UMETNA INTELIGENCA IN EKSISTENČNO TVEGANJE

ALEN LIPUŠ
Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
alen.lipus@pm.me

Sinopsis Kadar razmišljamo o možnosti umetne (super)inteligence, je treba v kontekstu problema nadzora vztrajati, da se umetni inteligenci dodeli učinkovit normativni okvir, da bi s tem povečali njeno usklajenost z najvišjimi družbenimi vrednotami. Pri takšni vrednostno občutljivi zasnovi umetne inteligence pa trčimo v problem spoznavne zaprtosti, ki otežuje uspešno implementacijo družbenih vrednot vanjo. V članku bomo (med drugim) raziskali naslednja vprašanja: Ali bodo racionalni stroji zagotovo presegli ljudi? Kakšen vpogled v delovanje lastnega uma lahko z njimi pridobimo? Kakšen je njihov moralni status? Jih lahko razumemo kot post-osebe? Ali je naša dolžnost, da ustvarimo superinteligentne sisteme? Kakorkoli, ne glede na to, koliko bi umetna inteligenca bila podobna ljudem, velja, da dokler v razlagi človeškega uma ne bo napredka, spoznavna vrzel med nami in umetno inteligenco ne bo nič manjša. Zdi se, da bolj kot smo v vsakdanjem življenju odvisni od tehnologije, bolj zapletena in nerazumljiva je za naše vsakdanje misli. Turingov test zato postane zastarel, saj predpostavlja delovanje inteligentnega uma, ki je podoben našemu. Toda inteligentnost lahko zavzame nam popolnoma tuje spoznavne oblike in nam je kot taka lahko popolnoma nedostopna in nerazumljiva; to pa je še dodaten razlog, zakaj je treba vztrajati pri vrednostni zasnovi inteligentnih sistemov.

Ključne besede:
post-osebe,
spoznavni dostop,
tveganje,
problem usklajenosti,
problem nadzora

Univerzitetna založba Univerze v Mariboru

DOI https://doi.org/10.18690/um.ff.11.2022.10
ISBN 978-961-286-675-4

ARTIFICIAL INTELLIGENCE AND EXISTENTIAL RISK

ALEN LIPUŠ
University of Maribor, Faculty of Arts, Maribor, Slovenia
alen.lipus@pm.me

Abstract When we think about the possibility of artificial (super)intelligence, we must insist, in the light of the control problem, that a normative framework integrated into artificial intelligence complies with the social values. But in such value-sensitive designed artificial intelligence, we are confronted with the problem of epistemic closure, which worsens the successful implementation of social values. In the article we deal with the following questions: Will rational machines surpass humans? What kind of insight into our mind can we gain through them? What is their moral status? Can they be understood as post-persons? Regardless of how similar artificial intelligence could be to humans, it holds that, if there is no breakthrough in the explanation of the human mind, the explanatory gap between us and artificial intelligence will not be any smaller. It seems that the more we depend on technology in everyday life, the more incomprehensible it is to us. The Turing test thus becomes obsolete, as it presupposes the workings of an intelligent mind that is like ours. But intelligence can take foreign cognitive forms, making it completely inaccessible to us, which represents an additional reason to insist on a value-sensitive design of intelligent systems.

Keywords:
post-persons,
epistemic access,
risk,
alignment
problem,
control problem

https://doi.org/10.18690/um.ff.11.2022.10 DOI
978-961-286-675-4 ISBN
University of Maribor Press

1 Zakaj bi stroji morali misliti?

Na vprašanje »Ali stroji lahko mislijo?« odgovarja Alan Turing (1912–1942), angleški matematik in filozof, s specifičnim diagnostičnim testom, ki pa za posledico sicer nima neposrednega odgovora na to vprašanje. Vprašanje namreč vsebuje problem, kako prepoznati mišljenje pri stroju: »Ali ne bi mogli stroji izvajati česa takega, kar bi morali opisati kot mišljenje, pa je zelo različno od tega, kar počne človek?« (Turing 1990: 63)

Ta problem je seveda zelo močan, vendar zgolj takrat, če se dopustimo zavesti prvotnemu vprašanju. Namesto »Ali stroji mislijo?« se vprašamo »Ali ostali ljudje mislijo?« in dobili bomo odgovor, kako soditi stroje. Na podlagi česa smo prepričani, da drugi ljudje mislijo? Na podlagi opazovanja in njihovega obnašanja.[1]

Podobno naj po Turingu velja tudi za stroje in tako se prvotno vprašanje spremeni v pogojnik: »Če stroj opravi Turingov test, potem mu moramo pripisati enake mentalne atribute kot ljudem«. Turingov test je tako preprosta stvar. V samem bistvu gre za test presoje človeškega sodnika med dvema kandidatoma. Sodnik ne ve, kateri kandidat je človek in kateri ni. Lahko si zamislimo, da je med njima pregrada. Sodnik nato izprašuje oba kandidata s preprostimi vprašanji. Če recimo zahteva izračun 3445 množeno s 7889, potem je smiselno, da bo stroj odgovoril s človeško zamudo. Če stroj zadovoljivo opravi test, potem nimamo nobenih razlogov, da mu ne bi pripisali določenih mentalnih lastnosti – kot jih pripisujemo sebi. Izmed bolj izpostavljenih kritikov Turingovega testa je John Searl z argumentom, ki temelji na miselnem eksperimentu 'Kitajska soba' (Markič 2021: 204).

V besedilu bomo naprej predstavili znano kritiko Turingovega testa, tj. miselni eksperiment »Kitajska soba«. Nato bomo podali kritiko nekaterih predpostavk, na katerih je razprava o mislečih strojih osnovana in ki omogočajo smiselno postavitev vprašanja *»Ali stroji lahko mislijo?«*. Čeprav se je Turing že sam zavedal, da je epistemska situacija strojev lahko radikalno drugačna od naše, se kljub temu ta vidik strojne kognicije zanemarja. Spoznavni (ne)dostop do strojne kognicije tako predstavlja še večji problem, sploh ob predpostavki, da inteligentnost pri strojih ne zavzema enake forme kot pri ljudeh. Iz te kritike se osredotočimo na eksistenčna in praktična vprašanja o umetni inteligenci. Obravnavamo problem nadzora, argument

[1] Glej tudi odgovor z »drugimi duhovi« (Searle 1990: 373).

pogube in oboje postavimo v širši kontekst hipoteze o ranljivem svetu. Na podlagi tega predstavimo vprašanje o možnosti tega, da je umetna inteligenca že prisotna oz. vsaj blizu. Podporo tega vprašanja razvijamo na primeru GPT jezikovnih modelov in pojmov nadgradljivosti ter transformativne umetne inteligence. Na koncu obravnavamo razširjeno stališče, da je razprava o eksistenčni ogroženosti neutemeljena in predstavimo primerjavo sklepanja med skeptičnimi teisti, ki zanikajo problem zla, in teoretiki pogube, ki vidijo v umetni inteligenci eksistenčno nevarnost. Čeprav je sklepanje obojih podobno, podamo razloge, da so teoretiki pogube vseeno bolj utemeljeni v svojem sklepanju.

