Pregledni znanstveni članek/Article (1.02)
Bogoslovni vestnik/Theological Quarterly 84 (2024) 4, 757—781
Besedilo prejeto/Received:09/2024; sprejeto/Accepted:12/2024
UDK/UDC: 004.8:14
DOI: 10.34291/BV2024/04/Zalec

Bojan Žalec

Človeški podobna umetna inteligenca in superinteligenca: verjetnost in glavne težave njunega oblikovanja

Human-like Artificial Intelligence and Superintelligence: The Probability and Main Challenges of Their Design

Povzetek: Avtor se ukvarja z vprašanjem verjetnosti nastanka umetne inteligence, podobne človeški, in nastanka superinteligence. Trdi, da ni verjeten nastop nobene. Verjetnost prve zavrne na podlagi argumentacije Erika J. Larsona, ki temelji na bistvenih značilnostih človeške inteligence, kot so splošnost, intuicija, zdravi razum in abdukcija. Verjetnost nastanka superinteligence zavrne na podlagi argumentov Françoisa Cholleta, ki poudarja ne-splošnost, situacijskost in kontekstualnost ter zunanjostnost inteligence. Avtor poudarja pomembnost pravilnega razumevanja pojma inteligence ter njene ustrezne opredelitve. Napačno razumevanje ima lahko negativne učinke, ki presegajo akademsko področje – vključno z neupravičenimi finančnimi dobički ter škodo pri organiziranju znanstvenega dela (velika znanost, znanost panja) in spodbujanju pogojev za ustvarjalnost.

Ključne besede: verjetnost človeški podobne umetne inteligence, verjetnost superinteligence, splošnost, intuicija, abdukcija, zdravi razum, partikularnost inteligence, zunanjostnost inteligence, velika znanost

Abstract: The author deals with the question of the probability of the emergence of humanlike artificial intelligence and the emergence of superintelligence. He argues that no emergence is probable. The probability of the first is rejected, following Erik J. Larson's argumentation based on the essential characteristics of human intelligence. These include generality, intuition, common sense, and abduction. The probability of the onset of superintelligence is rejected based on arguments provided by François Chollet, emphasizing the non-generality, situationality and contextuality, and externalism of intelligence. The author demonstrates the importance of a correct understanding of the concept of intelligence and its definition. Misunderstanding can have negative effects far beyond the academic realm, including unjustified financial gains and damage

in terms of organizing scientific work (big science, hive science), and fostering conditions for creativity.

Keywords: probability of humanlike artificial intelligence, probability of superintelligence, generality, intuition, abduction, common sense, particularity of intelligence, externalism of intelligence, big science

1. Uvod

Glavno vprašanje prispevka je: Ali je oblikovanje umetne inteligence (v nadaljevanju UI), podobne človeški, verjetno?[1] V nadaljevanju bom za trditev, da bo ustvarjena umetna inteligenca, podobna človeški, uporabljal kratico HIT. Ko govorim o umetni inteligenci, podobni človeški, mislim na UI, ki bi bila sposobna vsega, kar zmore človeška inteligenca. Vendar pa je vprašanje resničnosti HIT tesno povezano z verjetnostjo pojava superinteligence (v nadaljevanju SI) – inteligence, ki bi človeško inteligenco močno presegala. Zato bom razpravljal tudi o verjetnosti nastanka SI, ki ga napovedujejo nekateri transhumanisti – na primer Ray Kurzweil, ki napoveduje nastop singularnosti. (Kurzweil 2005) V nadaljevanju bom za trditev, da bo nastopila SI, uporabljal kratico SIT. HIT in SIT sta prepleteni. Po eni strani nastanek SI ni verjeten, če človeški podobna UI ni mogoča; po drugi strani pa bi pojav človeški podobne UI lahko razumno interpretirali kot pomemben dokaz v prid SIT. Tako sta tezi povezani – in argumenti za in proti HIT so posredno tudi argumenti za in proti SIT. Zato je smiselno obe tezi obravnavati skupaj.

V nadaljevanju bom zagovarjal stališče, da sklicevanje na dosedanje dosežke UI, ki so v mnogih pogledih res impresivni, glede na glavno vprašanje prispevka nima velike teže. Navajanje teh dosežkov kot prepričljivih dokazov za HIT in SIT temelji na nerazumevanju človeške inteligence. Zato bom v prispevku kritično razmišljal o razumevanjih (človeške) inteligence, na katerih temeljijo napovedi o verjetnosti obeh tez.

V prvem delu prispevka bom predstavil argumente proti HIT, kot jih je razvil Eric J. Larson v svoji knjigi *Mit o umetni inteligenci* (Larson 2021), v drugem delu pa argumente proti verjetnosti SIT, kot jih je predstavil François Chollet v svojem vplivnem prispevku „Neverjetnost eksplozije inteligence“ (Chollet 2017).

Larson se osredotoča na štiri temeljne in bistvene značilnosti človeške inteligence: splošnost (angl. *generality*), intuicijo, zdravi razum in zmožnost abdukcije. Trdi, da obstoječa UI teh zmožnosti nima in da v okviru trenutne (Turingove) paradigme razvoja UI take UI, ki bi jih imela, ne bomo mogli ustvariti. Prav tako je malo verjetno, da bi te sposobnosti UI razvila sama. Chollet trdi, da je nastanek SI malo verjeten. V svoji argumentaciji poudarja tri značilnosti inteligence: 1. ne-

[1] Prispevek je nastal v okviru raziskovalnega programa „P6-0269: Religija, etika, edukacija in izzivi sodobne družbe“ ter raziskovalnega projekta „J6-4626 (B): Teologija, digitalna kultura in izzivi na človeka osrediščene umetne inteligence“, ki ju financira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS).

-splošnost; 2. situacijskost oziroma kontekstualnost; 3. zunanjostnost (angl. *externalism*). Pomen teh izrazov bom pojasnil v nadaljevanju. Prvi dve značilnosti v nadaljevanju zajemam z izrazom partikularnost inteligence. Tretja značilnost partikularnost implicira.

