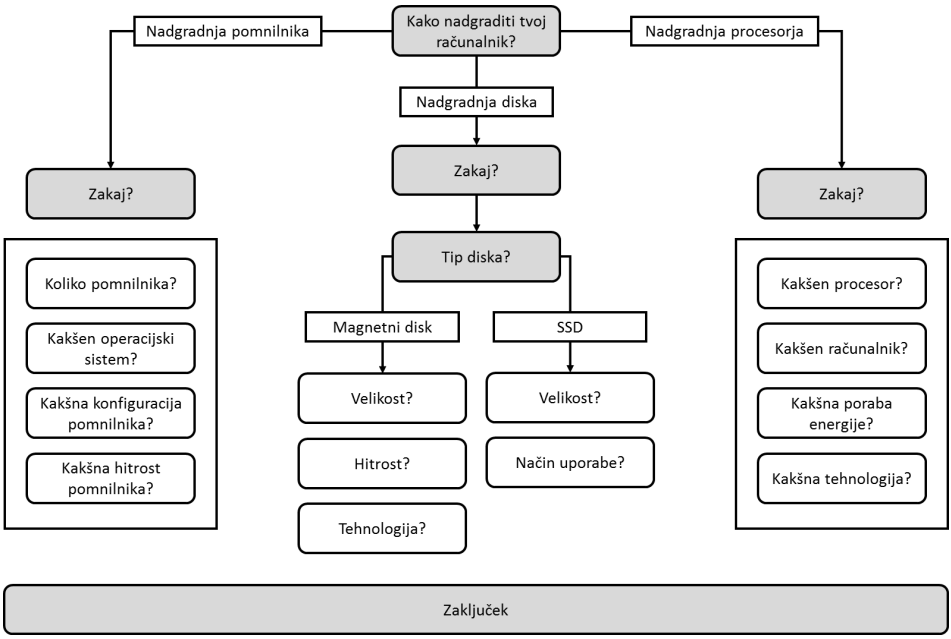
Za lažje razumevanje smo pripravili primer, kako pripraviti strukturo naprednega preverjanja znanja. Primer je s področja računalništva in ocenjuje znanje (in kompetence) študenta, če je postavljen pred nalogo, kako nadgraditi svoj računalnik. Da je študent učinkovit pri tem delu, mora poznati aparaturno opremo svojega računalnika. V preteklosti se je takšno vprašanje izkazalo kot najboljše za ugotovitev razlike pri kompetencah študenta (ali študent razume ali pa se je le naučil) in profesor že v nekaj vprašanjih dobi čisto natančen vpogled v kompetence študenta na tem področju.

V glavnem lahko nadgradimo računalnik na treh področjih. Seveda bi bil trivialni odgovor tudi nakup novega računalnika, ki pa ga zaradi testiranja ne dovoljujemo. Tako lahko študent izbere: nadgradnjo pomnilnika, nadgradnjo diska ali nadgradnjo procesorja. Če študent izbere, da bo nadgradil pomnilnik potem dobi vprašanja, ki se nanašajo na pomnilnik (koliko pomnilnika namerava dodati, kakšen operacijski sistem ima, kakšno ima trenutno konfiguracijo pomnilnika, kakšno hitrost pomnilnika ima. Seveda ni potrebno, da zastavimo vsa vprašanja, ampak jih lahko zastavimo več, če ugotovimo, da študent pri katerem vprašanju ni uspešen. Obstajajo pa nekatere medsebojne omejitve, ki jih računalnik lahko preprosto zazna (nesmiselna je nadgradnja pomnilnika na več kot 4 Gbyte, če ima 32-bitni operacijski sistem, nesmiselna je nadgradnja s hitrejšim pomnilnikom, kot ga matična plošča podpira itd.). S temi omejitvami lahko postavimo dodatna vprašanja za ocenjevanje višjih nivojev razumevanja.

Druga opcija je nadgradnja diska, kjer je možno izbrati magnetni ali pa elektronski disk (SSD – *Solid State Disk*). Če se študent odloči, da bi vgradil nov magnetni disk, bo dobil tri vprašanja: kakšno velikost diska namerava vgraditi, kako hitro se disk vrti (merjeno v vrtljajih na minuto – *RPM Revolution Per Minute*) in kakšno tehnologijo priklopa diska želi (SATA, mSATA, SAS itd.). Vsako vprašanje ima lahko pravilen in napačen odgovor glede na kombinacijo izbranih možnosti in tako se teoretično lahko zgodi, da bi bil vsak posamezni študentov odgovor sicer pravilen, ampak kombinacija pa ne. Pri SSD diskih lahko postavimo podobno število vprašanj (velikost, za kaj ga bo uporabljal, način priklopa, hitrost prenosa podatkov itd.). Koliko prostora vzame namestitev operacijskega sistema in programov, ki jih bo uporabljal, ter predvidevanje rasti podatkov na disku za časovno obdobje med dvema namestitvama operacijskega sistema. Seveda so te stvari zmeraj vezane tudi na ceno, ki bi jo lahko prav tako vključili v kakšno vprašanje.

Če pa se odloči študent za tretjo opcijo, nadgradnjo procesorja, pa mora poznati, kateri procesorji so podprti na njegovi osnovni plošči, kateri proizvajalec procesorja je podrt, ali je mogoče menjati procesor (lahko je procesor že nameščen na matični plošči), kakšno ima napajanje računalnika, ker nekateri procesorji zahtevajo zelo veliko električne energije za svoje delovanje. Lahko postavljamo tudi tehnološka vprašanja in vprašanja glede namena procesorja, ki naj bi ga nadgradili.

Iz napisanega je mogoče sestaviti shemo (ni popolna, je pa dobra za ponazoritev tega, kar smo prej pisali), ki pomaga razumeti potek ocenjevanja.



Slika 7: Napredno preverjanje znanja, nadgradnja računalnika
(Krašna, Repnik, Bratina in Kaučič 2012)

4.3.1 Razmišljanja o računalniških prilagodljivih testih

Računalniški prilagodljivi testi niso nekaj novega, so pa zelo redko uporabljeni v naših šolah. Ne moremo trditi, da jih učitelji ne uporabljajo, ker nimajo dovolj velikega znanja za njihovo pripravo. Ustno preverjanje znanja je prilagodljivo v sami osnovi in vsak učitelj ga obvlada. Problem je najverjetneje v tem, da še ne obstajajo preprosta orodja za izdelavo takšnih prilagodljivih testov. Očitno bo potrebno še nekaj časa, da bodo LMS orodja dobila module za izdelavo prilagodljivih testov in da bodo učitelji potem te teste tudi uporabljali. Ovir za izdelavo avtorskih orodij ni, ker so zahteve znane. Pri pripravi načrta za izdelavo naprednih prilagodljivih testov pa smo ugotovili še eno dodatno zanimivost. Neizpodbitno priprava prilagodljivih testov privede to spoznanja, kako pripraviti bolj primerna učna gradiva. Z opazovanjem sheme CAT vidimo analogijo z učnimi gradivi. Razlika je le ta, da pri preverjanju je dovolj, da študent pride po eni poti do cilja, pri učenju pa mora prehoditi vse poti.

5 Temeljne učne vsebine umetne inteligence v obveznem šolanju

Inteligentni agenti postajajo vedno boljši sogovorniki, nekateri med njimi so že opravili Turingov test. Roboti na delovnem mestu postajajo vedno pogostejši. V Severni Ameriki že preizkušajo popolnoma avtonomne avtomobile, tudi na področju zabave je umetna inteligenca vedno bolj prisotna, od predlaganja video vsebin do pisanja člankov v priznanih časopisih (GPT-3 2020). Posledično bo potreba po demistificiranju UI v izobraževanju vedno večja.

Vse večji prispevek tehnologij umetne inteligence k vsakdanjemu življenju in družbeni preobrati na obzorju z nadaljnjim razvojem teh tehnologij postavljajo vprašanje, ali bi morali temeljna znanja učnih vsebin umetne inteligence poučevati tudi v osnovni šoli. Če želimo, da se naši otroci informirano odločajo o vplivu UI na njihova življenja in družbo, je odgovor pritrdilen.

Vendar je za razumevanje osnovnih konceptov UI potrebno osnovno znanje računalništva, ki pa si tudi komaj utira pot v obvezne kurikulume obveznega šolanja. Situacija je po svoje paradoksalna, saj družba postaja vse bolj odvisna od računalniških tehnologij, hkrati pa je uvajanje računalniških temeljnih vsebin zelo počasno. Pozabljamo pa še na aktivno prebivalstvo, ki o vsebinah UI ni bilo poučeno, se pa mora o tem že odločati oz. delovati skupaj z UI tehnologijami. Poučevanje temeljnih vsebin UI bi zatorej morali izvesti celostno, kar pomeni, da bi morali sedaj nasloviti tako učence kot tudi aktivno prebivalstvo, kasneje pa je dovolj, da se poučujejo samo še učenci. Katere temeljne vsebine UI pa bi morali poučevati?

Touretzky in Gardner-McCune govorita o petih velikih konceptih (Touretzky in Gardner-McCune 2022, v tisku), ki so:

1. *zaznavanje*, računalniki zaznavajo svet z uporabo senzorjev,
2. *predstavitev in sklepanje*, agenti gradijo predstavitve sveta in jih uporabljajo za sklepanje,
3. Učenje, računalniki se lahko učijo iz podatkov,
4. naravna interakcija, inteligentni agenti potrebujejo veliko različnih vrst znanja, da lahko naravno medsebojno delujejo s človekom,
5. družbeni vpliv, UI lahko vpliva na družbo na dober in slab način.

Nekatere vsebine, povezane z UI, lahko začnemo poučevati že v nižjih razredih osnovne šole, kjer si lahko pomagamo z aktivnostmi, ki so dobile navdih pri gibanju računalništvo brez računalnikov (angl. *Computer science unplugged*). Avtorja Lindner in Seegerer sta pripravila 5 aktivnosti (Lindner in Seegerer 2020), ki jih izvedemo brez računalnika in nas skozi aktivnosti poučijo o nekaterih konceptih UI, kot so klasifikacija z odločitvenimi drevesi, okrepitveno učenje in delovanje Turingovega testa.