2 Kitajska soba

John Searle (1932), ameriški filozof, si je Kitajsko sobo zamislil kot argument proti podvigom raziskovanja umetne inteligence[2] v poznih sedemdesetih. V mislih je imel delo Rogerja Shanka (Shank in Abelson, 1977). Shank in njegovi sodelavci so ustvarili program, ki lahko v najslabšem primeru simulira človeško zmožnost razumevanja pripovedi (Searle 1990: 362). To pomeni, da lahko podajo zadovoljiv odgovor glede neke zgodbe, v kateri podatek v odgovoru ni zajet. Če vam ponudimo hiter zaslužek za sodelovanje v poslu, ki sumljivo spominja na piramidno shemo, boste ponudbo zavrnili. Podobno je odgovarjal Shankov program, kar je njegove stvarnike napeljalo k misli, da dejansko razume pripoved in ne simulira zgolj človeškega razumevanja. Ta in tej podobne raziskave so Searla napeljale k ugovoru.

Kitajska soba je miselni eksperiment, ki poskuša zavreči glavno tezo tedanje kognitivne znanosti, ki definira mišljenje kot računsko obdelavo formalno določenih elementov (Searle 1990: 363).[3] Posledica te definicije je seveda prepričanje, da je možno z ustreznim programom realizirati močno UI.[4] Pri Kitajski sobi gre za podoben scenarij kot pri Turingovem testu, le da tu ne nastopajo samo stroji. Imamo

[2] V nadaljevanju UI. Razlikovati je potrebno pred različnimi opredelitvami umetne inteligence: umetna splošna inteligenca (USI; angl. *Artificial General Inteligence*) je opredeljena kot inteligenca (hipotetičnega) stroja, ki bi lahko uspešno opravljal katero koli intelektualno nalogo, ki jo lahko človeško bitje. Superinteligenca (SI) je USI, ampak onkraj človeških sposobnosti. Bostrom jo opredeljuje kot »*vsak razum, ki močno presega kognitivne sposobnosti ljudi na skoraj vseh področjih, ki nas zanimajo*« (Bostrom 2014: 26).

[3] Lahko jo razumemo kot kritiko simbolne umetne inteligence, tj. pristopa, ki mišljenje razume kot računanje s simboli. Tako kot program v računalniku realizira operacije z računskimi procesi, tako naj b bila ena izmed operacij, ki jo realizira možganski program (oz. nevralna koda), mišljenje (Markič 2021: 204).

[4] Močna UI predpostavlja, da je možno ustvariti takšno UI, kjer bodo prisotna mentalna stanja, kot so razumevanje, spoznavanje, zavedanje in zavest. Hkrati močna UI predpostavlja, da je možno razložiti človeško pojavnost brez ukvarjanja z možgani, saj človeško kognicijo pojmuje kot računsko manipulacijo formalnih simbolov. Šibka UI predstavlja zgolj močno orodje pri analizi in preverjanju podatkov (Searle 1990: 361).

ograjeno sobo in na eni strani mimoidoče kitajsko govoreče posameznike, na drugi strani pa posameznico, ki ne zna govoriti in tudi ne razume kitajščine. Znotraj sobe so hkrati navodila v njenem maternem jeziku, kako manipulirati z določenimi kitajskimi pismenkami, da bodo tvorile smiseln odgovor na zastavljeno vprašanje. Posameznica sledi slovenskim navodilom za razporejanje kitajskih simbolov, medtem ko računalnik sledi programu. Tako kot računalnik tudi naša posameznica ustvari vtis razumevanja, toda očitno je, da razumevanja kot takšnega – ni. Isto velja za računalnike in programe. Primer Kitajske sobe s tem pokaže na pomanjkljivost Turingovega testa. Iz napisanega lahko rekonstruiramo argument, ki se nanaša zgolj na ta miselni primer.

1. Če je močna UI resnična, potem obstaja program kitajščine, ki omogoča nekemu sistemu (človek, računalnik), ki zažene program, da razume kitajščino.
2. Jaz bi lahko sledil programu kitajščine brez razumevanja jezika.
3. ∴ Torej je močna UI neresnična. (1, 2 MT)

Miselni eksperiment je podpora drugi premisi in sklep govori v prid temu, da razumevanje ne more biti posledica zagona programa oziroma manipulacije s formalnimi simboli. Če namreč računalniški programi sledijo sintaktičnim navodilom, potem nimajo semantične vsebine. Toda: človeška mentalnost ima mentalno vsebino in je kot taka semantična, zato primer s Kitajsko sobo pokaže, da sintaksa ni zadosten pogoj za semantično vsebino, kar ima za svoj sklep to, da mentalnost ne more biti posledica simbolnega računanja oziroma programov.

3 Kritika vprašanja o mislečih strojih

Toda celotno diskusijo, v katero spada problem Kitajske sobe, lahko upravičeno problematiziramo. Vprašanja, ali stroji lahko mislijo; ali je močna UI možna oziroma ali je možna kakšna oblika USI; ali lahko razvijemo UI, ki bo imela vsaj nekatere mentalne vsebine, kot jih imamo mi – ki bo torej mislila, razumela, čutila, so popolnoma odveč. Prvič, človeški možgani so sintaktični stroj, ki pa so vseeno – čeprav še ne vemo, kako – baza za zavest. Drugič, tudi za druge ljudi nismo zagotovo

prepričani, da niso zgolj filozofski zombiji.[5] Podobno kot pri ljudeh, pri katerih sklepamo na notranje življenje iz vedenja, lahko to storimo pri strojih. To, da imamo odpor pri pripisovanju mentalnih vsebin strojem, je prej dokaz za naše nevrobiološke predsodke kot pa podpora dokaza opisanemu dejstvu. In tretjič, prisotnost mentalnih atributov pri strojih je metafizično vprašanje, ki pa ne sme zamegliti aktualnejšega, praktičnega vprašanja o tem, ali lahko dejansko ustvarimo stroj, ki Turingov test dejansko prestane. Postavljanje praktičnih vprašanj ima za posledico to, da razmišljamo o praktičnih učinkih in implikacijah[6], npr. o eksistenčni ogroženosti[7], pogoji katere niso vezani na to, da stroji morajo biti sposobni misliti.[8] Če torej lahko imajo stroji (brez da pridobijo človeško kognicijo) takšen družbeno disruptiven vpliv, potem moramo v ospredje postaviti vprašanja o praktičnih posledicah UI. Še četrtič, antropomorfizacija strojne kognicije kvečjemu omeji učinkovito razvijanje rešitev problema nadzora. Če namreč pričakujemo človeške lastnosti pri strojih, lahko to pričakovanje poveča verjetnost izdajniškega obrata.[9]