Splošno gledano lahko to razpravo razumemo kot prispevek k obrambi antikartezijanske paradigme razumevanja (človeške) inteligence in uma (Žalec 2023b, 105). Za razumevanja in pristope znotraj te paradigme se uporabljajo različni izrazi: aristotelijansko in interaktivno stališče (Miščević 1988), enaktivizem (Gallagher 2017), ekologizem (Gibson 1979; Potrč 1993; 1996, 194–197; 2004, 55–61), eksternalizem (McCulloch 1995, 184ff; Rowlands, Lau in Deutsch 2020; Potrč 1993), teorija stika (Dreyfus in Taylor 2015), utelešeni humanizem in antropologija utelešenja (Fuchs 2021) in še drugi. Za različico antikartezijanstva, ki jo sam zagovarjam, je središčnega pomena Heideggerjev pojem ‚bit-v-svetu'. Zato ni presenetljivo, da omenjam tudi razumevanje Huberta Dreyfusa (Dreyfus 1991; Žalec 2023b, 19–29; 31–39; 107–108), vendar ga tu ne obravnavam podrobno.

Osredotočam se na inteligenco, podobno človeški. Trdim, da človeški podobna UI ne obstaja in da njen nastanek ni verjeten. Vendar to ne pomeni, da menim, da ne obstaja UI, ki ne bi bila inteligentna v pravem pomenu besede. Nasprotno – prepričan sem, da so obstoječi sistemi globokega učenja inteligentni. Vendar pa moramo razlikovati med človeško – ali človeški podobno – inteligenco in inteligenco kot tako. Človeška inteligenca ni edina vrsta inteligence. Sistemi globokega učenja so inteligentni, vendar njihova inteligenca ni podobna človeški (Žalec 2023a).

Dotikam se tudi družbene razsežnosti nerazumevanja potencialov UI. Ena od posledic tega je ideologija velike znanosti (angl. *big science*), ki je nevarna in škodljiva – med drugim kot negativen dejavnik pri negovanju (človeške) ustvarjalnosti v znanosti, pri razvoju tehnologije in tudi na drugih področjih. Škodljivost te ideologije kaže, da razprava o naravi, možnostih in potencialih razvoja UI ni le ozkega akademskega pomena, temveč je zelo pomembna s širšega – družbenega in moralnega – vidika.

V literaturi lahko najdemo različne druge argumente za verjetnost človeški podobne UI in SI in proti njej, s katerimi se ne ukvarjam – na primer obsežna in temeljita argumentacija Landgrebeja in Smitha (2023), da ustvarjanje ustreznega matematičnega modela človeškega uma presega človeške sposobnosti. To je pri mojem prispevku zagotovo omejitev. Po drugi strani pa verjamem, da že argumenti, ki jih bom predstavil, sami po sebi tvorijo dobro podlago za razumno zavrnitev verjetnosti HIT in SIT.

2. Bistvene značilnosti človeške inteligence, ki predstavljajo temeljni problem za razvoj človeški podobne UI

Naj razpravo začnem z nekaterimi splošnimi razlogi, zakaj – vsaj v bližnji prihodnosti in glede na trenutno znanstveno in tehnično znanje ter prevladujoči pristop

– ne moremo pričakovati, da bi kdo ustvaril UI, ki bi imela bistvene značilnosti človeške inteligence. Kot take značilnosti lahko omenimo zmožnosti, ki so tesno prepletene (Larson 2021): splošnost (kot nasprotje ozke specializacije za določene naloge in področje), sposobnost razumevanja, intuicijo, sposobnost učenja, izbiro problemov za reševanje, zdravi razum in sposobnost posebne oblike sklepanja, imenovane abdukcija.

Larson trdi, da trenutno nihče zares ne ve, kako ustvariti človeški podobno UI – nihče za to nima ustrezne znanstvene in tehnične zamisli. Namesto znanstvenih argumentov v prid možnosti ustvarjanja človeški podobne UI se mnogi zatekajo k različnim (znanstveno) slabo utemeljenim trditvam, ki širijo prepričanje, da človeški podobno UI bomo ustvarili. Vedno znova se najdejo (ugledne) osebe, ki trdijo, da smo ta cilj (načelno) že dosegli, da ga bomo kmalu dosegli, da bomo (kmalu) priča nastanku UI, ki bo človeško inteligenco močno presegla in tako naprej. Takšni ‚lažni preroki' ali promotorji UI ustvarjajo mit o UI (Larson 2023). Tukaj je ‚mit' mišljen v negativnem smislu: kot lažno prepričanje, da nekaj obstaja ali je verjetno, čeprav v resnici ne obstaja in ni verjetno. Kar ne obstaja in ni verjetno, je človeški podobna UI in SI. Podobno lahko v omenjenem smislu govorimo o ideologiji UI – to je o lažnem prepričanju, ki ga nekateri (zavestno ali nezavedno, namerno in nenamerno) ustvarjajo zaradi določenih interesov, med katerimi finančni niso zanemarljivi.

2.1 Problem splošnosti in past ozkosti

Nekateri napovedujejo, da ko bomo ustvarili človeški podobno UI, bo ta kmalu ustvarila inteligenco, ki bo človeške sposobnosti močno presegala – to pa bo privedlo do razvoja SI, katere zmožnosti si ne moremo predstavljati. Vendar pa o pravilnosti te domneve obstaja več utemeljenih pomislekov. Prvič: človeška raven inteligence obstaja že zelo dolgo, vendar se ni razvila v SI, ki bi človeško inteligenco presegala. Zakaj bi bilo razumno pričakovati, da bi človeški podobni UI to uspelo? Drugič: nihče ne ve, kako bi človeški podobna UI to dosegla. Nihče tega procesa ni znanstveno opisal. Toda kljub temu nekateri napovedujejo, da se to bo zgodilo.

Človeška inteligenca je zmožna intuicije in neodvisne oziroma samostojne izbire problemov. Zato človeška inteligenca ni ozka – ni omejena na reševanje zgolj določenih problemov ali nalog. Larson poudarja, da nihče nima znanstvene predstave, kako ustvariti UI, ki bi bila sposobna intuicije in neodvisne izbire problemov, potrebne za preseganje te ozkosti. Brez tega pa o človeški podobni UI ne moremo govoriti.