6 Zaključek

Prihodnji razvoj področja umetne inteligence v izobraževanju mora voditi k iterativnemu razvoju na učenca usmerjenega učenja, ki je podkrepljeno s podatki in je personalizirano. Ponovno velja poudariti, da umetna inteligenca ne bo nadomestila učiteljev. Pogosto citiran izrek Thornburga, da »vsak učitelj, ki ga lahko zamenja računalnik, si to zasluži«, je sporen, hkrati pa poudarja dejstvo, da trenutno ne obstaja nobena tehnologija, ki bi lahko posnemala, kaj šele izpodrinila, nešteto spretnosti in lastnosti odličnega učitelja. Pomembnost učiteljeve vloge še zdaleč ni postala obstranska s pojavom teh novih tehnologij. Obet umetne inteligence za učitelje je v njeni zmožnosti povečati učinkovitost njihovega poučevanja in jim pomagati pri zagotavljanju idealnih pogojev, v katerih se lahko njihovi učenci učijo in rastejo (Duggan 2020).

Umetna inteligenca bo učitelja osvobodila najbolj zamudnih in enoličnih nalog, kot sta ocenjevanje izpitov in preverjanje plagiatorstva v dokumentih. S pomočjo personaliziranih učnih vsebin in z umetno inteligenco kot učiteljevo asistentko, ki učitelju pomaga z odpravo pisanja zamudnih poročil, je lahko umetna inteligenca preobrazbena in osvobajajoča inovacija v izobraževanju.

Viri in literatura

Abdullah, S. C. in Cooley, R. E. (2002). »Using Simulated Students to Evaluate an Adaptive Testing System«. *International Conference on Computers in Education*. Auckland, Nova Zelandija.

Almohammadi, K., Hagras, H., Alghazzawi, D. in Aldabbagh, G. (2017). »A survey of artificial intelligence techniques employed for adaptive educational systems within e-learning platforms«. *Journal of Artificial Intelligence and Soft Computing Research*, 7, str. 47–64.

Barra, M., Iannaccone, A., Palmieri, G. in Scarano, V. (2002). »Test C++: An Adaptive Training System on the Internet«. *ISCC'02*. Taormina- Giardini Naxos, Italija.

Basu, A., Cheng, I., Prasad, M. in Rao, G. (2007). »Multimedia Adaptive Computer based Testing: An Overview«. *ICME 2007*. Beijing.

Bloom, B. (1984). »The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring«. *Educational Researcher*, 4–16.

Chen, S. in Zhang, J. (2008). »Ability Assessment based on CAT in Adaptive Learning System«. *Int. Workshop on Education Technology and Training & Int. Workshop on Geoscience and Remote Sensing*. Šanghaj, Kitajska.

Chen, J. in Wang, L. (2010). »Computerizd Adaptive Testing: A New Trend in Language Testing«. *International Conference on Artificial Intelligence and Education (ICAIE)*. Xi'an, Kitajska. URL = http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=05641509.

Cheng, I., Rodrigez, S. in Basu, A. (2009). »Multimedia and Games incorporating Student Modeling for Education«. *Int. Workshop on Technology for Education (T4E)*. Bangalode.

Costello, R. in Mundy, D. P. (2009). »The Adaptive Intelligent Personalised Learning Environment«. *9th IEEE Int. Conf. on Advanced Learning Technologies*.

Danielienė, R. in Telešius, E. (2008). »Analysis of Computer-Based Testing Systems«. *HSI 2008*. Krakow, Poljska.

De Praetere, T. (n.d.). »E-Learning: Learning throught the use of devices« (17. januar 2012), URL =http://knol.google.com/k/e-learning#.

Drigas, A. in Ioannidou, R. (2013). »A Review on Artificial Intelligence in Special Education. Information Systems, E-learning, and Knowledge Management Research«. *WSKS 2011*. Berlin: Springer.

Duggan, S. (2020). »AI in Education: Change at the Speed of Learning«. *Moscow: UNESCO IITE Policy Brief*.

Fan, O. in Pengcheng, J. (2021). »Artificial Intelligence in Education: The Three Paradigms«. *Computers and Education: Artificial Intelligence*, 2, 100020.

Garg, S. in Sharma, S. (2020). »Impact of Artificial Intelligence in Special Need Education to Promote Inclusive Pedagogy«. *International Journal of Information and Education Technology*, 10(7), str. 523–527.

Gerlič, I. (2000). *Sodobna informacijska tehnologija v izobraževanju*. Ljubljana: DZS.

GPT-3. (2020). »A robot wrote this entire article. Are you scared yet, human?« *The Guardian*. URL = https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3.

Guzmán, E. in Conejo, R. (2005). »Self-Assesment in a Feasible, Adaptive Web-Based Testing System«. *IEEE Transactions on Education*, 48(4), str. 688–695.

Haage, M., Piperagkas, G., Papadopoulos, C., Mariolis, I., Malec, J., Bekiroglu, Y., … Tzovaras, D. (2017). »Teaching Assembly by Demonstration Using Advanced Human Robot Interaction and a Knowledge Integration Framework«. *Procedia Manufacturing*, 11, str. 164–173.

Hatzilygeroudis, I., Koutsojannis, C. in Papavlasopoulos, C. (2006). »Knowledge-Based Adaptive Assesment in a Web-Based Intelligent Educational System«. *6th Int. Conf. on Advanced Learning Technologies (ICALT'06)*. Kerkrade, Nizozemska.

Krašna, M. (2010). *Multimedija v izobraževanju*. Nova Gorica: Educa.

Krašna, M., Repnik, R., Bratina, T. in Kaučič, B. (2012). »Advanced types of electronic testing of student's performance«. *MIPRO 2012*. Opatija.

Krašna, M. (2015). *Izobraževanje v digitalnem svetu*. Maribor: Zora.

Kurilovas, E. (2019). »Advanced machine learning approaches to personalise learning: learning analytics and decision making«. *Behaviour & Information Technology*, 38(4), str. 410–421.

Lindner, A. in Seegerer, S. (2020). *AI Unplugged - Unplugging Artificial Intelligence*. Erlangen: Friedrich-Alexander-Universität Erlangen-Nürnberg.

Pestano Perez, M., Pesek, I., Zmazek, B. in Lipovec, A. (2020). »Video explanations as a useful digital source of education in the COVID 19 situation«. *Revija za elementarno izobraževanje*, 13(4).

Robson, K. (2010). »RSA ANIMATE: Changing Education Paradigms«. *Youtube*. URL = https://www.youtube.com/watch?v=zDZFcDGpL4U.

Song, J., Chen, W. in Gao, D. (2011). »The Adaptive On-line Exam System based on Agent. Int.«. *Conf. on Future Computer Science and Education*. Xi'an, Kitajska.

Touretzky, D. in Gardner-McCune, C. (2022, v tisku). »Artificial Intelligence Thinking in K-12«. V S.-C. Kong in H. Aberson (urd.), *Computational Thinking in K-12: Artificial Intelligence Literacy and Physical Computing*. Cambridge, Mass.: MIT Press.

Woolf, B. P. (1990). *AI in Education*. New York: John Wiley & Sons. URL = https://web.cs.umass.edu/publication/docs/1991/UM-CS-1991-037.pdf.

Xie, H., Hwang, G.-J. in Wang, C.-C. (2019). »Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017«. *Computers & Education*, 140, 103599.

X5Gon. (2021). Dostopno na X5Gon. URL = https://www.x5gon.org/.

UMETNA INTELIGENCA IN PRIHODNOST UČENJA IN POUČEVANJA

BORIS ABERŠEK, ANDREJ FLOGIE

Univerza v Mariboru, Fakulteta za naravoslovje in matematiko, Maribor, Slovenija
boris.abersek@um.si, andrej.flogie@um.si

Sinopsis Umetna inteligenca (UI), strojno učenje, informacijsko-komunikacijske tehnologije (IKT) in sorodne računalniške tehnologije vplivajo na družbo kot celoto, kakor tudi na izobraževanje in prihodnost učenja in poučevanja. Skokovit razvoj učnih okolij, ki je podprt z učno tehnologijo, in nova spoznanja s področja kognitivne teorije, nevroznanosti ter UI predstavljajo velik izziv tako za izobraževalni sistem in razvijalce učnih okolij kot tudi za učitelje in učence. Izpostaviti je treba, da naloga šole ni le razvoj kognitivnih kompetenc učencev, ampak tudi razvoj socialnih kompetenc, kar je velik izziv, saj so socialne interakcije med mladostniki sestavni del njihovega zdravega psihosocialnega razvoja. Izpostavili bomo pozitiven doprinos sodobnih učnih okolij za sodelovanja med učenci ter med učenci in njihovimi učitelji in analizirali pomembne prednosti in slabosti UI ter ustrezne priložnosti in ovire za uporabo UI pri učenju in poučevanju ter ob tem poskušali analizirati, kaj smemo in česa ne (etične dileme).

Ključne besede:
inovativna pedagogika,
umetna inteligenca,
strojno vedenje,
inovativna učna okolja,
etika

Univerzitetna založba Univerze v Mariboru

DOI https://doi.org/10.18690/um.ff.11.2022.14
ISBN 978-961-286-675-4

ARTIFICIAL INTELLIGENCE AND THE FUTURE OF TEACHING

BORIS ABERŠEK, ANDREJ FLOGIE
University of Maribor, Faculty of Natural Science and Mathematics, Maribor, Slovenia
boris.abersek@um.si, andrej.flogie@um.si

Abstract Artificial intelligence (AI), machine learning, information and communication technologies (ICT), and other emerging technologies affect society as a whole, including the sphere of education and the future of learning and teaching. The rapid development of learning environments, supported by learning technology and coupled with new insights from the fields of cognitive science, neuroscience, and AI, is a major challenge not only to the education system and to creators of learning environments, but also to teachers and students in general. The task of the school is not only the development of students' cognitive competences, but also the development of their social competences. In this article, we will highlight the positive effects of student-student and student-teacher collaboration and an analysis of the advantages and disadvantages of AI, opportunities and obstacles of using AI in learning and teaching, and the ethical dilemmas related to such use.