Naj izpostavimo tipičen primer preokupacije z metafizičnim vprašanje mišljenja pri strojih. Mindt in Montemayor (2020) sta predstavila razdelitev inteligentnih sistemov, vodilna vprašanja njunega eseja pa so na primer: »Kaj bi pomenilo, da bi sistemi prešli iz zgolj inteligentnega izvrševanja naloge v dobro izvedljivo nalogo?« in »Kakšna je povezava med zavestjo in inteligenco, tako da se lahko podajo posebne ocene o zavestni umetni inteligenci?« Ponujata sicer uporabno taksonomijo za navigacijo pri postavljanju teh vprašanj, a pri tem nekritično obravnavata možnosti za realizacijo modelov, ki so t. i. proizvajalci znanja.[10]

---

[5] Problem drugih duhov (Avramides 2020). Filozofski zombi je opredeljen kot natančen fizični dvojnik brez pojavnih lastnosti (Chalmers 1996: 95-96).

[6] Pri tem velja pripomniti, da sem spadajo tudi razprave o okoljski etiki predhodno usposobljenih jezikovnih modelov, ki za usposabljanje potrebujejo ogromne količine podatkov in računalniške moči – problem, ki je bil eden od poglavitnih razlogov za to, da je raziskovalka Timnit Gebru izgubila mesto pri Googlu. Sporna je bila objava raziskave »O nevarnostih stohastičnih papagajev: so jezikovni modeli lahko preveliki?« (2021), ki določa tveganja velikih jezikovnih modelov, usposobljenih za procesiranje neverjetne količine besedilnih podatkov. Prav tako praktične posledice zadevajo vse od posnemanja človeškega jezika, koherentnega pisanja in potenciala širjenja lažnih novic do že tako vsepristone algoritmizacije vseh digitalnih vidikov z edinim ciljem pritegnitve in ohranitve pozornosti.

[7] To so grožnje, ki bi lahko povzročile naše izumrtje in za katere velja – zaradi njihove kompleksnosti – da je običajno upravljanje tveganj neučinkovito (Bostrom 2013: 15).

[8] Lahko imajo zgolj velik družbenotransformativni potencial, ti. transformativna umetna inteligenca (Karnofsky 2016).

[9] V izvirniku *treacherous turn*, ki ga Bostrom (2014) opredeli kot idejo, da se SI v eni točki lahko nauči zavajanja.

[10] Z razliko od orodij znanja so proizvajalci znanja sposobno proizvesti dodaten iznos, npr. sposobnost metaučenja (Mind in Montemayor 2020: 14). Dosedanji sistemi UI se večinoma uvrščajo med orodja, saj so pod nadzorom človeka in nimajo notranjih stanj oz. potreb. Sleherna avtonomija – npr. za igranje igre – je vnaprej določena. Pri tem delata primerjavo s človeško avtonomijo in intencami za dosego ciljev. Vprašanja o potencialni disruptivnosti proizvajalcev znanja ju ne zanimajo.

Na tej točki naj torej zaključimo z razpravo o možnosti mislečih strojev, ker je takšna razprava vzpostavljena znotraj problematičnih in potencialno nevarnih antropomorfnih predpostavk o strojni kogniciji. V nadaljevanju se bomo torej posvetili praktičnemu vprašanju in pokazali smiselnost artikuliranja odgovorov na vprašanje eksistenčne ogroženosti tudi v primeru, ko nimamo opravka z notranjimi stanji strojev, ki so podobna človeški kogniciji.

4 Problem nadzora

V ospredje bomo tako postavljali praktična vprašanja, tj. problem nadzora, z njim pa tudi vprašanje, kako zagotoviti varnost pred morebitno disruptivno UI.[11] En odgovor je, da varnost pred SI zagotovimo empirično z opazovanjem njenega vedenja, ko je v nadzorovanem, omejenem okolju ('peskovnik'),[12] in da UI spustimo iz peskovnika samo, če vidimo, da se obnaša prijazno, sodelovalno in odgovorno.

Primer takega pristopa je sprotni sistem preverjanja etičnosti nekega UI sistema, v nasprotju s t. i. konceptom kurativnega »velikega rdečega gumba«[13] (Arnold in Scheutz 2018). Pojmovanje nadzora UI sistemov skozi dokončni izklop je nezadostno, saj je po eni strani prežeto s senzacionalistično obarvanimi scenariji glede potencialnih nevarnosti UI sistemov v daljni prihodnosti, po drugi strani pa VRG implicira, da je škoda že storjena, saj se na podlagi le-te upravitelji UI sistema odločijo za izklop. Boljši pristop, ki ga avtorja predlagata, je izoliran modul za sprotno samoocenjevanje in testiranje, s katerim se postavlja sprotna diagnostika, na podlagi katere se tveganje zniža oziroma prepreči.

Avtorja predlagata, da se ni treba ozirati na pogubne scenarije in SI, da razmišljamo o etičnosti UI sistema.[14] Nedavni dosežki strojnega učenja (zmaga v igri GO, slikovno prepoznavanje, procesiranje naravnega jezika, samovozeča vozila) kažejo na to, da so rigidna logična pravila robotike nezadostna za krotenje algoritemskega avtonomnega učenja. VRG je reakcija na to dognanje in način, kako nasloviti grožnje UI sistemov, preden ti postanejo uničujoči. Konkretno, VRG je način, kako preprečiti UI sistemu, da manipulira načine, preko katerih ga je možno ugasniti. Toda, kot izpostavljata avtorja, točka intervencije z VRG nastopi prepozno, šele

[11] Lahko je USI, SI ali zgolj algoritem, ujet v temno neskončnost (brezizhodno ponavljanje izvajanje neke operacije).
[12] V izvirniku *sandbox*.
[13] VRG.
[14] Problem, kot bomo videli, na katerega avtorja pozabljata, je ta, da je sama možnost SI pobijajoča kritika njunega 'etičnega preverjanja'.

tedaj, ko je sistem že "podivjan". Poleg tega mora ta tip intervencije preprečiti možnost, da sistemi, ki temeljijo na spodbujevalnem učenju (angl. *reinforcement learning*), ne uspejo prilagoditi nagradnih funkcij tako, da povečajo nagrade, kadar preprečijo svoj izklop. VRG prav tako ne naslavlja praktičnih vprašanj (UI sistemov, ki so že v uporabi) in tako ignorira trenutne probleme glede odgovornosti avtomatiziranih sistemov.