Eden od načinov za zagovarjanje teze o SI je sklicevanje na evolucijo: SI se bo v evoluciji, ki proizvaja vedno višje in bolj kompleksne sisteme, zagotovo pojavila. Tudi če ne vemo, kako, lahko na podlagi evolucijskih razlogov trdimo, da se SI sčasoma bo pojavila. Vendar takšna napoved ne more veljati za znanstveno utemeljeno. Eden od znakov njene neznanstvenosti je, da je načelno neovrgljiva. Poleg tega ta napoved ni odvisna od napak pri napovedih. Kurzweil je pojav človeški podobne UI napovedal že za leto 2029, singularnost pa do leta 2045. Toda tudi če se to ne zgodi, lahko zagovorniki SI pri svoji napovedi še vedno vztrajajo – zgodilo

se bo pač nekaj desetletij kasneje. In če se ne bo zgodilo, se bo zgodilo še malo kasneje.[2]

Nekateri zagovarjajo tezo o SI na podlagi načela (nujnega) preskoka iz količine v kakovost. Neprestano izboljševanje ‚kvantitativnih' sposobnosti UI mora sčasoma pripeljati do preskoka v kakovost, do pojava kvalitativno višje ravni UI in tako naprej do SI. Načelo nujnega preskoka iz količine v kakovost so v preteklosti že zagovarjali,[3] vendar ga izkustvo ne potrjuje.

Trditve o pojavu človeški podobne UI in SI so bile v preteklosti izrečene že večkrat. Doslej se nobena od teh napovedi ni uresničila. Poleg tega, kot že omenjeno, trenutno nihče nima znanstvene ideje, kako ustvariti sistem, ki bi izkazoval značilnosti človeški podobne inteligence, kot so navedene zgoraj. Brez kakšnega novega temeljnega teoretičnega vpogleda, preboja ali revolucije pa ni možnosti, da bi to dosegli – zagovorniki možnosti ustvarjanja UI v okviru trenutno razpoložljivega teoretičnega okvira znanstvenih argumentov za svojo trditev dejansko ne ponujajo, kar kaže, da gre za ideologijo.

Da bi razumeli izzive in ovire za razvoj človeški podobne UI, se moramo osvoboditi ozkega pojma inteligence, ki vso inteligenco zvaja na (neodvisno) reševanje problemov. Neodvisno reševanje problemov je za pripisovanje inteligence zadosten pogoj, vendar to še ne pomeni, da je zadosten za vsako vrsto inteligence. Je le minimalen – nujen in zadosten – pogoj za inteligenco (Žalec 2023a), kar pa za človeško inteligenco ni dovolj. Človeška inteligenca je več kot minimalna inteligenca. Človeška inteligenca vključuje ne le neodvisno reševanje problemov, ampak tudi neodvisno izbiro problemov. »Gospodar mora samo znati ukazati tisto, kar mora suženj znati izvesti,« pravi Aristotel (1995, 1992). V tem smislu je človeška inteligenca inteligenca gospodarja, medtem ko je UI inteligenca sužnja. Ne more si ukazovati sama, ne more si sama postaviti svojih problemov (Peirce 1887, 165; Larson 2021, 233) – potrebuje gospodarja.[4] Zato ne more biti ustvarjalna v smislu postavljanja novega problema. Definiranje inteligence kot neodvisnega reševanja problemov – recimo temu minimalna definicija – je zelo jasno in zato inženirsko uporabno. Po tej definiciji lahko za UI že rečemo, da je inteligentna. Toda če ne upoštevamo, da ta definicija za opredelitev človeške inteligence ni dovolj, saj ne upošteva njenih razlikovalnih značilnosti, kot je splošnost, lahko ta definicija posebnost človeške inteligence zamegli – in posledično vodi k napačnemu razumevanju potencialov UI. Vendar je opustitev minimalne definicije kot zadostne za definicijo katerekoli inteligence (ne samo minimalne inteligence) med programerji in inženirji bržkone precej težavna naloga, saj je takšno razumevanje del Turingove zapuščine – Turing pa je utemeljitelj paradigme, ki pri raziskavah in razvoju UI še danes močno prevladuje.

---

[2] Na podobno neovrgljiv način so nekateri predstavniki *new agea* napovedovali začetek nove, ‚višje' zavesti (Sire 2002, 164ff).

[3] Friedrich Engels (1975, 118–119; 351–352; 482; 510–511; 516–517; 552) ga je zagovarjal kot eno glavnih načel svoje ‚filozofije' narave ali dialektike narave.

[4] UI svojemu gospodarju ne more reči »Ne!« v smislu zavrnitve ukaza. Pripisovanje takšne zavrnitve UI nima nobenega smisla. Pri človeškem sužnju pa takšna možnost obstaja.

Če računalnik uspešno ali celo bolje opravlja naloge, ki so prej zahtevale človeško inteligenco, je to v razvoju UI zagotovo napredek – in bolj zapletene kot so te naloge, večji je napredek. S tem se lahko strinjamo. Vendar pa je vprašanje, kakšno težo ima ta napredek kot evidenca za HIT ali SIT. Igranje iger, kot so šah in go, pravilno odgovarjanje na vprašanja kviza, medicinske diagnoze itd., ki jih izvaja UI, nekateri razglašajo za dokaze v prid HIT in SIT. S tem se ne strinjam – in namen tega prispevka je utemeljitev tega nestrinjanja.

Obravnavanje inteligence kot reševanja problemov vodi k ozkim aplikacijam. Če bi se stroj svojo omejitev na ozke naloge naučil preseči in bi bil sposoben reševati probleme na splošno, bi to pomenilo prehod v višjo ali (bolj) človeški podobno inteligenco. Toda zaenkrat smo od oblikovanja splošne UI še daleč. Da bi dosegel svoje cilje, se mora vsak sistem strojnega učenja naučiti nekaj specifičnega. Raziskovalci temu pravijo pristranskost sistema. Pristranskost v strojnem učenju pomeni, da je sistem zasnovan in prilagojen za učenje nečesa posebnega – to pa je ravno proizvodnja aplikacij za reševanje ozkih problemov. Zato se sistemi globokega učenja, ki jih je Facebook uporabljal za prepoznavanje človeških obrazov, niso hkrati naučili, kako izračunati naše davke. Situacija je v resnici še slabša: raziskovalci so ugotovili, da če je sistem strojnega učenja pristranski in specializiran za priučitev določeni aplikaciji ali nalogi, potem slabše opravlja druge naloge. Obstaja obratna korelacija med uspehom sistema pri učenju določene stvari in uspehom njegove priučitve drugi nalogi – in to velja celo za zelo podobne naloge. Sistem strojnega učenja, ki se uči igrati šah na visoki ravni, se goja na visoki ravni ne bo naučil igrati – in obratno. Sistem za go je bil namreč zasnovan s posebno pristranskostjo za učenje pravil goja. Težava je, da se pristranskosti ne moremo znebiti, ker je sestavni del strojnega učenja. Pregovor »nič ni zastonj« velja tudi za strojno učenje. V tem primeru to pomeni, da ko bo uporabljen za poljubne probleme oziroma naloge, noben sistem strojnega učenja, ki ni pristranski, ne bo deloval nič bolje od naključnega ugibanja. Resnično nepristranski sistem – sistem, ki ga programerji ne oblikujejo kot pristranskega – je neuporaben. Po drugi strani se pristranski sistem nauči samo tega, česar si želijo njegovi oblikovalci. Tako s pristranskostjo sistem naredimo ozek v smislu, da potem ne posplošuje na druga področja. Ozkost je torej od uspeha sistema neločljiva: ozkost in uspeh sta v sistemih strojnega učenja dve plati istega kovanca (Larson 2021, 28–30).