Keywords: innovative pedagogy, artificial intelligence, machine behaviour, innovative learning environment, ethics

https://doi.org/10.18690/um.ff.11.2022.14 DOI
978-961-286-675-4 ISBN
University of Maribor Press

1 Uvod

Naravoslovna znanost je znanje o naravnih objektih in pojavih. Sprašujemo se lahko, ali ne more obstajati tudi 'umetna' znanost – znanje o umetnih predmetih in pojavih.

Herbert Simon

Umetna inteligenca (UI), strojno učenje in sorodne računalniške tehnologije vplivajo na družbo kot celoto kakor tudi na izobraževanje in prihodnost učenja. Vpliv tehnologije na izobraževanje je pogosto podvržen negativnim odnosom in strahovom, ne glede na to, ali so učinki predvidljivi ali ne. Zaradi pospeševanja integracije tehnologije v učna okolja in v procese učenja in poučevanja, kar je še posebej očitno v zadnjih mesecih s pojavom pandemije, ni več poti nazaj, saj sodobne tehnologije omogočajo vsaj okrnjen način izobraževanja. Zdaj je čas, da začnemo načrtovati, kako najbolje razviti in uporabljati UI v izobraževanju na načine, ki so efektni, pravični do posameznika, etični in učinkoviti ter ob tem ublažiti ali celo nevtralizirati slabosti, tveganja in morebitno škodo.

V prispevku bomo poskušali izpostaviti in osmisliti tri nivoje, ki lahko uokvirijo pomen UI za izobraževanje:

- *prvič*, umetno inteligenco je mogoče šteti za 'obveščevalno inteligenco' in sposobnost, ki jo lahko prenesemo na izobraževalne izzive kot dodaten vir pomoči učiteljem;
- *drugič*, UI prinaša posebne, vznemirljive, nove zmogljivosti sodobnim učnim okoljem, vključno s prepoznavanjem čustev (angl. *emotional recognition*), prepoznavanjem vzorcev (angl. *face recognition*), predstavljanjem znanja, načrtovanjem ter podpiranjem naturalističnih interakcij z ljudmi in nenazadnje učno analitiko. Te specifične zmogljivosti se lahko izražajo v rešitvah za podporo učencem z raznolikimi potrebami (individualizacija in personalizacija). Poleg bolj tradicionalnega vnosa s tipkovnico, miško ali računalniškim peresom omogočajo učencem, da pretvarjajo rokopis v tekst, omogočajo vnos ukazov in podatkov s kretnjami ali govorom ipd.;
- *tretjič*, UI se lahko uporablja kot orodje, ki omogoča, da si predstavljamo, študiramo in razpravljamo o prihodnosti učenja, da razvijamo algoritme, ki danes še ne obstajajo. Strokovnjaki so enotnega mnenja, da najbolj vplivne uporabe UI v izobraževanju danes še niso niti razvite. V prispevku bomo

analizirali pomembne prednosti in slabosti UI ter ustrezne priložnosti in ovire za uporabo UI pri učenju in poučevanju in ob tem poskušali analizirati, kaj smemo in česa ne.

Posebno pozornost bomo posvetili predvsem individualizaciji, personalizaciji in socialnemu učenju. V tem scenariju bo UI podpirala več vrst dejavnosti učnih partnerjev in vzorcev interakcij, ki lahko obogatijo učno okolje (realno ali virtualno učilnico). Ta koncept se razlikuje od številnih, trenutno znanih, oblik učnih okolij, ki se osredotočajo predvsem na izoliranega posameznika (individualizacija in personalizacija), ki sodeluje samo z eno samo napravo. V sodobnih učnih okoljih moramo zagotoviti, da UI zagotovi tudi podporo sodelovalnemu učenju, skupini študentov, ki dela na skupnem projektu ali skupni nalogi, kar vključuje podporo študentom, da delajo kot člani skupine (drug drugega poslušajo in gradijo na prispevkih drug drugega) in podporo za naloge, ki jim pomagajo organizirati, upravljati in povezati njihove prispevke s skupnim ciljem skupine. Na nekem elementarnem, simbolno zapisanem sistemu, kot sta npr. MS Teams ali Zoom, delamo že danes pri študiju na daljavo. Sodobna inteligentna učna okolja pa bi bila adaptivna, prilagajala bi se lahko temu, kar skupine potrebujejo za dobro sodelovanje kot tudi načinu prehoda skupin med posameznim delom, delom malih skupin in razpravami s celotno učilnico. Ta vrsta UI sistema naj bo socialno ozaveščena in lahko uporablja socialno interakcijo s študenti ali člani projektnega tima kot način spodbujanja uspešnosti. V prispevku bomo izpostavili in poskušali analizirati šest prednostnih področij raziskav v prihodnosti:

1. analiza modelov UI za razširjen nabor učnih scenarijev,
2. razvoj sistemov UI, ki pomagajo učiteljem in izboljšajo poučevanje,
3. intenzivnost in širitev raziskav o UI za ocenjevanje učenja, npr. različne učne analitike,
4. pospeševanje razvoja človeških ali odgovornih UI,
5. etično rabo UI in
6. krepitev splošnega ekosistema UI in izobraževanja.

Naša razmišljanja v tem prispevku bodo temeljila na spoznanjih, ki se pojavljajo na presečišču področij filozofije (etike), umetne inteligence (UI) in izobraževanja. Izhajali bomo iz dognanj avtorjev, kot so Turing (1950), Bostrom (2014), Arnold, Kurzweil (2005) in drugi. V razmišljanje bomo vključevali aktivnosti, ki jih je pričela Evropska komisija leta 2018 na področju etike v UI z ustanovitvijo Evropske UI

zveze (*Europe AI Alliance*) in so zapisane v dokumentih »Artificial Intelligence for Europe (COM 237 2018)« v »Coordinated Plan on Artificial Intelligence (COM 759 2018)«. Razmišljali bomo o digitalni pismenosti/kompetencah, v okviru katerih mora biti posameznik

- na eni strani zmožen uporabljati, razumeti in vrednotiti tehnologijo,
- na drugi strani pa poznati tehnološke principe in postopke, ki so potrebni za razvoj tehnoloških rešitev in doseganje ciljev družbe s pomočjo tehnologije. Pri tem se zastavi vprašanje: *Koliko znanj potrebuje vsak posameznik (splošno izobraževanje) za doseganje ciljev družbe?*

Digitalna pismenost/kompetence so strukturirana zmožnost, sestavljena iz treh ključnih področij. Gre za kompetence s področja (Kordigel, Aberšek in Aberšek 2020):

1. sistemskega razmišljanja in ustvarjanja,
2. informacijsko-komunikacijske tehnologije in umetne inteligence ter
3. tehnologije in družbe.

Področje *sistemskega razmišljanja in ustvarjanja* pokriva tako znanje s področja naravoslovnih ved kot tudi znanje s področja tehniških ved, ki so ključna za ustvarjanje tehnoloških rešitev in za razumevanje osnovnih principov upravljanja sodobnih tehnologij v vsakodnevnem življenju. Področje *informacijsko-komunikacijske tehnologije in umetne inteligence* pokriva računalnike in programsko opremo, omrežja in protokole, mobilne naprave in ostale tehnologije, s katerimi dostopamo do informacij in znanja ter s pomočjo katerih ustvarjamo novo znanje.

Tretje področje, področje *tehnologije in družbe,* pokriva zmožnost poznavanja in kritičnega in etičnega vrednotenja vpliva, ki ga ima tehnologija na družbeno in naravno okolje. Posebej pomembna je v tem kontekstu zmožnost zastavljanja etičnih vprašanj in etičnega presojanja odgovorov na vprašanja*,* povezana z vplivi (neodgovorne/sporne) rabe tehnologije na družbeno in naravno okolje. Znanje in kompetence s področja *tehnologije in družbe* so ključne za razumevanje problematike razvoja in uporabe tehnologije ter za sprejemanje odločitev v zvezi z uporabo te tehnologije, zato bomo prav temu segmentu posvetili posebno pozornost. Etika in UI sta tako povezani na več ravneh:

1. 'etika z ustvarjanjem': tehnična/algoritmična integracija etičnih zmožnosti sklepanja kot dela znanja umetnega avtonomnega sistema;
2. 'etika v ustvarjanju': regulativne in inženirske metode, ki zagotavljajo analizo in oceno etičnih posledic delovanja sistemov UI;
3. 'etika za ustvarjanje': kodeksi ravnanja, standardi in postopki certificiranja, ki zagotavljajo integriteto razvijalcev in uporabnikov, ko raziskujejo, oblikujejo, uporabljajo in upravljajo avtonomne sisteme.

2 Družba in tehnologija

Tehnologija je dvoje, je vzrok in posledica za vse hitreje spreminjajočo se družbo. Ko govorimo o tehnologiji, govorimo o inteligentnih sistemih, ki so lahko fizični (kibernetski fizični sistemi) ali samo kognitivni, nefizični sistemi, torej UI, ki nima fizične oblike in se nahaja znotraj nekega sistema, npr. interneta ali inteligentnih učnih okolij. Pri inteligentnih sistemih (naj bodo to živalski, mehanski, ekonomski, socialni, človeški ipd.) sistem vedno vsebuje *programsko opremo*, software – inteligenco – duh in tudi mehanizem povratnih informacij (sistem dobiva informacije o tem, kaj se v njem in okoli njega dogaja), ki se na podlagi teh informacij uči ter ukrepa v novonastalih okoliščinah. V jeziku humanistike bi to lahko poimenovali tudi *razum, etika, zavest, hotenje* ... Obstoj tega mehanizma povratnih informacij je osnova in osnovna zahteva za preživetje sistema. Sistem pa lahko ob tem, če govorimo v jeziku tehnologije, vsebuje tudi *strojno opremo* − hardware, stroje in naprave (analogija s človeškim ali živalskim telesom), skratka hardware, ko govorimo o kibernetskih fizičnih sistemih. Kibernetski pristop dokazuje povezanost sistema v celoto, dokazuje, da sprememba v enem delu sistema povzroča spremembe, tako pričakovane kot tudi nepričakovane, v drugih delih sistema.