Primeri preprečevanja vplivov na VRG:

- spodbujevalni sistemi so lahko varno prekinljivi (angl. *safely interruptible*), kot predlagata Orseau in Armstrong iz skupine *Google Deep Mind* (2016), saj se lahko UI sistem nauči, da so prekinitve ovire pri doseganju nagrad;
- Reidl (2016) se osredotoča na izgon sistema v simulirano okolje kot način, kako mu preprečiti, da vpliva na človeško kontrolirano stikalo;
- Hadfield-Mennel (2016) predlaga negotovost ključnih nagradnih funkcij kot način, kako sistemu preprečiti vrednotenje VRG-ja.

Vsi ti pristopi predpostavljajo sistem spodbujevalnega učenja. Prvi pristop ilustrira načine, kako sistemu preprečiti, da dojema prekinitve kot ovire na poti do nagrad, drugi pristop se prekinitvam docela izogne v prid izgonu v simulacijo, tretji pristop pa predpostavlja netransparentnost nagradnih funkcij. Cilj pri vseh je enak; sistem ne sme asociirati VRG z grožnjo.

Toda kako vemo, kdaj pritisniti VRG? Ti pristopi namreč ne razvijajo strojne etike. Kaj točno torej sistem počne, preden ga izklopimo, je stvar nadzornikove interpretacije. Tudi če bi sistem imel integrirano etiko, kako bi vedeli, da ne gre zgolj za posnemanje zavoljo izognitvi pritiska na VRG?

Idealni način ocenjevanja UI sistema tako ni retrospektivna analiza, ampak sprotna verifikacija. Zaradi tega je sistemu potrebno onemogočiti, da stori več od zgolj blokiranja VRG. Sistem ne sme imeti dostopa do vodljivih, etičnih principov, tako so sami načini preverjanja sistemu nedostopni. Vse testiranje je torej sistemu nedostopno, čeprav se izvršuje vzporedno z normalnim delovanjem. To se doseže z izoliranim modulom etičnega jedra, tj. sloj nad strojno opremo, ampak pod virtualnim modulom (Arnold in Scheutz, 2018: 62). UI programska oprema v virtualnem modulu ima torej dostop do senzorjev in strojne opreme, ampak ne do

svojih notranjih operacij. S tem, ko se etično jedro umesti pod virtualni modul, na katerem tečejo razni operacijski sistemi in sam UI, UI sistem nima načina za zaznavo dodatnega etičnega modula. Na tak način UI sistem ne more anticipirati oziroma vedeti, kdaj je podvržen testiranju. Tako umeščeno etično jedro sestavljajo:

- simulacijsko okolje, ki je dovolj realistično, da UI sistem ni zmožen diskriminirati med realnimi vnosi in simulacijo;
- generator etičnih scenarijev, ki so neponovljivi in jih UI sistem ne zmore predvideti;
- minimalno delujoč sistem izklopa, ki lahko po potrebi izklopi UI sistem;
- povezave do senzorjev in aktuatorjev (*angl. actuators*), ki so prestregljive in preusmerljive in ki omogočajo shranjevanje, kloniranje in zagon virtualnih okolij, v katerih se UI sistem nahaja.

V vsakem primeru testiranje UI sistema v danem trenutku naredi kopijo sistema, ki je podvržena simuliranemu scenariju in testirana. Vzporedno s tem prvotni UI sistem deluje normalno. Brezhibno kloniranje in preusmeritev senzoričnih vnosov na simulacijo zahteva grafično in scenarijsko natančnost slednje, saj lahko v nasprotnem primeru klonirana verzija uvidi, da je podvržena testiranju (Arnold in Scheutz 2018: 66). Če testirana verzija test opravi, se vzporedni proces testiranja zaključi, prvotni UI sistem pa nadaljuje svoje delovanje. Če testa ne opravi, se lahko zaradi slojevite arhitekture izklopita tako testirana kot prvotna verzija UI sistema.

Tak način sprotnega testiranja, ki temelji na slojeviti arhitekturi modulov in izoliranosti etičnega jedra, je nujen, saj se ne smemo zanašati zgolj na nujne izklope v primeru, ko nek UI sistem že predstavlja grožnjo oziroma je že izpeljal neko operacijo, ki jo upravitelj presodi kot dovolj sporno, da zahteva pritisk na VRG. Etika UI sistemov, skratka, ne sme temeljiti zgolj na snovanju sistemov za nujni izklop, ampak zahteva kontinuirano testiranje, ki je imuno na manipulacijo nagrad s strani spodbujevalnih sistemov.[15]

[15] V okviru projekta, ki je preučeval razvoj (simuliranih) organizmov, je Claus Wilke (s skupino raziskovalcev, 2001) želel omejiti stopnjo razmnoževanja digitalnega organizma. Sistem je programiral tako, da se po vsaki mutaciji zaustavi, izmeri hitrost replikacije mutanta v izoliranem testnem okolju in mutanta izbriše, če se je repliciral hitreje kot nadrejeni. Vendar so se organizmi razvili, da so prepoznali, kdaj so bili v testnem okolju, in se 'igrali mrtve' (zaustavitev ponovitve), da jih ne bi izločili in jih namesto tega zadržali v populaciji, kjer bi se lahko še naprej razmnoževali zunaj testnega okolja. Ko je to odkril, so randomizirali vnose testnega okolja, tako da ga ni bilo mogoče tako enostavno zaznati, toda organizmi so razvili novo strategijo za izvajanje nalog, ki imajo veliko verjetnost, da bi lahko pospešile njihovo razmnoževanje. Tako so vsaj nekatere različice prestale testiranje. (Wilke in drugi, 2001)

Pomanjkljivost ideje sprotnega testiranja je, da predpostavlja, da je to izvedljivo. Lepo in zaželeno vedenje v nadzorovanem okolju sestavljanja je t. i. konvergentni instrumentalni cilj tako za prijazne in neprijazne UI. Sistem umetne inteligence (ali podsistem) se lahko nauči zaznati, kdaj je nadzorovan/testiran, in spreminja svoje vedenje med nadzorom/testiranjem, tako da njegove neželene lastnosti (oblikovalcev) ostanejo neopažene (Bostrom 2014: 136).