Sistemi UI, ki smojih razvili doslej, lahko probleme uspešno rešujejo samo tako, da jih oblikujemo kot pristranske. Tudi učeči se sistemi UI so zgolj taki ozki sistemi. Doslej ni znan noben znanstven ali tehnični preboj iz ozkih sistemov do splošne inteligence, kot jo izkazujejo ljudje. S tega vidika se je razvoj splošne UI ustavil: razumevanje inteligence le kot (neodvisnega) reševanja problemov postopoma – vendar neizogibno – vodi v teoretično slepo ulico v samem osrčju raziskovanja UI, v ‚past ozkosti',[5] kot jo imenuje Larson. To kaže, da je razumevanje inteligence,

[5] Eden prvih, ki je na problem izvorne pobude in ozkosti jasno opozoril, je bil Charles Sanders Peirce: »Vsak stroj za sklepanje, to se pravi vsak stroj, ima dve inherentni nezmožnosti. Prvič, je brez vsake izvirnosti, brez vsake pobude. Ne more si najti svojih lastnih problemov; ne more se sam napajati. Ne more se sam usmerjati med različnimi mogočimi postopki. /…/ [S]troj bi bil popolnoma brez izvirne

ki se jo zvede na (neodvisno) reševanje problemov, preozko in za doseganje širših ciljev od proizvodnje ozkih aplikacij neustrezno (30–32).

Središčni problem za razvoj človeški podobne UI lahko opišemo s pojmom intuicije. Če bi želeli, da bi UI bila podobna človeški, bi morala biti sposobna intuicije. Če bi želeli, da bi UI imela intuicijo, kot jo imajo ljudje in raziskovalci, bi morali to intuicijo znanstveno opisati tako, da bi bil ta opis za programerje in inženirje uporaben. Vendar pa si nihče ne predstavlja, kako to narediti. Brez intuicije UI ne more preseči ‚prekletstva' ozkosti – kar pa bi bilo potrebno, če bi se želela učiti tako, da bi postala podobna človeški inteligenci. Takšno učenje zahteva, da sistem probleme izbere sam, za to pa je potrebna intuicija. Vemo, da oblikovalci sistemov UI za usmerjanje teh sistemov, katere specifične probleme naj rešujejo ali se naj se naučijo reševati, uporabljajo lastno intuicijo. Toda da bi bila UI resnično inteligentna, bi morala imeti svojo intuicijo. Tega nima in kot že omenjeno – nihče nima znanstvene predstave, kako ustvariti UI, ki bi jo imela (31–32).

Povzemimo glavne doslej navedene razloge za zavrnitev napovedi o verjetnem nastanku človeški podobne UI ali SI, še zlasti v bližnji prihodnosti. SI bi se iz človeške ali človeški podobne UI lahko razvila ‚sama od sebe' v teku evolucije, ne da bi jo tehnično razumeli. A to se ni zgodilo, čeprav človeška inteligenca že dolgo obstaja. Zakaj bi se to zgodilo v bližnji prihodnosti? Seveda imamo zdaj sisteme, ki določene dejavnosti opravljajo veliko hitreje in natančneje kot ljudje. Toda za človeški podobno inteligenco je potrebno veliko več kot hitrost in natančnost – potrebni sta intuicija in sposobnost samostojne izbire problemov. Tega, kako ustvariti takšno UI, pa si nihče ne predstavlja.

Turing je razmišljal o formalizaciji intuicije, da bi jo računalnik lahko uporabil, vendar ni vedel, kako to doseči. In tega nihče ne ve niti danes, kar ni presenetljivo, saj je prevladujoči znanstveni in tehnološki pristop UI dedič Turingovega razumevanja in deluje v okviru Turingove paradigme. Za preboj bi bila potrebna korenita sprememba pristopa in razumevanja UI, vendar nihče ne ve, kakšna naj bi ta sprememba v resnici bila. Podobno tudi še nihče ni opisal, kako bi se človeški podobna inteligenca – umetna ali naravna – lahko sama razvila v SI. Zato lahko rečemo, da je človeški podobna UI (za zdaj) le ‚mit' (Larson 2021) in znanstvena fantastika.

---

pobude in bi opravljal le tisto posebno vrsto dela, za katero je bil narejen. Vendar to ni pomanjkljivost stroja; ne želimo, da opravlja svoje delo, ampak naše. /…/ Drugič, zmogljivosti stroja imajo absolutne omejitve; zasnovan je bil za opravljanje določene stvari in ne more storiti nič drugega.« (Peirce 1887, 168–169; Larson 2021, 233) Turing se je problema pobude zavedal: »V naslednjem stoletju je Turing predlagal, da sprejmemo izziv in strojem vcepimo ‚izvirno pobudo' tako, da jih najprej programiramo, da se pogovarjajo z nami. Turing se je zavedal Peircevega ugovora, ki ga je v svojem članku iz leta 1950 pripisal lady Lovelace. Prav tako se je igral s preprostimi učnimi algoritmi, v petdesetih letih prejšnjega stoletja pa so se pojavile enoslojne nevronske mreže (imenovane perceptroni). Razumljivo je, da je Turing mislil, da bi se ugovoru Peircea in lady Lovelace morda lahko izognili tako, da bi ustvarili stroje, ki se učijo, zasnovane po človeških možganih. Če beremo *Computing Machinery and Intelligence*, dobimo vtis, da je učenje predstavljalo edini resnični izhod iz inherentnih omejitev strojev in edino resnično upanje, da bo stroj prestal Turingov preizkus. Vendar ga ni, to se ni zgodilo. Prepričanje, da se bo, da se mora, ima posledice za družbo, ki so zdaj postale več kot očitne. V zadnjem delu te knjige si bomo ogledali nekaj posledic mita o neizbežnosti – zlasti njegove škodljive učinke na znanost samo.« (233–234)

2.2 Abdukcija, zdravi razum in razumevanje

Lahko razlikujemo tri osnovne oblike sklepanja, ki jih ni mogoče zvesti eno na drugo: dedukcijo, indukcijo in abdukcijo.