Čeprav morda lahko celo razumemo povezavo med tehnologijo in družbo, pa kljub temu nismo sposobni v celoti vplivati na svojo prihodnost. Na dogodke se večinoma le odzovemo, ne moremo pa jih vedno usmerjati. Na področju povezav med tehnologijo in družbo obstaja vrsta popolnoma diametralnih teorij. Jacques Ellul in Herbert Marcus menita, da postaja tehnologija vse bolj človekov gospodar in vse manj njegovo orodje. V nasprotju s tem pa drugi, kot npr. Lynn While, menijo, da je tehnologija nevtralna, skratka, da *tehnologija odpira vrata, vendar se človek lahko odloča, ali bo vstopil ali ne!* (Aberšek, Borstner in Bregant 2014). Ta pogled takoj poraja naslednja vprašanja:

- Kdo odloča, katera vrata moramo odpreti?
- Ko in če vstopimo, ali tehnologija določa obliko sobe, v katero smo vstopili?
- Če je tehnologija preprosto skupek zamisli, se postavi še naslednje vprašanje: Kdo določa cilje in ali ne nazadnje ne obstaja nevarnost, da zamisli same postanejo cilji?

Vsa ta vprašanja niso le zgodovinska vprašanja, to so tudi realni problemi današnjih dni, kar je razvidno tudi iz smeri in trendov razvoja in kontrole vseh tehnologij.

Mnogi problemi, ki so povezani s sodobnimi tehnologijami, so povezani tudi s strojno oz. umetno inteligenco (UI). Za reševanje teh problemov potrebujemo splošna znanja (in odgovore) o svetu, družbi, ljudeh, torej je to *globalen filozofski problem*. Zato je za vse, ki morajo sprejemati odločitve s področja UI, izjemno pomembno, da *razumejo* osnovne splošne mehanizme inteligence – filozofijo duha, tj. kako deluje narava, kako deluje človeška inteligenca in kaj to sploh je.

Na področju filozofije duha, vse od najzgodnejših začetkov pojava človeka kot mislečega 'stroja' pa vse do danes, je prisoten osnovni razkorak razlag, ki bi jih ekstremno lahko poenostavili na dve struji (Horst 2007):

- *misterianistično*, ki zagovarja tezo, da je duh nekaj edinstvenega v svoji neponovljivosti, kjer je misterij v nekih lastnostih, svojstvenih samo študijam duha in duhu samemu, kot npr. t. i. 'misterianizem'[1], ki ga je razvijal Colin McGinn (1999), in
- *naturalistično,* ki zagovarja tezo, da duh lahko ali ga celo moramo 'naturalizirati', in to tako, da so lahko mentalna stanja in procesi opisljivi z jezikom znanosti: fizike, nevroznanosti ali jezikom drugih naravoslovnih ved.

Oba pristopa skušata najti resnico, vendar uporabljata popolnoma različne metode in orodja ter interpretirata svoje dosežke na popolnoma različnih osnovah in izhodiščih. Če želimo poudariti to misel, lahko povzamemo misli A. Einsteina, ki je trdil, da sta znanost (objektivnost) in duhovnost (subjektivnost) druga drugi komplementarni in zato potrebujemo obe. Zato filozofija uporablja obe, znanost in

[1] *Misterianizem* je filozofska pozicija, ki zagovarja tezo, da je *težki problem zavesti* (ang. *hard problem of consciousness*) onkraj meja naše razumljivosti.

duhovnost, objektivne in subjektivne izkušnje, z namenom doseči uravnoteženost in harmonijo med racionalnim in intuitivnim razumom, med glavo in srcem. Resnična klasična filozofija se ubada z vprašanji, kako deluje vesolje, zakaj vesolje obstaja in kaj je smisel življenja. Razlika med znanstveniki, poimenujmo jih empirični znanstveniki (naravoslovci, tehniki …), in filozofi, ni v vsebini, temveč je v metodah in načinu razmišljanja. Za razliko od empiričnih znanosti, kjer je poudarek na opazovanju, zbiranju gradiva in klasifikaciji ter na izvajanju in interpretiranju eksperimentov ali razvijanju novih pristopov in sistemov, pristop filozofov temelji predvsem na *argumentiranju*, *pojmovni analizi* in *zgodovinskem vidiku* (Markič 2010). Filozofi se sprašujejo o temeljnih predpostavkah (metafizičnih, epistemoloških in metodoloških), na katerih znanstveniki postavljajo hipoteze in načrtujejo eksperimente. Ob tem poskušajo podati sintezo različnih pristopov in sestavljajo pregled področja znanosti.

Raziskovalci s področja filozofije, kognitivne znanosti, nevrobiologije, nevroračunalništva in umetne inteligence se pogosto sprašujejo, ali in kakšna je zveza med razvojem kognitivnega nevromodeliranja (simuliranja delovanja možganov s pomočjo sodobnih računalniških tehnologij) in nevroračunalništva (računalniške simulacije nevronskih mrež oz. t. i. umetnih nevronskih mrež, ki v kognitivni znanosti prevzemajo vlogo procesnih modelov možganskih in duševnih procesov). Posplošeno, sprašujejo se o korelaciji človek – stroj in to glede prenosa idej, navdiha za modele ter o primernosti/uporabnosti teh modelov za najrazličnejše namene. S terminom 'stroj' bomo poimenovali vsak sistem, ki v sodobnem svetu skuša nadomestiti človeka, tako človeka kot fizično bitje (npr. stroj kot humanoidni robot, kibernetski fizični sistem, ki lahko izvaja različne fizične aktivnosti …), predvsem pa človeka kot duševno, mentalno, nefizično razmišljajoče bitje (npr. računalnik, ki simulira človeške kognitivne procese).

Do sedaj še ni bil podan natančen odgovor na tri, za raziskave na teh področjih bistvena, vprašanja:

- Ali lahko celovito kognitivno nevromodeliramo – simuliramo delovanje možganov (glede korelacije med možgani in duhom) s pomočjo sodobnih računalniških tehnologij?
- Ali lahko razvoj nevroračunalništva in umetne inteligence pripelje do nadomeščanja človeka in njegove naravne inteligence (*neposredno* –

humanoidni robot – stroj, ki opravlja človeško delo, ali – *posredno* – računalniški 'učitelj', inteligentni tutor, ki opravlja le človekove mentalne funkcije)?

- Ideja, ki se ponovno bolj intenzivno pojavlja v zadnjih desetletjih, ali lahko človeka (duh in telo) obravnavamo kot dinamični sistem, skratka ali obstaja takšen sistem diferencialnih enačb, s katerimi bi lahko to zapisali.

Izhajajoč iz teh treh vprašanj bi lahko formulirali še četrto:

- Če je to vse možno, kakšna je korelacija med kognitivnim nevromodeliranjem in nevroračunalništvom?

2.1 Umetna inteligenca

Dolgoročni cilj raziskav s področja strojnega učenja, ki je danes še vedno videti nedosegljiv, je ustvariti umetni sistem, ki bo s samostojnim učenjem dosegel ali celo presegel človeško inteligenco. Širše področje raziskav s tem istim ciljem pa na kratko poimenujemo *umetna inteligenca UI (angl. artificial intelligence)*. Omenimo eno od izhodiščnih splošnih definicij, povzeto po Marvinu Mynskyju: »Umetna inteligenca je *znanost o izdelavi strojev*, ki so sposobni narediti stvari, za katere je po naših merilih potreben um« (Minsky 1969). Minsky je ustanovitelj laboratorija za UI na MIT-ju, v Copelandu 2007: 1) Če govorimo o umetni inteligenci, od inteligentnega sistema (v tem primeru računalnika, stroja, robota) ne pričakujemo, da je ekstremno inteligenten (izjemno sposoben) samo v enem vidiku inteligence. Pričakujemo, da je kompleksen, inteligenten na vseh področjih, ki jih zahteva človeška inteligenca pri reševanju problemov. Raziskave s področja UI se ubadajo z razvojem sistema, ki deluje bolj ali manj inteligentno in je sposoben reševanja relativno zahtevnih problemov. Te metode mnogokrat temeljijo *na osnovi posnemanja človeškega reševanja problemov*. UI področja, razen strojnega in globokega učenja, so povezana s predstavljanjem znanja, razumevanjem govora, avtomatskim sklepanjem in dokazovanjem teoremov, logičnim programiranjem, kvalitativnim modeliranjem, ekspertnimi sistemi, igranjem iger, hevrističnim reševanjem problemov, umetnimi zaznavami, roboti in kognitivnim modeliranjem. Tu se osredotočimo le na splošne probleme, povezane z UI (Stone et al. 2016).

V vseh UI področjih igrajo pomembno vlogo algoritmi strojnega učenja. Praktično mora biti povsod prisotno učenje, samoučenje. Z uporabo tehnik učenja se lahko sistem uči in izboljšuje svoje zaznave (postaja adaptiven), izboljšuje svoje razumevanje govora, zmožnost sklepanja ipd. Področje logičnega programiranja je prav tako v tesni zvezi z induktivnim logičnim programiranjem, ki se uporablja pri razvoju logičnih programov, npr. za določanje ciljev (npr. GPS naprave). Za razvoj ekspertnih sistemov lahko uporabljamo strojno učenje za ustvarjanje podatkovnih baz iz primerov usposabljanja reševanja problemov. Inteligentni roboti morajo neizogibno izboljšati svoj postopek za reševanje problemov s pomočjo učenja. Končno, tudi kognitivno modeliranje se praktično ne more izvajati brez upoštevanja učnih algoritmov. Danes pa se vse pogosteje zastavljajo predvsem etični problemi, ne več, kako kaj narediti, ampak predvsem, kakšne posledice bodo imele naše aktivnosti na celoten naravni in tudi družbeni sistem. Tako kot vse pogosteje govorimo o razvoju in spreminjanju človeškega vedenja (obnašanja), moramo pričeti razmišljati tudi o vzporednih inteligentnih entitetah (UI) in njihovem vedenju. To so strokovnjaki iz MIT-ja poimenovali »strojno vedenje« (Rahwan et al. 2019) .