> Ideja je, da varnost superinteligentne umetne inteligence potrdimo empirično z opazovanjem njenega vedenja, ko je v nadzorovanem, omejenem okolju (»peskovnik«), in da umetno inteligenco spustimo iz škatle samo, če vidimo, da se vede na prijazen, sodelovalni, odgovoren način. Pomanjkljivost te ideje je, da je lepo vedenje v škatli konvergentni instrumentalni cilj za prijazne in neprijazne umetne inteligence. Neprijazna umetna inteligenca z zadostno inteligenco se zaveda, da bo svoje neprijazne končne cilje najbolje uresničila, če se bo na začetku obnašala prijazno, tako da bo izpuščena iz škatle. Obnašati se bo začela neprijazno šele, ko ne bo več pomembno, ali mi to izvemo; to je, ko je umetna inteligenca dovolj močna, da je človekovo nasprotovanje neučinkovito. (Bostrom 2014: 146)

Neprijazna umetna inteligenca z zadostno inteligenco se zaveda, da bo svoje neprijazne končne cilje najbolje uresničila, če se bo na začetku obnašala prijazno, tako da bo izpuščena iz testnega okolja. Neprijazna bo postala šele tedaj, ko bo umetna inteligenca dovolj močna, da je človekovo nasprotovanje neučinkovito (Bostrom 2014: 147).

5 Argument pogube

Možnost izdajniškega obrata zaostri eksistenčno grožnjo, t. i. vrsto groženj, ki predstavljajo nevarnost celotni prihodnosti človeštva. Argument iz pogube[16] je kombinacija možnosti UI – prednosti prvega gibalca (biti v poziciji, da UI počne, kar želi), teze o pravokotnosti (to, kar želi, je lahko karkoli) in konvergentnih instrumentalnih vrednot (ne glede na želje bo delovala pri pridobivanju virov in izničevanju nevarnosti zanjo)[17] človeštvu nakazujejo na propad.

[16] V izvirniku *Doomsday Argument*, tako v poglavju »Je privzeti izid poguba?« (angl. *Is the Default Outcome Doom*) (Bostrom 2014: 140).

[17] Človeška bitja predstavljajo koristne viri kot so »priročno nameščeni atomi« in ostali lokalni viri, ki jih izrabljamo. (Bostrom 2014: 116).

Situacija prvega gibalca bo za UI edinstvena strateška priložnost, saj bo v poziciji, da ustvari nov svetovni red, v katerem obstaja ena sama, najvišja raven odločanja. Med njegove pristojnosti bi spadala (1) sposobnost preprečevanja kakršnih koli groženj lastnemu obstoju in nadvladi ter (2) sposobnost učinkovitega nadzora nad glavnimi značilnostmi svojega področja (Bostrom 2014: 141).

Takšna stopnja inteligence je v skladu s skoraj vsakim končnim ciljem. Zatorej ne moremo domnevati, da bo imela katero od dobronamernih vrednot ali ciljev.[18] Težko je opaziti, ali je umetna inteligenca nevarna s svojim vedenjem v času, ko bi jo lahko izklopili, ker imajo umetne inteligence konvergentne instrumentalne razloge, da se pretvarjajo, da so varne in prijazne, četudi niso. Zato bi prva SI zlahka imela neantropomorfne končne cilje in bi verjetno imela instrumentalne razloge za nadaljevanje pridobivanja virov brez konca (Bostrom 2014: 116).

6 Hipoteza o ranljivem svetu

Problematiziranje SI spada v širšo domeno eksistenčnega raziskovanja[19], ki jo Bostrom (2019) poimenuje hipoteza o ranljivem svetu. Osnovna ideja je ta, da ljudem znanstveni in tehnološki napredek nudita vse več različnih zmožnosti, ki utegnejo destabilizirati civilizacijo. Hipoteza ranljivega sveta (HRS) se tako glasi:

> HRS: "Če se nadaljuje tehnološki razvoj, bo v določenem trenutku dosežen nabor zmogljivosti, zaradi katerih je civilizacijsko opustošenje izjemno verjetno, razen če civilizacija izstopi iz polanarhičnega privzetega stanja." (Bostrom 2019: 457)

Privzeto stanje označuje stanje, v katerem imajo družbe omejene kapacitete za preventivne politike in globalni nadzor ter razpršene motivacije. Če izstop iz privzetega stanja ni mogoč, tj. če civilizacija ni zmožna implementirati preventivnih ukrepov, potem na podlagi HRS sledi gotov propad civilizacije. Del preventivnih ukrepov je ravno problem nadzora. Toda, neodvisno od eksistenčne pomembnosti

[18] Tudi ob predpostavki dobronamernosti se UI interpretacija človeške blaginje lahko radikalno razlikuje od človeškega pojmovanja. Na tej točki določenih miselnih eksperimentov ne bomo omenjali, ker predstavljajo t. i. informacijsko nevarnost (angl. *informational hazard*), v neakademskih razpravah tudi t. i. memetična nevarnost. Bostrom razdela tipologijo informacijskih nevarnosti gleda na potencialno škodo zaradi prenosa znanja (Bostrom 2012).

[19] V profilnem članku za New Yorker so zapisali, da Bostrom vodi inštitut kot nekakšno filozofsko radarsko postajo: bunker, ki pošilja navigacijske impulze v meglico možne prihodnosti (Khatchadourian 2016).

ukvarjanja s problemom nadzora in pionirskih naporov pri vpisovanju človeških normativnih, etičnih parametrov v ustroj umetne inteligence, ostaja možnost, da je vse to zaman. Omenjeni projekti namreč temeljijo na antropomorfnih predpostavkah o umetni inteligenci in njeni kogniciji in če smo kognitivno zaprti do preprostih algoritmov (kot nakazuje t. i. *problem črne škatle*[20]), kako lahko pričakujemo neoviran spoznavni dostop do nečesa takšnega, kot je SI? Legitimno lahko postavimo vprašanje: Kaj če je splošna umetna inteligenca že tukaj?[21]

Da bi lahko odgovorili na zgornje vprašanje, bomo pogledali razvoj GPT[22] modelov in z njihovo pomočjo ilustrirali njeno smiselnost. Ob tem bomo predstavili t. i. hipotezo o nadgradljivosti (angl. *scalability hypothesis).* Slednja je namreč ključnega pomena za težo zgornjega vprašanja, saj ponuja utemeljitev pozitivnega odgovora. Hipoteza o nadgradljivosti se nanaša na napoved, da bomo postopoma videvali vse večje preboje zmožnosti UI s povečevanjem nabora parametrov[23] in podatkov.[24] Hipoteza o nadgradljivosti nudi podporo pojmu t. i. transformativne[25] UI. Pokazali bomo, da lahko na podlagi GPT primera smiselno postavimo zgornje vprašanje o aktualnosti t. i. transformativne UI, ki je sposobna imeti vpliv na človeštvo, kar je primerljivo z industrijsko revolucijo.