V simbolih stavčne logike jih lahko predstavimo takole (Larson 2021, 172):

Dedukcija:	Indukcija:	Abdukcija:
P -> Q	P	P -> Q
P	Q	Q
Q	P -> Q	P

Sistemi globokega učenja temeljijo na statistiki. Statistika je v bistvu indukcija. Abdukcija je podobna ugibanju, vendar ni zgolj ugibanje. Je več kot le asociacija ali korelacija. Vendar nihče ne ve, kako ta ‚več' opisati na način, ki bi omogočil inženirstvo človeške inteligence. Abdukcija, zdravi razum in intuicija predpostavljajo razumevanje, ki ga nihče ne zna opisati na tehnično uporaben način. Razumevanje pa je jedro človeške inteligence. In dokler ga ne bomo znali ustrezno opisati, človeški podobne inteligence ne bomo mogli ustvariti.

Edino znanje, ki ga sistemu strojnega učenja lahko zagotovimo, je tisto, kar sistem iz podatkov lahko pridobi na povsem skladenjski način. To pomeni, da v sistemu obstaja slepa pega, ki povzroča napačne napovedi – saj tistega, česar sistem v podatkih ne more opaziti, ne pozna. (173) To slepo pego in napako v sklepanju lahko ponazorimo z zgledom purana (121–124), ki je induktivist in induktivno pride do usodno napačnega sklepa:

> »Ta puran je ugotovil, da je bil v svojem prvem jutru na farmi puranov nahranjen ob 9. uri. Vendar kot dober induktivist ni delal prenaglih sklepov. Počakal je, da je zbral veliko število opažanj dejstva, da je bil nahranjen ob 9. uri – in to se je zgodilo v različnih okoliščinah: ob sredah in četrtkih, v toplih in mrzlih dneh, na deževne in sušne dni. Vsak dan je na svoj spisek dodal novo opažanje. Končno je bila njegova induktivistična vest zadovoljna in izvedel je induktivni sklep, da ›sem vedno nahranjen ob 9:00‹. Žal se je ta sklep na svet večer nedvomno izkazal kot napačen, saj so mu, namesto da bi ga nahranili, prerezali vrat. Induktivni sklep z resničnimi premisami je privedel do napačnega sklepa.« (Chalmers 1982, 41–42)[6]

Ta primer lepo ponazarja »neumnost oblikovanja ‚navad asociacije' brez globljega znanja« (Larson 2021, 123). Strojno »učenje« ne temelji na razumevanju in ni učenje razumevanja, ampak zgolj na asociaciji ali korelaciji. Zato strojno učenje ni človeškemu podobno učenje, saj človeško učenje temelji na razumevanju in je učenje razumevanja. Napredek v učenju za ljudi pomeni večje razumevanje.

[6] Chalmersov primer purana, ki ga navajam, je različica znanega primera, ki ga je pred tem uporabil že Bertrand Russell (2001, 35–36).

Pri sistemih globokega učenja ne moremo govoriti o človeškemu podobnem učenju, ker jim ne moremo pripisati razumevanja. Njihov napredek ni napredek v razumevanju, ampak v količini informacij ali »dejstev«, ki jih imajo, in v (simulaciji) prilagodljivosti (Fuchs 2021, 33).

Razlika med človeško inteligenco in UI je jasno vidna pri prepoznavanju obrazov. Medtem ko UI za prepoznavanje obrazov potrebuje ogromno število ogledov, dojenčku za prepoznavo obraza lahko zadostuje le nekaj ogledov. Seveda dojenček ne more razložiti, zakaj je obraz prepoznal. To znanje je ‚znanje, kakó'. Hubert Dreyfus (1972) je zagovarjal dve trditvi, ki sta za to znanje pomembni (Oettinger 1972, xii-xiii): 1. Ljudje najprej zaznamo celoto,[7] ki jo šele potem, dodatno – če je potrebno – razčlenimo na dele. 2. Dojenčkovo prepoznavanje temelji na njegovem človeškem telesu. Zato bi morali roboti imeti dovolj človeškim podobna telesa, da bi imeli tudi človeški podobno inteligenco.

3. Partikularnost in zunanjostnost inteligence

Cholletov članek (2017) o neverjetnosti tako imenovane eksplozije inteligence je vzbudil veliko zanimanja. Doslej je doživel že več kot osemnajst milijonov ogledov. Izraz ‚eksplozija inteligence' je v šestdesetih letih prejšnjega stoletja uporabljal britanski matematik Irving John Good, ko je napovedoval nastop UI, ki da se bo razvila do nepredstavljive mere. Ta razvoj naj bi potekal s silno rastjo. Chollet dokazuje, da je takšna eksplozija malo verjetna. Njegova argumentacija je pomembna danes, ko se trditve o eksploziji inteligence – formulirane kot napovedi o pojavu singularnosti in podobnih pojavov – pojavljajo znova.

Cholletov pristop v članku je izkustven. Sklicuje se na podatke o evoluciji in zgodovinskem razvoju inteligence ter na različne druge empirične podatke, na primer o uspehu in dosežkih ljudi z nadpovprečnim inteligenčnim količnikom (v nadaljevanju IQ) in tako dalje.

Chollet pravi, da je pojem inteligence zelo pomembno najprej opredeliti – da bo jasno, o čem razpravljamo in kaj dokazujemo. Uvodoma jo opredeljuje kot reševanje problemov. Pravi, da je to začetna definicija, tako rekoč delovna definicija za začetek razprave. V celotnem članku to definicijo naprej razvija in dopolnjuje.

Jedro Cholletove argumentacije v članku sestavljajo naslednje tri trditve:

- Ni splošne inteligence.
- Inteligenca je situacijska in kontekstualna.
- Inteligenca je zunanjostna (eksternalistična).

Vse tri značilnosti so medsebojno povezane in prepletene.

Prva teza trdi, da je vsaka inteligenca prilagojena entiteti, ki to inteligenco ima – njenim potrebam, njenemu načinu življenja itd. To velja za inteligenco živali, za človeško inteligenco in tudi za UI, ki je prilagojena svojim namenom. Ni splošne

[7] To celoto lahko v smislu Gestalt psihologije označimo kot lik (nem. *Gestalt*).

inteligence v smislu, da bi bila primerna za vse naloge in vse namene. Človeška inteligenca je za nekatere živalske potrebe neuporabna – in obratno. Veliko je že prirojeno. Tretja trditev pomeni, da naša inteligenca ni (le) v naših možganih, ampak v znatni meri tudi v naših civilizacijah (Larson 2021, 27).