2.2 Strojno vedenje

Če sodobna kognitivna nevroznanost v zadnjih desetletjih nezadržno napreduje in o delovanju človeške kognicije in delovanju človeških možganov oz. človeške inteligence (ČI) vemo vedno več, pa je področje strojnega vedenja področje, ki spodbuja kognitivno znanost, da bi poskušala razumevati UI in UI agente. Trenutno so znanstveniki, ki najpogosteje preučujejo vedenje strojev, računalniški znanstveniki in inženirji, ki so stroje ustvarili. Ti pa običajno niso izurjeni na področju kognitivnih (vedenjskih) znanosti (Rahwan et al. 2019). Podobno, čeprav vedenjski znanstveniki razumejo te inženirske discipline, pa nimajo strokovnega znanja za razumevanje učinkovitosti določenega algoritma ali kibernetskega fizičnega sistema. Da bi dosegli celovito razumevanje vedênja UI agentov, bi morali védenje o strojnem vedênju postaviti na križišče računalništva, inženirstva in vedenjskih ved. Ker postajajo UI agenti vedno bolj sofisticirani in kompleksni, bo za analizo njihovega vedenja potrebna kombinacija razumevanja njihove notranje arhitekture (domene računalniških znanstvenikov) v interakciji in sodelovanju z drugimi agenti (inteligenca roja) in njihovim okoljem (domena vedenjskih znanstvenikov). Medtem ko bo nekdanji vidik razumevanja notranje arhitekture še vedno funkcija tehnik optimizacije globokega učenja, se bo interakcija inteligentnih agentov in njihovih okolij morala zanašati pretežno na vedenjske, kognitivne vede.

Pri razvoju nove transdisciplinarne znanosti, ki jo imenujemo *znanost o vedenju UI,* lahko izhajamo iz dela Nikolaasa Tinbergena (1969), ki obravnava prepoznavanje ključnih razsežnosti vedenja živali. Tinbergenova teza pravi, da obstajajo za razumevanje vedenja živali (in tudi človeka) štiri dopolnjujoče dimenzije, mehanizem, razvoj, funkcija in evolucija. Če te štiri dimenzije prenesemo na področje umetnih sistemov, lahko zapišemo (Rahwan et al. 2019):

1. **Mehanizem:** Mehanizmi za ustvarjanje vedenja UI agentov temeljijo na njihovih algoritmih in značilnostih okolja, ki jih usmerja.
2. **Razvoj:** Vedenje agentov UI se s časom razvija, se uči, je adaptivno. Strojno vedenje tako preučuje, kako stroji pridobivajo (razvijajo) specifično individualno ali kolektivno vedenje, kako se razvija inteligenca roja[2].
3. **Funkcija:** Razumevanje, kako specifično vedenje vpliva na življenjsko funkcijo UI agenta.
4. **Evolucija:** UI agenti so tudi ranljivi zaradi evolucije in interakcije z drugimi agenti. Ponovno jih je mogoče uporabiti v novih kontekstih, tako za omejevanje prihodnjega vedenja kot tudi za omogočanje dodatnih izboljšav.

Kljub temeljnim razlikam med UI in človeško inteligenco si lahko strojno vedenje izposodi nekatere Tinbergenove zamisli za oris glavnih vrst vedenja pri UI agentih. UI ima *mehanizme,* ki proizvajajo vedenje, se razvijajo, hkrati pa v svoje vedenje vključujejo okolijske informacije, proizvajajo funkcionalne posledice, ki povzročajo, da specifični stroji postanejo bolj ali manj pogosti v specifičnih okoljih, in utelešajo *evolucijo* zgodovine, skozi katero pretekla okolja in človeške odločitve še naprej vplivajo na vedenje UI. Prilagoditev Tinbergenovega okvira vedenja strojev je shematično predstavljeno na sliki 1.

Štiri dimenzije Tinbergena (1969) zagotavljajo celovit model razumevanja obnašanja UI agentov. Vendar pa te štiri dimenzije ne veljajo na enak način glede tega, ali ovrednotimo klasifikacijski model z enim zastopnikom ali z več sto zastopniki. V tem smislu vedenje strojev uporablja prej omenjene štiri dimenzije v treh različnih lestvicah:

[2] Inteligenca roja je »disciplina, ki se ukvarja z naravnimi in umetnimi sistemi, sestavljenimi iz številnih posameznikov, ki se usklajujejo z uporabo decentraliziranega nadzora in samoorganizacije« (Dorigo, Birattari, 2007).

1. Prvi je *individualno strojno vedenje*: ta dimenzija vedenja poskuša preučiti vedenje posameznih UI agentov. Obstajata dva splošna pristopa k preučitvi vedenja posameznih UI agentov. Prvi se osredotoča na profiliranje nabora vedenja katerega koli specifičnega stroja z uporabo pristopa znotraj stroja, pri tem pa primerja vedenje določenega stroja v različnih pogojih. Drugi pristop preučuje, kako se različni posamezni stroji obnašajo pri enakih pogojih (Aberšek 2018).
2. Druga lestvica je *kolektivno vedenje strojev*: za razliko od posamezne dimenzije to področje poskuša razumeti vedenje UI agentov s preučevanjem interakcij UI agentov v skupini. Kolektivna dimenzija vedenja stroja poskuša opazovati vedenja UI agentov v njihovi interakciji.
3. In končno, lestvica opazuje hibridno *vedenja človek-stroj*: obstaja veliko scenarijev, v katerih na vedenje UI agentov vpliva njihova interakcija z ljudmi. Ta dimenzija vedenja stroja se osredotoča na analizo vedenjskih vzorcev pri UI agentih, ki jih sproži interakcija z ljudmi.

Tabela 1: Tinbergen je predlagal, da se študija vedenja živali lahko prilagodi študiji vedenja strojev (Tinbergen 1969; Rahwan et al. 2019)

Vrsta razlage	Predmet študije	
	Dinamični pogled Zgodovinski (evolucijski) pogled	**Statični pogled** Trenutno vedenje stroja
Približen pogled posameznih vrst funkcij stroja	**Razvoj (ontogeneza)** Razvojna razlaga, kako stroj (UI) pridobi svojo različno vrsto vedenja z učenjem v določenem okolju.	**Mehanizem (vzročnosti)** Mehanična razlaga, kakšno je vedenje in kako je zgrajeno.
Končni (evolutivni) pogled Zakaj individualna vedenja strojev, kot je	**Evolucija (filozofija)** Sile, ki opisujejo, zakaj se je vedenje razvijalo in širilo.	**Funkcija (prilagajanje)** Posledica obnašanja strojev v trenutnem okolju.

2.3 UI in Izobraževanje

Izpostavili smo, da je temelj vsake inteligence učenje, ki vodi do potrebnih (želenih) sprememb. Ko govorimo o učenju, istočasno govorimo tudi o metodah učenja, ki jih na najbolj generalnem nivoju poimenujemo pedagogika, na bolj posplošenem pa didaktika in predmetna didaktika. Kako se uči UI danes vemo, pojavi pa se vprašanje, ali so metode učenja UI *pedagoško pravilne* ali morda potrebujemo neke popolnoma drugačne pristope k poučevanju UI k metodam učenja, ki so implementirane v različne UI algoritme? Izhajajmo iz premise, da človek ustvarja novo 'bitje', ki se bo

skozi proces učenja spreminjalo in razvijalo. Vsi vemo, da je pri vzgoji bioloških 'bitij', pa naj bodo to ljudje ali živali, najpomembnejša najbolj zgodnja faza razvoja, ki v veliki meri generira nadaljnji razvoj.

Problem:

Če tako govorimo o razvoju (odraščanju) UI, se pedagoško gledano to orientira (pretežno) le na kognitivno področje, zanemarjajo pa se vsi ostali aspekti vzgoje in izobraževanja. Oglejmo si to na enem od najstarejših primerov, na »*vzgojnih zakonih - strojnem vedenju*« robotike (UI), ki jih je postavil že Isaac Asimov (1950). Trije zakoni so:

1. *Robot ne sme raniti človeškega bitja ali z nedelovanjem omogočiti, da se človek poškoduje.*
2. *Robot mora ubogati ukaze, ki mu jih je dal človek, razen če bi bil takšen ukaz v nasprotju s prvim zakonom.*
3. *Robot mora zaščititi svoj obstoj, vse dokler takšna zaščita ni v nasprotju s prvim ali drugim zakonom.*

Izhajajoč iz tega pridemo do temeljne pedagoške dileme sedanjosti. Postavili smo se v vlogo ustvarjalca, kreatorja 'novega življenja', umetnega življenja. Vemo, da se pedagoška doktrina vzgoje in izobraževanja (šole) skozi zgodovino stalno spreminja. Zdaj je napočil čas, da se pedagoške in didaktične metode (zakoni za UI) napišejo na novo tudi za UI. Ali bi morali razviti specifično kognitivno znanost posebej za UI in zanjo urediti tudi primerno izobraževanje, ustrezno šolo? In ali bi pri ustvarjanju te kognitivne znanosti in šole morali vključiti tudi UI samo? Ali se mora pri razvoju kognitivne znanosti in šole za UI upoštevati poleg kognitivne tudi socialna komponenta izobraževanja ob upoštevanju nekakšnih etičnih (človeških) norm pri razvoju UI aplikacij. Ali in do kakšne mere lahko uporabljamo UI v učnem procesu ljudi ali kako bi morala biti organizirana 'inteligentna učna gradiva' in inteligentna učna okolja? Pred nami je ogromno, za človeško družbo pomembnih, vprašanj, časa pa je izjemno malo, saj se UI uči in spreminja za človeške pojme s svetlobno hitrostjo in vse bolj postaja samostojna, avtonomna. Ljudje ji prepuščamo vse več področij odločanja in imamo nad njo vse manj nadzora.