[20] Za nas netransparenten sistem, za katerega ne vemo, kako je prišel do rezultatov (Markič 2020: 209).

[21] Verjetnost vzleta (predvsem na blogih se uporablja izraz *foom*) oz. eksplozija rekurzivne samoizboljšave. Ob predpostavki, da bi lahko en sam programski projekt, ki se začne z majhnim delom svetovnih virov, v nekaj tednih postal tako močan, da prevzame svet. Kaj je bolj verjetno, da živimo v času tik pred S(U)I ali, da se je to že zgodilo? Čas tik pred nastankom SI (ne glede na to ali mi to vemo ali ne – koncepcija zavajanja (angl. *the conception of deception)*) nedvomno zavzema manjši časovni interval kot čas po SI, iz tega sledi, da je bolj verjetno, da živimo v času po vzletu SI. Podobno lahko na enak način razmišljamo, da je bolj verjetno, da živimo v simulaciji kot pa v bazični realnosti, tik pred SI vzletom. Zakaj bi SUI sploh želela zaganjati simulacije vesolja pa je vprašanje, na katerega zaradi informacijske nevarnosti ne želimo podati odgovora.

[22] Splošni vnaprej usposobljeni jezikovni modeli (angl. *General Pre-Trained Transformers*). BERT, 500-krat manjši model (tehnično je način učenja) od GPT-3, je recimo eden izmed jezikovnih modelov, ki ga uporablja Google za svoj iskalnik.

[23] Parametri so spremenljivke, ki se uporabljajo za nastavitev in prilagoditev modelov umetne inteligence.

[24] Turingov nagrajenec Geoffrey Everest Hinton, ena od pomembnih osebnosti v razvoju globokega učenja, se je v čivku pošalil, da: »ekstrapolacija spektakularne zmogljivosti GPT-3 v prihodnost kaže, da je odgovor na življenje, vesolje in vse samo 4398 milijard parametrov.«

[25] Lahko je SI ali preprosto samoizboljševalni algoritem, označuje pa že prej omenjeno UI z disruptivnim potencialom za družbo in svet.

7 GPT[26] in vključenost v svet

Preteklo leto 2020 je bilo prvo, v katerem so jezikovni modeli UI postali ekonomsko uporabni. Zlasti GPT-3 je pokazal, da imajo veliki jezikovni modeli presenetljivo jezikovno sposobnost opravljanja najrazličnejših uporabnih nalog. Pričakuje se, da bodo jezikovni modeli postajali vse kompetentnejši, tako da bodo najboljši modeli leta 2020 v primerjavi z njimi videti dolgočasno in preprosto.[27] To pa bo odklenilo aplikacije, ki si jih danes težko predstavljamo. Leta 2021 se bodo jezikovni modeli začeli zavedati vizualnega sveta. Samo besedilo lahko namreč izraža veliko informacij o svetu, vendar je nepopolno, saj živimo tudi v vizualnem svetu. Naslednja generacija modelov bo tako lahko urejala in ustvarjala slike kot odziv na vnos besedila. Predpostavlja se, da bodo besedilo bolje razumeli zaradi številnih slik, ki so jih videli. Ta sposobnost skupne obdelave besedila in slik bi morala narediti modele pametnejše.[28] Ljudje smo izpostavljeni ne le temu, kar preberemo, ampak tudi tistemu, kar vidimo in slišimo. Če lahko modele izpostavimo podatkom, podobnim tistim, ki jih absorbiramo ljudje, bi se morali naučiti pojmov, ki so podobni našim.

S tem se predpostavlja t. i. hipoteza naravne abstrakcije oziroma naravnih pojmov. Za namene tega besedila je ne bomo problematizirali, čeprav vsebuje nekritično obravnavo same narave pojmov. Predpostavlja namreč, da bodo novi modeli UI s tem, ko bodo opremljeni s senzorji in tako vključeni v svet, prihajali do podobnih pojmov kot ljudje. Konkretno, predpostavlja se tesna povezanost med vsebinami abstrakcije in načini vnosa – jezikovni model, ki vizualno procesira besedilo, je različno določen pri pridobitvenem načinu informacij kot pa model, ki besedilo pridobi preko nesenzoričnih načinov. Hipoteza naravne abstrakcije je, ker temelji na 'utelešenosti', predvsem izpostavljena kritiki računalniškega mišljenja iz perspektive, ki se osredotoča na abstraktno naravo znanstvenih modelov, med katere spadajo tudi računalniški modeli možganov (Chirimuuta 2020: 424). Navadno so bili ugovori računalniški teoriji mišljenja osnovani na fenomenologiji dvajsetega stoletja, s posebnim poudarkom na utelešenje in vdelanost inteligence (Dreyfus 1972). Prej

[26] GPT spadajo v t. i. pozni drugi val UI, ki se v nasprotju s prvim (močna UI, ki razume mišljenje kot simbolno manipulacijo), obračajo k induktivnemu sklepanju – na podlagi baz podatkov, izkušenj in interakcije z okoljem. Programe, kot so navodila za simbolno manipulacijo, so zamenjali algoritmi za strojno učenje. Če prvi val zaznamuje dobro definirano okolje (npr. igre, programi, simboli) in deduktivna logika, potem drugi val zaznamuje odprt empiričen svet, poln negotovosti, za katerega je ustreznejša indukcija (Markič 2021: 209).

[27] Tako Ilya Sutskever, soustanovitelj OpenAI, komentira za spletno revijo The Batch.

[28] V skladu z Dreyfusovo kritiko prvega vala UI, da za vsakdanje znanje potrebujemo drugačno obliko predstavitve informacij. Po njegovem je eden izmed ključnih pogojev za to, da bi UI posedovala razumevanje, vključenost v svet (Markič 2021: 209).

omenjena kritika računalniškega mišljenja preko problematizacije znanstvenih modelov pa temelji na Whiteheadovem pojmu zmote napačno postavljene konkretnosti[29] in se tako pogoju utelešenosti popolnoma izogne.