Druga izjava pravi, da na uporabo, izkoriščanje in razvoj naše inteligence vpliva situacija, v kateri se znajdemo. Chollet navaja primer ‚divjih' otrok, ki so odraščali med živalmi, na primer človeka, ki je živel med opicami. Na neki način je postal ‚opica' in se nikoli ni mogel popolnoma počlovečiti. Nikoli se ni naučil jezika, skakal je po mizah in drugod, zavračal kuhano hrano itd. Enako velja za druge ‚divje' otroke. Nekateri so se počlovečili bolj, drugi manj – a v vseh primerih precej pomanjkljivo.

Le malo ljudi z nadpovprečnim IQ v življenju doseže kaj izjemnega. Večina teh ljudi dela v ‚banalnih' poklicih, opravlja precej običajne in ‚povprečne' funkcije in naloge. Kaj in koliko bo kdo s svojo inteligenco dosegel, kako jo bo izkoristil in razvil itd., je odvisno od številnih dejavnikov: njegovega zgodovinskega in družbenega položaja, vzgoje, izobrazbe, naključnih srečanj s pravimi ljudmi, nahajanja na pravem mestu ob pravem času, ustrezne motivacije in prizadevanj oziroma teženj itd. Chollet navaja primere znanstvenikov, ki so dosegli pomembne rezultate, vendar njihov IQ ni bil nič posebnega – na primer fizik Feynman in Watson, soodkritelj DNA. Imeli so enak IQ kot mnogi drugi znanstveniki, ki nikoli ne bodo odkrili česa podobnega. In koliko ljudi z višjim IQ kot oni bo doseglo kaj njihovim dosežkom primerljivega? Navaja tudi besede slavnega biologa Stephena Jaya Goulda, ki ga je teža in velikost Einsteinovih možganov skrbela manj kot misel na to, koliko ljudi z Einsteinovim IQ je trpelo in še trpi v različnih okoljih, kjer svojih danosti nimajo možnosti razviti. Ali pa če se vrnemo v človeško zgodovino pred 10.000 leti – kako so ljudje živeli, kako so lahko razvijali svojo inteligenco. Govorili so preprost jezik, večina jih ni znala niti brati niti pisati itd. Kaj bo kdo dosegel ali odkril, je odvisno še od drugih zgodovinskih dejavnikov. Sklepna Cholletova ugotovitev – na podlagi različnih podatkov iz preteklosti in sedanjosti – je, da bo rast inteligence tudi v prihodnje zmerna, linearna in ne eksplozivna, eksponentna ali celo nepredstavljivo strma. Napovedi Raya Kurzweila in podobnih ‚prerokov' o nastopu SI se ne bodo uresničile.

Cholletov sklep je v nekaterih pogledih lahko presenetljiv. Vzemimo razvoj znanosti in tehnologije, ki močno vpliva na razvoj na vseh drugih področjih. Če upoštevamo, kako močna orodja, katerih zmogljivosti se nenehno povečujejo, imamo zdaj na voljo, kakšna je prefinjenost komunikacij, možnost izmenjave znanja in mreženja, kako se je število ljudi, ki se ukvarjajo z razvojem znanosti in tehnologije itd., povečalo, bi pričakovali hitrejši napredek. Vendar Chollet opozarja, da moramo videti celoto in upoštevati učinke, ki jih takšen razvoj prinaša. Z razvojem se kompleksnost znanstvenih problemov povečuje. Z naraščanjem znanja se povečuje čas, potreben za obvladovanje tega znanja, za izobraževanje znanstvenikov in spremljanje dosežkov v znanosti – da lahko ostanemo na tekočem in razvijamo nove stvari. Če upoštevamo vse navedeno, ni presenetljivo, da se po vseh metodološko merljivih kazalnikih znanstveno znanje na področju fizike v drugi polovici 20. stoletja ni povečalo bolj kot v prvi polovici istega stoletja, lahko pa bi navedli še druge primere. Po eni strani je res, da bolj ko ste izobraženi, hitreje bo vaše

znanje naraščalo, saj boste lahko uporabljali kompleksna orodja, matematično znanje, ustrezne simbole itd. Vendar je treba upoštevati celoto ter inteligenco in rast znanja razumeti kontekstualno in zunanjostno, da lahko vidimo povezanosti stvari in razumemo, zakaj je rast inteligence in znanja linearna. Podobno na področju ekonomije – po eni strani je res, da več ko imate denarja, več lahko dodatno pridobite. Toda če pogledamo celoto, z rastjo bogastva in naložb rastejo tudi drugi negativni dejavniki rasti – in končno empirični podatki kažejo, da je rast bogastva linearna, postopna, ne pa ‚eksplozivna'.

Chollet torej na več načinov ugotavlja, da rast inteligence ne bo eksplozivna, pač pa bo inteligenca rasla tako kot doslej – postopoma, linearno. Civilizacija in inteligenca se bosta razvijali še naprej: znanstveno znanje, število tehnoloških orodij in njihove zmogljivosti itd. bodo naraščali. Vendar do razvoja supermožganov, SI ali česa takšnega, kar bi pomenilo konec oziroma ukinitev človeštva, ne bo prišlo. Niti ne bo UI prevzela vseh nalog in oblasti – UI namreč nima lastnih potreb, namenov, motivov, želja, namer itd. V vseh teh pogledih in pomenih »eksplozija inteligence« po Cholletovem prepričanju ni verjetna.