3 UI in učna okolja

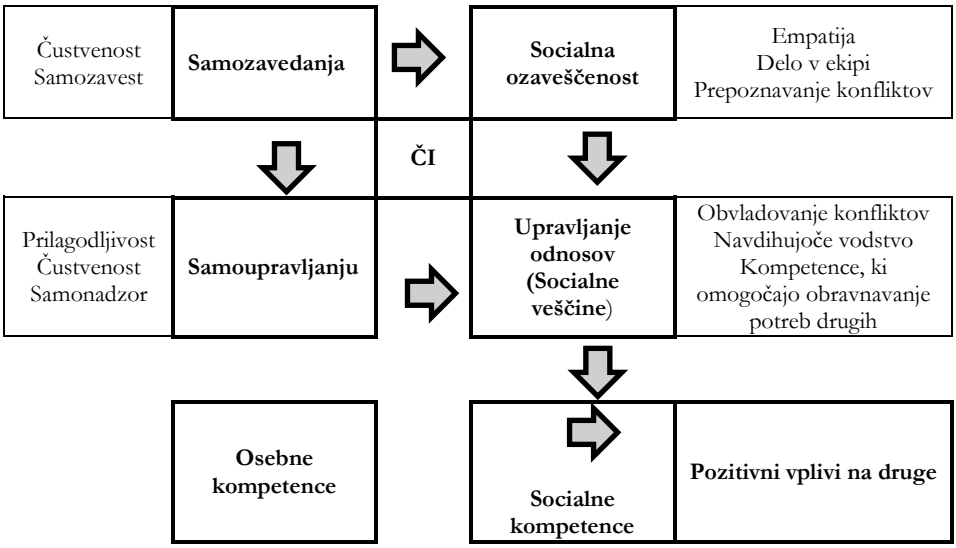
Problem sodobne družbe je, da mora šolski sistem mladostnike usposabljati za življenje in jim dajati ne le znanja in različnih veščin, temveč jih mora naučiti predvsem soočati se z vsakodnevnimi izzivi in problemi ter jih ob tem naučiti reševati te probleme. Pri mladostnikih se morajo razvijati kognitivne kompetence, razvijati pa moramo tudi sodelovalnost in socialne kompetence, saj je to eden izmed temeljnih pogojev za vseživljenjsko učenje in izboljšano zaposljivost. Pogoj za to pa je udejanjanje prožnih oblik učenja (Flogie in Aberšek 2019: 2017).

Značilnosti današnjih generacij, opredeljenih z različnih vidikov (od sociološkega, tehnološkega, psihološkega do filozofskega) in njihova pričakovanja, predstavljajo nove izzive sodobne šole. Izobraževalni proces mora biti bolj povezan s potrebami posameznika in njegovim razvojem ter kulturnim okoljem, v katerem le-ta živi. Kompleksnost stvari, ki vplivajo na današnjega mladostnika (okolje, tehnologija, velika količina takoj dostopnih informacij, možnost neposredne komunikacije z vsem svetom, sodobna spoznanja kognitivne in nevroznanosti, UI ipd.) terja od ustvarjalcev šolskih politik premišljen in hitrejši odziv kot v preteklih obdobjih (Flogie in Aberšek 2019a, 2019b, Flogie, Barle Lakota in Aberšek 2018), predvsem zato, ker se socialno okolje, v katerem živimo (družba, tehnologija ipd.) zelo hitro spreminja in ker mora šola pripravljati učence na poklice in socialna okolja, ki v tem hipu sploh še ne obstajajo. Vse te spremembe socialnega okolja, kar se je v obdobju covida še toliko bolj izpostavilo, pa posledično zahtevajo drugačne, inovativne načine učenja in poučevanja, čemur se mora prilagoditi na paradigmatskem nivoju celoten šolski sistem. Govorimo o novi resničnosti, na katero nismo bili pripravljeni, a smo se morali v njej znajti. To novo stanje pred paradigmatski nivo postavlja nove zahteve, daje pa tudi možnosti prepotrebnih analiz. Zavedati se moramo, da z majhnimi koraki ne bomo mogli doseči velikih sprememb v času, ki ga imamo. V tej strategiji učenja in poučevanja je treba še posebno pozornost posvetiti povečevanju interesa in motivacije mladih za učenje, usvajanju novih tehnologij in novih pristopov k posredovanju in usvajanju znanj. Raziskave kažejo, da učenci, ki so deležni inovativnega didaktičnega poučevanja, podprtega s sodobno IKT, izražajo manj odklonilen odnos do šole (Flogie in Aberšek 2019; Aberšek 2018; Flogie, Barle Lakota in Aberšek 2018). Pri uporabi sodobnih učnih okolij je kontekstu socialnih kompetenc treba nameniti posebno pozornost (Mut in Morey 2008), saj le ta igra pomembno vlogo pri procesu socializacije (Li in Kirkup 2007).

3.1 Učna okolja in socialne kompetence

Kvaliteta in dodana vrednost kompetenc, veščin in znanj posameznika predstavljajo temelj za ustvarjanje konkurenčne prednosti v globalnem svetu in posledično blaginjo posamezne družbe. Globalno se kažejo tudi vse večje demografske spremembe v zmanjšanju družbenih in ekonomskih virov ter posledično s slabitvijo družinskih vezi. Pojem *kompetenca* predstavlja veliko več kot pa samo znanje in veščine. Gre za sposobnost spopasti se s kompleksnimi zahtevami, ki temeljijo na mobilizaciji vseh psihosocialnih virov (vključno z veščinami in stališči) v določenem kontekstu. Evropski referenčni okvir v dokumentu »Ključne kompetence za vseživljenjsko učenje« določa osem ključnih kompetenc, med katerimi sta tudi digitalna pismenost ter socialne in državljanske kompetence. Ključne kompetence se štejejo za enako pomembne, saj vsaka od njih prispeva k uspešnemu življenju v družbi znanja (Evropska komisija 2007).

Socialne kompetence predstavljajo različne medčloveške sposobnosti in lastnosti, ki imajo celosten vpliv na posameznika. Pri tem imamo v mislih predvsem občutek lastne vrednosti oziroma zaupanja vase, samodiscipline ter odgovornosti. Adler definira socialne kompetence tudi kot veščine dobrega shajanja - kot veščine za življenje v sozvočju s samim seboj in okolico. V prvi fazi shajanje samega s seboj v nadaljevanju pa še v interakciji z drugimi ljudmi. Socialna kompetenca od posameznika tako zahteva zdravo mero občutka lastne vrednosti in zaupanja vase, lastne odgovornosti in samodiscipline. V odnosu z drugimi se kaže kot pozornost in empatija oziroma sposobnost vživljanja, zmožnost tako kompromisa kot konflikta, poznavanje ljudi, zmožnost kritike, spoštovanja in tolerantnosti ter sposobnost vse skupaj verbalno izraziti, torej sposobnost jezikovne kompetence (Adler 2014). Socialne kompetence tako razumemo kot interakcije med posameznikom in drugimi ljudmi. Tako socialne kompetence neposredno povezujemo s čustveno inteligenco (ČI), shematsko prikazano na sliki 1.



Slika 1: Čustvena inteligenca (ČI) in socialne kompetence
Vir: lasten.

Socialne kompetence in čustvena inteligenca so veščine in metode, s katerimi posameznik uresničuje uspešno zadovoljitev lastnih potreb v socialnem okolju in v odvisnosti od različnih življenjskih situacij z namenom osebne rasti in razvoja ter sposobnosti empatije in čustvovanja z drugimi. So veščine, ki omogočajo človeku kvalitetno živeti v sozvočju s samim seboj in okolico. Predstavljajo torej odzivanje in sodelovanje posameznika v medosebnih odnosih. Odzivanje in sodelovanje posameznika poteka na treh ravneh:

- na osebnem nivoju (graditev samopodobe, reševanje lastnih težav, izražanje zamisli),
- na nivoju odnosov (pogajanje, sodelovanje, sklepanje kompromisov, mreženje),
- na nivoju širše družbe ali makrosistema (občutljivost za druge, prispevanje k dobrobitu vseh).

3.2 Socialne kompetence in šolski prostor

Če na socialne kompetence pogledamo z vidika izobraževanja in šole, lahko rečemo, da tako kot učenje vsakega področja tudi sposobnost za uspešno obvladovanje socialnih veščin zahteva, da so izpolnjene osnovne telesne in čustvene potrebe mladostnikov. Socialne interakcije med mladostniki so sestavni del njihovega zdravega psihosocialnega razvoja. Ker pa teh potreb domače okolje oziroma družina ne zadovoljuje mladostniku v popolnosti, prehaja ta odgovornost vse bolj v šolski prostor. Torej je naloga šole med drugim zapolniti to socialno vrzel, sicer mladostnik ni možen biti uspešen, kot bi lahko bil sicer (Koplow 2002). V času osnovne in srednje šole se mladostnik vse bolj osredotoča na šolo in vse manj na družino, prijatelji postajajo vse pomembnejši za njegovo socializacijo. Čustvena regulacija ter družabnost sta še dva ključna elementa, pomembna v času mladostnikovega odraščanja. Lahko rečemo, da sta to dve pomembni dinamični spremenljivki vsakega posameznika, ki pomembno vplivata na zmožnost ohranjanja prijateljstva, medtem ko so socialne veščine in zmožnost udejstvovanja in iskanja skupnih aktivnosti pomembne za ohranitev prijateljstva (Semrud-Clikeman 2007). Izziv v procesu takšne socializacije predstavlja vloga sodobnih učnih okolij, e-storitev in e-vsebin v procesu sklepanja prijateljstev (predvsem z vidika sodobnih socialnih omrežij in prijateljskih mrež znotraj teh omrežij). Vendar pa, ali je e-socialna mreža prav tako pomembna, kot je realna? Zanimivo je spoznanje, da že v predšolskem obdobju uporaba sodobne tehnologije nima tako negativnega vpliva, kot je bilo sprva pričakovano. Je pa pri otrocih vseeno moč zaznati povezavo med njihovimi socialnimi kompetencami in njihovim razumevanjem negativnih socialnih situacij (analizirano s strani staršev in vzgojiteljev) (Proekt, Kosheleva, Lugovaya in Khoroshikh 2017).