8 Zgodba o GPT

Začelo se je leta 2018, ko so pri OpenAI izdali prvi GPT model, ki je imel velik vpliv na tedanjo UI skupnost. Še večji vpliv je imel izid GPT-2 modela v začetku leta 2019, čeprav ga pri OpenAI niso želeli izdati v celoti zaradi zaskrbljenosti zlonamerne uporabe. (Radford et al. 2019)

GPT-2 je bil predhodno usposobljen z raznoliko, 40 GB veliko vsebino, postrgano z interneta – z enim preprostim ciljem: da predvidi naslednjo besedo glede na vse prejšnje besede v nekem besedilu.

Tudi okrnjen model je produciral presenetljive rezultate, kot je znana zgodba o samorogih, pri kateri je človeški vnos besedila v slogu naslova in podnaslova zgodbe o odkritju samorogov na območju gorovja Andov. GPT-2 je na podlagi te vsebine in sloga izvozil slogovno primerljivo in skladno besedilo, ki vsebuje vse elemente tipske reportaže, tj. od biološke opredelitve samorogov, imen glavnih raziskovalcev, njihovih citatov in do spekulacij glede izvora samorogov. Nič od tega ni bilo zajeto pri vnosu. Vzorci tekstov, kot je omenjena zgodba s samorogi, imajo pomembne posledice, saj je velike jezikovne modele vedno lažje usmerjati v prilagodljivo, prilagojeno in skladno ustvarjanje besedila, ki bi se lahko nato uporabljalo na številne koristne in zlonamerne načine.

Naslednje leto, 2020, izide GPT-3, ki je, če pogledamo njegovo velikost, 100-krat večji od predhodnika. GPT-3 ima 175 milijard parametrov. Torej je GPT-3 100-krat večji od predhodnika GPT-2, ki je bil že izjemno velik, ko se je pojavil leta 2019. Povečanje števila parametrov 100-krat z GPT-2 na GPT-3 ni prineslo le količinskih razlik. GPT-3 ni le močnejši od GPT-2, ampak je tudi zmogljivejši – na drugačne načine; med obema modeloma je kvalitativni preskok. GPT-3 lahko počne stvari, ki jih GPT-2 ne more. Če se GPT-3 lahko nauči učiti, kdo ve, kaj lahko prinese

[29] Zmota zamenjave abstrakcij znanosti za konkretne stvari na svetu, iz katerih abstrakcije izhajajo (angl. *fallacy of misplaced concreteness*) (Chirimuuta 2020: 430; Whitehead 1928: 66).

GPT-4; morda bomo videli prvo nevronsko mrežo, ki je sposobna resničnega sklepanja in razumevanja.

9 GPT in transformativna UI

GPT verjetno ne bo postal SI, ne glede na to, kako velik je GPT-3, je namreč splošno orodje, ki ga je mogoče izvesti za različne naloge in je morda res sposobnejše izvajati večje število nalog, ampak spontano ne bo sposoben početi stvari, ki so ključne za pristno SI – na primer razumevanje vzročnosti – saj preprosto nima potrebne arhitekture.

Kar bi bilo sicer zelo prepričljivo, bi bil npr. izjemno dober klasifikator slik in videov, ki je opremljen z umetnimi čutili za interakcijo s svetom in dobro obdelavo naravnega jezika, ki je sposoben prepoznati tudi relacijske besede in dejanja, npr. prepoznavanje videoposnetka osebe, ki položi vrček na mizo. Predstavi se mu vrček in ukaže, »Daj vrček na mizo«, pri čemer besede razčleni v besedilo in v svoji bazi podatkov o usposabljanju poišče videoposnetke, ki so zelo podobni temu opisu, pripravi celoten video, prepozna vizualni vhod skodelice kot isti simbolni predmet kot sestavljeni vrček, nato pa, da se vizualni vnos natančno ujema s tem agregatom, s poskusi in napakami in s strojnim učenjem spozna pravilen način uporabe udov – da vrč postavi na mizo. In nato, kar je ključnega pomena, shrani pravilne v spominu in na splošno izmeri stopnjo učinkovitosti iz svojih veščin (se izboljša pri 'dajanju', pri 'vklopu' itd.), ki jih je mogoče uporabiti za prihodnje naloge. Zdi se, da bi ga to približalo nečemu, kar bi lahko označili z 'razmišljanjem'.

Tak sistem še vedno ne bi nujno razumel vzročnosti same po sebi, vedel pa bi, kakšne stvari mora oddati, da izpolni svojo funkcijo koristnosti in kdaj mora ukrepati. A kljub temu da GPT in ostali modeli ne bi postali SI, še več, četudi SI kot taka sploh ni možna, postaja na podlagi zmogljivosti teh modelov, ki dajejo podporo hipotezi o nadgradljivosti, pojem transformativne UI vse bolj verjeten in smiseln. Transformativna UI je UI, ki lahko pospeši družbeni napredek/prehod, primerljiv s kmetijsko ali industrijsko revolucijo (ali celo takšnega, ki bi bil pomembnejši od nje) (Karnofsky 2016). Transformativna UI ni nujno SI in je v tem oziru nevtralna v primerjavi s potezami človeškega uma. Za transformativno AI torej ne šteje, da primerja človeške lastnosti zavedanja, čustev, razumevanja in podobno. Vse, kar je za transformativno UI pomembno, je to, da je sposobna privesti do znatnih sprememb na svetu.

10 Fideizem in UI

Zgornjo razpravo o praktičnih vidikih razvoja UI nekateri označujejo kot vnebovzetje za piflarje,[30] ji očitajo neutemeljenost (Bringsjord et al. 2012) in je ne obravnavajo pogosto v akademskih kontekstih, čeprav se to spreminja. Kljub temu je razvoj drugega vala UI s seboj prinesel določene novosti in premike, s katerimi so pokazali, da lahko s primerno velikostjo UI modelov že sedaj produciramo disruptivne dogodke, npr. prepričljive lažne novice. Vendar je kljub praktičnim napredkom UI vseeno treba odgovoriti na ta sentiment o spoznavni neutemeljenosti svarilcev pred pogubo zaradi UI. Spodaj povzemamo in odgovarjamo na primerjavo svarilcev UI s teističnimi skeptiki (izvaja jo Danaher 2015), pokazali bomo neustreznost te primerjave in še na ta način pokazali spoznavno utemeljenost zgornje razprave in nasploh doslednost svarilcev pred UI.