4. Veliki podatki in znanost panja

Sestavni del današnje ideologije SI je ideologija velikih podatkov (angl. *big data*) in velike znanosti. Zagovorniki te ideologije trdijo, da bo zgolj kopičenje in združevanje oziroma povezovanje velike količine podatkov z uporabo trenutnega znanja in teorij privedla do nastanka UI, ki bo človeško raven najprej dosegla in jo nato (hitro) neskončno presegla – dosegla singularnost. Trdijo, da je nadaljnji razvoj teorije nepotreben ali celo nemogoč in da bo manjkajoče koščke za ustvarjanje UI priskrbela integracija velike količine podatkov s pomočjo zmogljivih računalnikov. Poleg poudarjanja potenciala velikih podatkov pristaši velike znanosti zagovarjajo koncept znanosti panja (angl. *hive science*). Ne potrebujemo več visoko inteligentnih in ustvarjalnih posameznikov z izvirnimi zamislimi in uvidi – ne potrebujemo več Einsteinov.[8]

Takšno stališče pomeni degradacijo človeškega uma in ljudi zgolj na asistente velikega računalnika, zgolj na njegove služabnike in vzdrževalce, ki ga morajo oskrbovati s potrebnimi podatki in storitvami za delovanje. Ne potrebujemo več ustvarjalnih osebnosti, ki bi razmišljale in imele nove ideje, ampak samo visoko usposobljene tehnike, servisno osebje in operaterje. Posledica uresničevanja takšnega stališča je, da se ljudje vse bolj prilagajajo stroju – in postajajo strojem vse bolj podobni. V smislu tipičnih mehanskih nalog stroj seveda ljudi presega, zato je pri takšni zasnovi razumljivo, da se ljudje stroju podredijo in postanejo njegovi strežniki. Takšno dojemanje pomeni zmanjševanje pomena posameznika in spodbuja oblikovanje družbe, ki ne vlaga več v ustvarjanje pogojev za razvoj posameznikov z izvirnimi za-

[8] Eden od zagovornikov takšnega koncepta je Henry Markram, ki je izjavil, da nas ovira splošno razširjeno prepričanje, da za razlago, kako delujejo možgani, potrebujemo Einsteina. Dejansko pa je tisto, kar resnično potrebujemo, da »postavimo na stran svoje ege in ustvarimo novo vrsto kolektivne nevroznanosti« (Markram 2013; navajam po Larson 2021, 246).

mislimi in teorijami, ki omogočajo resnične znanstvene in tehnološke preboje in napredek. Zato ni presenetljivo, da v okoljih in obdobjih, kjer prevladuje ideja velike znanosti, pravega napredka v znanosti in tehnologiji ni – čeprav zagovorniki velike znanosti stanje skušajo prikazati drugačno. Če se bo v prihodnosti model velike znanosti uveljavil še bolj, se bo ogroženost znanstvenega in tehnološkega razvoja in napredka povečala. Mnogi znanstveniki se tega zavedajo in opozarjajo na škodljivost koncepta velike znanosti ter aktivno organizirajo in usmerjajo nasprotovanje njenemu financiranju. Pogosto so projektom, ki potekajo po modelu velike znanosti, namenjeni ogromni zneski. Primer nasprotovanja veliki znanosti je peticija, ki jo je podpisalo več kot 500 uglednih znanstvenikov in jo naslovilo na Evropsko komisijo. Zahtevali so velike spremembe projekta Human Brain Project (Larson 2021, 267), ki se je uradno začel leta 2013 pod vodstvom že omenjenega nevroznanstvenika, dr. Henryja Markrama iz Švicarskega zveznega inštituta za tehnologijo v Lozani. Projekt je sprva vključeval več kot 150 institucij po vsem svetu (243).

5. Zaključek

Predstavil sem štiri Larsonove argumente proti HIT – imenujemo jih lahko argument splošnosti, argument intuicije, argument abdukcije in argument zdravega razuma –, ki so po mojem mnenju dobri razlogi za zavrnitev HIT. Poleg tega sem predstavil Cholletovo argumentacijo proti SIT. Ta temelji na izkustveni razčlenitvi, pri čemer upošteva partikularnost in zunanjostnost inteligence. Cholletovi argumenti v povezavi s predstavljenimi Larsonovimi argumenti zagotavljajo solidno utemeljitev za zavrnitev SIT. Pokazal sem, kako pomembno je inteligenco pravilno razumeti in opredeliti, saj ima nepravilno razumevanje lahko škodljive posledice, ki daleč presegajo zgolj akademsko področje. Takšno razumevanje lahko na primer zagotavlja osnovo za pridobivanje neupravičenih finančnih in drugih koristi za škodljive in nevarne ideologije, kot je denimo radikalni transhumanizem,[9] ter za instrumentalistična, razčlovečujoča in antipersonalistična stališča o pomenu in naravi človeške ustvarjalnosti in znanstvene dejavnosti.

Kratice

HIT – trditev, da bo prišlo do nastopa človeški podobne UI.
IQ – inteligenčni količnik.
SI – superinteligenca.
SIT – trditev, da se bo pojavila SI.
UI – umetna inteligenca.

[9] Za kritično razpravo o (radikalnem) transhumanizmu prim. Dittrich 2021; Sturm 2021; Pouliquen 2022; Petkovšek in Žalec, ur. 2021; Globokar 2019; Gregorčič 2023, 914; Klun 2019; 2024, 28; Osredkar 2019; Platovnjak in Svetelj 2019; Pohar 2019; Stegu 2019; Strahovnik 2019; Štivić 2021; 2023, 889–891; Vodičar 2019; Žalec 2019; 2023b.

Reference

Aristotle. 1995. Politics. Prev. Benjamin Jowett. V: Jonathan Barnes, ur. *The Complete Works of Aristotle*. Zv. 2, 1986–2129. Princeton: Princeton University Press

Bostrom, Nick. 2020. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Chollet, François. 2017. The implausibility of intelligence explosion. Medium. 27. 11. Https://medium.com/@francois.chollet/the-impossibility-of-intelligence-explosion-5be4a9eda6ec (pridobljeno 22. 8. 2023).

Chalmers, Alan. 1982. *What Is This Thing Called Science*. St. Lucia, AU: University of Queensland Press.

Dittrich, Tristan Samuel. 2021. Transhumanistisches Glückstreben und christliche Heilshoffnung: Ein Vergleich. *Zeitschrift für Theologie und Philosophie* 143, št. 3:452–474.

Dreyfus, Hubert L. 1972. *What Computers Can't Do: A Critique of Artificial Intelligence*. New York: Harper and Row.

– – –. 1991. *Being-in-the-World: A Commentary on Heidegger's* Being in Time, *Division I*. Cambridge, MA: The MIT Press.

Dreyfus, Hubert, in Charles Taylor. 2015. *Retrieving Realism*. Cambridge, MA: Harvard University Press.

Engels, Friedrich. 1975. Dialektik der Natur. V: Karl Marx in Friedrich Engels, *Werke*. Zv. 20, 307–570. Berlin: Dietz Verlag.

Fuchs, Thomas. 2021. *In Defense of Human Being: Foundational Questions of an Embodied Anthropology*. Oxford: Oxford University Press.

