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.