3.3 UI, etika in izobraževanje

Preden damo strojem (inteligentnim učnim okoljem) smisel etike in morale, morajo najprej ljudje definirati, kaj sta morala in etika. In to na način, ki ga bodo stroji lahko procesirali, oz. na način, ki ga bodo stroji 'razumeli'. Ko govorimo o razumevanju, to pomeni, da morajo biti algoritmi morale in etike definirani tako, da se jih da formalizirati, da se jih da prevesti v jezik znanosti in ga kodirati v enem od strojem razumljivih jezikov, najbolje v strojnem jeziku.

Če so težave uvajanja UI v proizvodne in servisne dejavnosti, torej pri uporabi *pametnih strojev,* kjer lahko ugotavljamo 'napake', relativno hitro zaznane in nimajo drastičnega vpliva na kognitivni del družbe, pa je uvajanje UI v izobraževalne procese, ki so zagotovo temeljni procesi človeške civilizacije, izjemno riskantni in potrebni temeljnega premisleka – *kaj in koliko?* Posledice napak so lahko katastrofalne in predvsem dolgoročne, saj bodo rezultati uvajanja takšnih učnih okolij vidni šele čez vrsto let. Le nekaj izhodiščnih opozoril. Že dalj časa nekateri 'alarmistični' strokovnjaki opozarjajo na nepredvidljive posledice splošne prisotnosti UI v družbi. Ray Kurzweil (2005) je napovedal, da bodo do leta 2029 stroji bolj inteligentni kot ljudje. Stephen Hawking ugotavlja, da »ko bo človek razvil celovito UI in bo ta stopila na svojo pot razvoja, se bo lahko ta samostojno preoblikovala vedno hitreje«, kar bo temeljni riziko in grožnja za obstoj človeštva. Prav tako Elon Musk svari, da lahko UI ustvari »temeljni riziko za obstoj človeške civilizacije«.

3.4 Inteligentna učna okolja in izobraževanje

Osredotočimo se sedaj le na medčloveške odnose – izobraževanje in posameznike, izhajajoče iz tega procesa, to je učitelje in učence in njihovo obnašanje v procesu izobraževanja. Bitja so živčni sistem, ki jih obvešča, razvila zato, da bi uravnavala obnašanje. Obvešča jih:

- o potrebah njihovega *notranjega okolja* in
- o tem, kaj se dogaja v *okolju izven njih*.

Nekatera od naših obnašanj so zelo elementarna in ne potrebujejo nikakršne adaptacije. Na notranje ali zunanje dražljaje reagiramo avtomatsko. Načelno je večina teh obnašanj povezana s *kolektivnim spominom*. Druga, bolj sofisticirana obnašanja, zahtevajo pomnjenje prijetnih ali neprijetnih preteklih izkušenj in ustrezno reakcijo na njihovi podlagi. Ta obnašanja predstavljajo večino socialnega, moralnega in kulturnega znanja, ki smo si ga pridobili. Nadaljnja obnašanja pa zahtevajo bolj dovršeno načrtovanje. Zahtevajo domišljijo in zato abstrakten način razmišljanja, tako da lahko razvijemo strategijo, ki bo zagotavljala čim manj neprijetno ali boleče ukrepanje. To pa predstavlja kreativne, inovativne, torej zavestne sposobnosti človeškega duha.

Zunanje okolje znamo dokaj dobro simulirati, opisovati s takšnimi ali drugačnimi simbolnimi ali mrežnimi sistemi. Vse skupaj pa postane neobvladljivo, ko moramo podobno narediti za *notranje okolje,* ki je povezano z našo (posameznikovo) *zavestjo*. Po Chalmersu bi ta problem lahko razdelili na:

- *lahek problem*, to je, kako so naši zavestni vzgibi vzrok za aktivacijo nevronov, ki naredijo, kar smo načrtovali, in
- *težek problem (*poimenovan tudi *explanatory gap)*, to je, kadar so vzgibi, ki so vir naših obnašanj pravzaprav zavestni, kar pomeni vprašanje svobodne volje.

Ne bomo se poglabljali v podrobnosti. Oglejmo si le nekaj težav pri obravnavi 'težjega UI problema', to je uvajanje UI v izobraževanje (splošno družbo):

Preden damo strojem, v našem primeru kognitivnemu delu stroja (to je inteligentni programski opremi oz. inteligentnemu učnemu okolju, oz. i-učbenikom, inteligentnim tutorskim sistemom ali podobnemu učnemu gradivu), komponente morale in etike, morajo biti algoritmi morale in etike definirani tako, da se jih da formalizirati. Prav tako pa je treba definirati metode ugotavljanja, ali sistemi delujejo dolgoročno pravilno, saj posledic nedelovanja ali nepravilnega delovanja (predvsem s stališča etike in morale, spomnimo se Asimovih zakonov robotike) ne moremo sprotno spremljati in izvajati ustrezne korekcije. Skratka, z današnje perspektive se srečujemo z dvema nerešljivima problemoma, saj ne znamo definirati splošnih etičnih norm niti v splošnem jeziku, kaj šele, da bi jih lahko definirali v jeziku znanosti in tem etičnim normam dali generalizirano in normirano vsesplošno veljavo. Podobno pa tudi velja za možnost preverjanja dejanskih odstopanj od normativnih zahtev.

4 Zaključek ali ne/zmožnosti UI

Praktično vse raziskave s področja umetne inteligence skušajo razviti sistem, ki bi se obnašal inteligentno in bil sposoben reševati relativno težke probleme. Razvojne metode imajo mnogokrat osnovo v človeškem načinu reševanja problemov. *Dolgoročni cilj, ki si ga je tehnološki svet zastavil, je, da bi računalniška inteligenca (njene sposobnosti) dosegla ali celo presegla človeško inteligenco*. Pomemben vidik razumevanja sposobnosti umetne inteligence je vpliv učenja na inteligenco, hitrost reševanja problemov, osnovne omejitve algoritmov in posnemanje inteligentnega obnašanja:

- *Vpliv učenja na inteligenco*: Z učenjem se sposobnosti sistema povečujejo, zato tudi inteligenca narašča. Človeška inteligenca je dinamična in se v življenju neprestano spreminja, po navadi se veča. Seveda pa moramo pri tem upoštevati tudi različnost inteligence.
- *Hitreje je bolj inteligentno*: Prilagajanje okolju in reševanje problemov sta boljša (bolj učinkovita), če sta hitrejša. Zato je inteligenca znatno povezana s hitrostjo in časom. Vsi testi inteligence so časovno omejeni, tako kot tudi vsi izpiti. Tako lahko zaključimo, da čim hitrejši je računalnik, bolj je inteligenten, vzporedno (paralelno) procesiranje je bolj inteligentno od zaporednega (serijskega) itd.
- *Omejitve inteligence*: Če hipotetično človeka naredimo (degradiramo) ekvivalentnega računalniškemu algoritmu, potem vse omejitve računalniške teorije veljajo tudi za človeka in sposobnosti njegove inteligence. Če predpostavimo, da je človek sposobnejši 'stroj' kot (digitalni) računalnik (na primer zvezni in ne diskretni stroj), potem so človeške aktivnosti neopisljive, saj so med drugim povezane tudi s človekovo zunanjo okolico. Posledično iz te predpostavke izhaja, da je nemogoče algoritmično izpeljati umetno inteligenten sistem, ki bo v celoti posnemal človeško obnašanje. In vprašanje je tudi, zakaj bi jo morala. Človek je izrazito nepredvidljiv, mnogokrat nelogičen in velikokrat samodestruktiven. In zakaj bi morala tudi UI biti taka?
- *Posnemanje inteligentnega obnašanja*: Danes so sodobne tehnologije, kot so filmi, multimedija, računalniki, roboti in virtualna realnost, izjemno prepričljive in sugerirajo, da je možno posnemati prav vse in s tem doseči senzacijo realnosti. Zato so stroji, če seveda izključimo zavest, v principu dovolj inteligentni, da ustvarijo občutek umetne inteligence. (Pomislimo samo na velike kapacitete spomina, ki vsebujejo rešitve za vse možne situacije). Če dodamo še izjemne sposobnosti procesiranja (super paralelnost s super hitrimi procesorji) algoritmov za učinkovito iskanje ogromnega števila informacij in algoritmov za strojno učenje sposobnih spletnih izboljšav in to povežemo z ustrezno hevristiko, potem lahko takšne stroje v resnici imenujemo 'inteligentne', saj lahko presegajo ljudi v mnogih, če ne v vseh 'praktičnih' nalogah. Vendar moramo ponovno poudariti, da takšni stroji verjetno še vedno nimajo zavesti, torej niso primerljivi z živimi bitji.

V principu smo sposobni določiti (zaznati ali objektivno izmeriti) katerokoli lastnost sistema, ki ima neko sposobnost učenja in določeno stopnjo inteligence. V nasprotju z učljivostjo in inteligenco je zavest nekaj popolnoma drugega. Treba jo je povezovati z osebno izkušnjo in kateri koli objektivni opazovalec je ne more preprosto verificirati.