Gibson, James Jerome. 1979. *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Globokar, Roman. 2019. Normativnost človeške narave v času biotehnološkega izpopolnjevanja človeka**.** *Bogoslovni vestnik* 79, št. 3:611–628. https://doi.org/10.34291/bv2019/03/globokar

Good, Irving John. 1965. Speculations Concerning the First Ultraintelligent Machine. *Advances in Computers* 6:31–88. https://doi.org/10.1016/s0065-2458(08)60418-0

Gregorčič, Rok. 2023. Tehnološki razvoj v luči Habermasove etike diskurza. *Bogoslovni vestnik* 83, št. 4:911–922. https://doi.org/10.34291/bv2023/04/gregorcic

Klun, Branko. 2019. Transhumanizem in transcendenca človeka. *Bogoslovni vestnik* 79, št. 3:589–600. https://doi.org/10.34291/bv2019/03/klun

– – –. 2024. Problem religioznega izkustva v digitalno transformiranem svetu: Eksistencialno fenomenološki pristop. *Bogoslovni vestnik* 84, št. 1:19–32. https://doi.org/10.34291/bv2024/01/klun

Kurzweil, Ray. 2005. *The Singularity is Near: When Humans Transcend Biology*. London: Duckworth Overlook.

Landgrebe, Jobst, in Barry Smith. 2023. *Why Machines will Never Rule the World: Artificial Intelligence without Fear*. London: Routledge.

Larson, Erik J. 2021. *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do*. Cambridge, MA: The Belknap Press of Harvard University Press.

Markram, Henry. 2013. Seven Challenges for Neuroscience. *Functional Neurology* 28:145–151.

McCulloch, Gregory. 1995. *The Mind and Its World*. London: Routledge.

Miščević, Nenad. 1988. *Radnja i objašnjenje*. Zagreb: Hrvatsko filozofsko društvo.

Oettinger, Anthony G. 1972. Preface. V: Hubert Dreyfus. *What Computers Can't Do: A Critique of Artificial Intelligence*, xi-xiii. New York: Harper and Row.

Osredkar, Mari Jože. 2019. Religija kot izziv za transhumanizem**.** *Bogoslovni vestnik* 79, št. 3:657–668. https://doi.org/10.34291/bv2019/03/osredkar

Peirce, Charles Sanders. 1887. Logical Machines. *The American Journal of Psychology* 1, št. 1:165–170.

Petkovšek, Robert, in Bojan Žalec, ur. 2021. *Transhumanism as a Challenge for Ethics and Religion*. Zürich: Lit.

Platovnjak, Ivan, in Tone Svetelj. 2019. To Live a Life in Christ's Way: The Answer to a Truncated View of Transhumanism on Human Life. *Bogoslovni vestnik* 79, št. 3:669–682.

Pohar, Borut. 2019. Transhumanizem v službi človekove odgovornosti do stvarstva. *Bogoslovni vestnik* 79, št. 3:643–656. https://doi.org/10.34291/bv2019/03/pohar

Potrč, Matjaž. 1993. *Phenomenology and Cognitive Science*. Dettelbach: Verlag J. H. Röll.

– – –. 1996. Phenomenology and organic unity. V: Elisabeth Baumgartner, Wilhelm Baumgartner, Bojan Borstner, Matjaž Potrč, John Shawe-Taylor, Elisabeth Valentine, ur. *Handbook Phenomenology and Cognitive Science*, 185–197. Dettelbach: Verlag J. H. Röll.

– – –. 2004. *Dinamična filozofija*. Ljubljana: Filozofska fakulteta.

Pouliquen, Tanguy Marie. 2022. *Čar novih tehnologij in transhumanizem: 115 vprašanj*. Ljubljana: Družina.

Rowlands, Mark, Joe Lau, in Max Deutsch. 2020. Externalism About the Mind. V: Edward N. Zalta, ur. The Stanford Encyclopedia of Philosophy. Https://plato.stanford.edu/archives/

win2020/entries/content-externalism/ (pridobljeno 1. 2. 2024).

Sire, James W. 2002. *Izazov svjetonazora: pregled temeljnih svjetonazora* [*The Universe Next Door,* 1997]. Zagreb: STEPress.

Russell, Bertrand. 2001. *The Problems of Philosophy.* Oxford: Oxford University Press.

Stegu, Tadej. 2019. Transhumanizem in krščanska antropologija. *Bogoslovni vestnik* 79, št. 3:683–692. https://doi.org/10.34291/bv2019/03/stegu

Strahovnik, Vojko. 2019. Vrline in transhumanistična nadgradnja človeka. *Bogoslovni vestnik* 79, št. 3:601–610. https://doi.org/10.34291/bv2019/03/strahovnik

Sturm, Wilfried. 2021. Transhumanismus und Digitalisierung: Theologisch-anthropologische Perspektiven. *Zeitschrift für Theologie und Philosophie* 143, št. 3:425–451.

Štivić, Stjepan. 2021. Upanje v krščanstvu in transhumanizem. *Bogoslovni vestnik* 81, št. 4:849–856. https://doi.org/10.34291/bv2021/04/stivic

–––. 2023. Kiborgizacija: nastanek in razvoj pojma. *Bogoslovni vestnik* 83, št. 4:883–892. https://doi.org/10.34291/bv2023/04/stivic

Vodičar, Janez. 2019. Transhumanizem in katoliška vzgoja. *Bogoslovni vestnik* 79, št. 3:693–704. https://doi.org/10.34291/bv2019/03/vodicar

Turing, Alan. 1950. Computing Machinery and Intelligence. *Mind* 59, št. 236:433–460. https://doi.org/10.1093/mind/lix.236.433

Žalec, Bojan. 2019. Liberalna evgenika kot uničevalka temeljev morale: Habermasova kritika. *Bogoslovni vestnik* 79, št. 3:629–641. https://doi.org/10.34291/bv2019/03/zalec

–––. 2023a. Ali je umetna inteligenca inteligenca v pravem pomenu besede? Vprašanje psihičnih značilnosti in splošnosti. *Bogoslovni vestnik* 83, št. 4:813–823. https://doi.org/10.34291/bv2023/04/zalec

–––. 2023b. *Človečnost v digitalni dobi: Izzivi umetne inteligence, transhumanizma in genetike.* Ljubljana: Teološka fakulteta. Https://www.teof.uni-lj.si/uploads/Zalozba/ZnK86-Zalec-clovecnost_elektronska.pdf (pridobljeno 1. 2. 2024).