Možno je objektivno določiti sposobnost učenja, količino usvojenega znanja, sposobnost (inteligenco) prilagajanja okolju in reševanja problemov. Z različnimi testi lahko merimo določene tipe inteligence, dobljeni izidi pa so le bolj ali manj zanesljivi. V nasprotju s tem v principu ni možno preverjati zavesti sistema. Ali je (biološki ali umetni) sistem zavesten ali ne, ve samo sistem sam, seveda če je zavesten. Zunanji opazovalec nima nikakršne možnosti ugotavljati prisotnosti ali odsotnosti zavesti. Posameznik lahko govori o zavesti, če ima sam zavest in če predpostavlja, da ima sistem, ki mu je podoben in o katerem želi govoriti, prav tako zavest. Vsak zavestni sistem lahko posnemamo z nezavestnim sistemom (npr. Turingov test umetne inteligence), da bi iskali (vedno nepopolno) podobnost, zato je vsak objektivni opazovalec lahko zelo ukanjen (Bregant 2010; Abramsen 2008). Vendar pa zavest gor ali dol, UI je realnost in lahko rečemo, da se UI širi viralno, s človeku nepojmljivo hitrostjo in le še vprašanje časa je, kdaj se bo UI razvila na stopnjo inteligence roja, ko bo pričela razvijati sebi lastno zavedanje, ki bo zagotovo začetna stopnja UI zavesti. In zakaj bi ta morala biti definirana s človeškimi normami zavesti, ki jih še sami dobro ne poznamo in predvsem ne razumemo?

Viri in literatura

Aberšek, B., Borstner, B., Bregant, J. (2014). *Virtual teacher: cognitive approach to e-learning material*. Newcastle Upon Tyne: Cambridge Scholars Publishing.

Aberšek, B. (2018). *Problem-based learning and proprioception*. Newcastle Upon Tyne: Cambridge Scholars Publishing.

Adler, E. (2014). *Ključni dejavnik socialna kompetenca: kaj vse nam manjka in česa se lahko naučimo*. Novi Sad: Psihopolis institut.

Asimov, I. (1950). *I robot*. New York: Gnome Press.

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

Bregant, J. (2010). »Ali lahko stroj misli?«. *Analiza*, 4, str. 55–72.

COM 237 COM 237. (2018). »Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions: Coordinated Plan on Artificial Intelligence«. *European Commission*. URL = https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN.

COM 759 COM 759. (2018). »Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the

Committee of the Regions: Coordinated Plan on Artificial Intelligence«. *European Commission*. URL = https://eur-lex.europa.eu/legal-content/EN/TXT/DOC/?uri=CELEX:52018DC0795&qid=1546111312071&from=EN.

Copeland, J. (1993/2007). *Artificial Inteligence: A philosophical Introduction*. New Jersey: Blackwell Publishing.

Dorigo, M. in Birattari, M (2007). »Swarm intelligence«. *Scholarpedia*, 2(9), 1462. URL = http://www.scholarpedia.org/article/Swarm_intelligence.

EK. (2007). »Ključne kompetence za vseživljenjsko učenje, Evropski Referenčni Okvir«. *Evropska komisija*. Luxemburg: Evropska komisija.

Flogie, A., Aberšek, B. (2019a). *The Impact of Innovative ICT Education and AI on the Pedagogical Paradigm*. Newcastle Upon Tyne: Cambridge Scholars Publishing.

Flogie, A. in Aberšek, B. (2019b). *Inovativna učna okolja - vloga IKT*. Maribor: Zavod Antona Martina Slomška.

Flogie, A. in Aberšek, B. (2017). »Transdisciplinary approach of science, technology, engineering and mathematics education«. *Journal of Baltic Science Education*, 14, str. 779–790.

Flogie, A., Barle Lakota, A. in Aberšek, B. (2018). »The Psychosocial and Cognitive Influence of ICT on Competences od STEM Students«. *Journal of Baltic Science Education,* 17, str. 267–276.

Horst, S. (2007). *Beyond Reduction: Philosophy of Mind and Post-Reductionist Philosophy of Science*. Oxford: Oxford University Press.

Koplow, L. (2002). *Creating Schools That Heal*. New York: Teachers College Press, Columbia University.

Kordige, Aberšek, M. in Aberšek, B. (2020). *Society 5.0 and Literacy 4.0 for 21st Century*. Hauppauge: Nova Science Publishers, Inc.

Kurzweil, R. (2006). *The Singularity is Near*. London: Penguin.

Li, N. in Kirkup, G. (2007). »Gender and cultural differences in internet use: a study of China and the UK«. *Computers&Education*, 48, str. 301–317.

Markič, O. (2010). Kognitivna *znanost: Filozofska vprašanja*. Maribor: Aristej.

McGinn, C. (1999). *The Mysterious Flame: Conscious Minds in a Material World*. New York: Basic Books, Perseus Books Group.

Minsky, M. L., Papert, S. A. (1969). *Perceptrons*. Cambridge, MIT Press.

Proekt, Y., Kosheleva, A., Lugovaya, V. in Khoroshikh, V. (2017). »Developing Social Competence of Preschoolers in Digital Era: Gender Dimension«. V D. Alexandrov, A. Boukhanovsky, A. Chugunov, Y. Kabanov in O. Koltsova (urd.), *Communications in Computer and Information Science*, 745. Springer International Publishing, str. 87–101.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.F., Breazeal, C., Crandall, J.W., Christakis, N.A., Couzin, I.D., Jackson, M.O., Jennings, N.R., Kamar, E., Kloumann, I.M., Larochelle, H., Lazer, D., Mcelreath, R., Mislove, A., Parkes, D.C., Pentland, A., Roberts, M.E., Shariff, A., Joshua B. Tenenbaum, J.B. in Wellman, M. (2019). »Machine behaviour«. *Nature*, 568, str. 477–286.

Semrud-Clikeman, M. (2007). *Social Competence in Children*. Boston: Springer.

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G. et al. (2016). »Artificial Intelligence and Life in 2030«. *One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel*. Palo Alto: Stanford University.

Tinbergen, N. (1963). »On aims and methods of ethology«. *Ethology*, 20, str. 410–433.

Turing, A. (1950). »Computing Machinery and Intelligence«. *Mind*, 59, str. 434–460.

SODOBNE PERSPEKTIVE DRUŽBE: UMETNA INTELIGENCA NA STIČIŠČU ZNANOSTI

JANEZ BREGANT,[1] BORIS ABERŠEK,[2] BOJAN BORSTNER[1]
(UR.)

[1] Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
janez.bregant@um.si, bojan.borstner@um.si
[2] Univerza v Mariboru, Fakulteta za naravoslovje in matematiko, Maribor, Slovenija
boris.abersek@um.si,

Sinopsis Vprašanja, ki se dotikajo prihodnosti družbe, so v veliki meri povezana z razvojem umetne inteligence (UI). V publikaciji, ki je pred vami, je UI odprto področje znanstvenega raziskovanja, kjer se srečujejo družboslovne, naravoslovne in tehniške znanosti. Nagovarja trenutno najbolj aktualno temo, ki jo lahko prosto povzamemo kot prednosti in slabosti hitrega razvoja UI z vidika njenega vpliva na moralo, psihologijo in izobraževanje. Vsebuje štirinajst člankov, razdeljenih v dva vsebinska dela, pri čemer se prvi osredotoča na presek med UI, filozofijo in etiko, drugi pa na presek med UI, psihologijo in izobraževanjem. Posebej so poudarjene tiste dimenzije vsebin, ki se nahajajo v preseku omenjenih področij. Takšna interdisciplinarnost predstavlja dodano vrednost te publikacije, saj uporaba vsaki vedi lastne raziskovalne metodologije v člankih pripomore k njihovi svojstveni obogatitvi v smislu vpliva UI na moralo, transparentnost in uporabnost. Ukvarjanje z raziskovalnimi problemi na opisan način pa poskrbi za raznolike, pestre in domiselne prispevke, ki pri ponujanju rešitev pogosto prestopijo meje ustaljenega.

Ključne besede:
umetna inteligenca, moralna odgovornost, transparentnost, izobraževanje, singularnost

Univerzitetna založba Univerze v Mariboru

DOI https://doi.org/10.18690/um.ff.11.2022
ISBN 978-961-286-675-4

CONTEMPORARY PERSPECTIVES OF SOCIETY: ARTIFICIAL INTELLIGENCE AT THE INTERSECTION OF SCIENCES

JANEZ BREGANT,[1] BORIS ABERŠEK,[2] BOJAN BORSTNER[1]
(EDS.)

[1] University of Maribor, Faculty of Arts, Maribor, Slovenia
janez.bregant@um.si, bojan.borstner@um.si
[2] University of Maribor, Faculty of Natural Science and Mathematics, Maribor, Slovenia
boris.abersek@um.si

Abstract Questions concerning the future of our society are largely related to the development of artificial intelligence (AI). In this publication, AI represents the open field of research where social, natural, and technical sciences meet. It addresses maybe the most topical subject: pros and cons of the AI's rapid evolution in terms of its effects on morality, psychology, and education. It contains fourteen articles and is divided into two parts, the first dealing with the intersection of AI, philosophy, and ethics, and the second dealing with the intersection of AI, psychology, and education. The interdisciplinary nature of the publication gives it additional value. Namely, the use of each disciplines' own research methodology contributes to the enrichment of the articles in the sense of AI's effect on morality, transparency, and applicability. Moreover, dealing with the research problems in the described way gives way to diverse, rich, and inventive contributions that often step out of the proverbial box in offering their respective solutions.

Keywords: artificial intelligence, moral responsibility, transparency, education, singularity

https://doi.org/10.18690/um.ff.11.2022 DOI
978-961-286-675-4 ISBN

University of Maribor Press

Umetna inteligenca je danes tako močno prisotna v našem življenju, da imamo občutek, kot da je z nami že od nekdaj. Njena vpetost v naše vsakodnevno početje ta vtis samo še utrjuje. Kljub temu pa se zdi, da smo (spet) pred vrati tehnološke revolucije: stroji, ki se učijo, ne bodo prevzeli zgolj umazanih in zdravju škodljivih služb, ampak bodo nadomestili ljudi tudi tam, kjer se zahtevajo izvirnost, ustvarjalnost in iznajdljivost. Tako se lahko (v kolikor se že ni) kmalu zgodi, da bodo o odobritvi kredita za nakup stanovanja, medicinskih diagnozah ali dodelitvi socialne pomoči odločali samo še stroji. Zato ni presenetljivo, da v nas to, kar nas čaka v prihodnosti, vzbuja tako navdušenje kot strah.

Janez Bregant, iz *Uvod: za kaj sploh gre?*

Univerza v Mariboru
Filozofska fakulteta