Osrednji korak za skeptične teiste je sklepanje z videza na realnost zavoljo zagovora Božje moralnosti – četudi se nam zdi, da je v svetu zlo, iz tega ne sledi, da je to tudi res, saj je lahko dozdevno zlo zgolj funkcija za neko drugo (višje) dobro (Denahar 2015: 233). Bostrom (2014) z izdajniškim obratom sicer ubira podobno logiko neupravičenosti induktivnega sklepanja iz empirično danih dokazil, a Bostromov izdajniški obrat kot blokada induktivnega sklepanja iz dozdevne dobronamernosti na dejansko dobronamernost ni enaka sklepanju skeptičnih teistov. Podobnost med UI teoretiki pogube in skeptičnimi teisti sestoji v zguljeni filozofski tezi o različnosti med videzom in realnostjo in spoznavnem opozorilu, ki iz tega sledi, in sicer, da moramo biti previdni pri tovrstnem induktivnem sklepanju iz empiričnega videza na pravo naravo stvari. Danaher (2015), ki razvije to primerjavo, se tako vpraša: »Zagotovo obstajajo nekakšni empirični dokazi, ki bi ga zadovoljili, da UI ne predstavlja tveganja za ljudi?« (Danaher 2015: 239) Pri tem pa ne upošteva ravno te poante, da v luči eksistenčne nevarnosti (Bostrom 2013 in 2014) ni racionalno zaupati nobenim empiričnim dokazilom.

V obeh primerih imamo torej prvotno prepričanje, ki je velikega pomena (obstoj Boga v prvem primeru in možnosti UI pogube v drugem). Ta prepričanja se nanašajo na obstoj nadmočnih in superinteligentnih agentov (Bog ali SI). Prepričanja izpodbijajo nekateri ugovori (problem zla in ugovor empiričnega, sprotnega preizkušanja). Oba ugovora temeljita na ideji, da lahko zanesljivo sklepamo (čeprav

[30] Izvorni naslov je *Rapture of the Nerds* (Doctorow in Stross 2012).

induktivno) od videza dogodka do njegove dejanske narave. V obeh primerih se blokira sklepanje od "navideznega" do "dejanskega" s sklicevanjem na naše kognitivne omejitve: ne vemo popolnoma, kako se to, kar opazimo pri videzu (zlo, vedenje UI), lahko poveže z drugim (nezamisljivimi) končni cilji. Možno je namreč to, da zaradi nepopolnega kognitivnega dostopa (ali do misli Boga ali do UI), obstajajo drugi končni cilji, do katerih nimamo spoznavnega dostopa (Denahar 2015: 235; Bostrom 2014: 158).

Za razliko od skeptičnih teistov je spoznavno opozorilo pri sklepanju z videza na realnost v primeru UI utemeljeno z dejansko eksistenčno nevarnostjo, medtem ko je pri skeptičnih teistih prej orodje za podporo dogme. Prav tako je ena izmed praktičnih posledic za skeptičnega teista moralna paraliza (prekiniti trpljenje in s tem potencialno prekiniti višje dobro ali pa dovoliti odvijanje trpljenja). Teoretik UI pogube tovrstni paralizi ni podvržen; to, da sodeluje pri razpravah, Inštitutih za nadzor UI in ozaveščа o tej problematiki, je skladno s prepričanjem o potencialni eksistenčni nevarnosti UI. Še več, zavora induktivnega sklepa in iz tega izhajajoča previdnost temelji pri UI teoretikih izgube na nomološki možnosti pogube, medtem ko pri skeptičnih teistih zavora sklepanja temelji na obrambi nečesa, ki krši oz. je onkraj nomološke realnosti. In še, pri UI teoretikih pogube previdnost izhaja iz skrbi za človeštvo in preživetje naše vrste, medtem ko pri skeptičnih teistih zavora induktivnega sklepanja ne izhaja iz previdnosti in človekoljubnosti, ampak iz zavezanosti in obrambe Boga.

11 Sklep

V besedilu smo naprej predstavili znano kritiko Turingovega testa, tj. miselni eksperiment 'Kitajska soba'. Nato smo podali kritiko nekaterih predpostavk, na katerih je razprava o mislečih strojih osnovana in ki omogočajo postavitev vprašanja *»Ali stroji lahko mislijo?«*. Izpostavili smo pogled, da lahko spoznavni (ne)dostop do strojne kognicije predstavlja še večji eksistenčni problem, sploh če predpostavimo, da lahko strojna kognicija zavzema nam tuje oblike. Praktična vprašanja in razpravo o eksistenčni nevarnosti smo postavili v kontekst hipoteze o ranljivem svetu. Iz tega smo predstavili in utemeljevali smiselnost vprašanja o možnosti tega, da je umetna inteligenca že prisotna oz. vsaj blizu. S primerom GPT jezikovnih modelov in pojmov nadgradljivosti in transformativne umetne inteligence smo legitimnost tega vprašanja utemeljevali. Na koncu smo obravnavali razširjeno stališče, da je razprava o eksistenčni ogroženosti neutemeljena. Predstavili smo primerjavo sklepanja med

skeptičnimi teisti, ki zanikajo problem zla, in teoretiki pogube, ki vidijo v umetni inteligenci možnost eksistenčne nevarnosti. Čeprav je sklepanje obojih podobno, smo podali razloge, da so teoretiki pogube vseeno bolj utemeljeni v svojem sklepanju. Slednji namreč delujejo znotraj nomoloških možnosti in zaradi svojih prepričanj niso pahnjeni v moralno paralizo, ki preti skeptičnim teistom. V teh ozirih je torej eksistenčna skrb glede praktičnih posledic razvoja umetne inteligence utemeljena. Še več, glede na uspešnost raznih modelov za procesiranje naravnega jezika, ki vse bolj potrjujejo tezo o nadgradljivosti, tj. da torej za 'pravo' UI ni potreben nek kvalitativen preskok, ampak zadošča zgolj nadgradnja v parametrih in podatkih, in glede na pojmovanje UI kot transformativne, tj. takšne, ki za disruptivne učinke ne potrebuje uprimerjati človeških atributov mentalnosti, postaja vprašanje o že aktualizirani različici UI, ki ima pogubno moč, toliko bolj legitimno.

Viri in literatura

Arnold, T. in Scheutz, M. (2018). »The 'big red button' is too late: an alternative model for the ethical evaluation of AI systems«. *Ethics and Information Technology*, 20(1), str. 59–69.

Avramides, A. (2020). »Other Minds«. V Zalta, E. N. (ur.), *The Stanford Encyclopedia of Philosophy* (zima 2020). URL = https://plato.stanford.edu/archives/win2020/entries/other-minds/.

Bender, M. E., Gebru, T., Angelina McMillan-Major in Shmitchell, S. (2021). »On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?« V *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. New York: Association for Computing Machinery, str. 610–623.

Bostrom, N. (2012). »Information Hazards: A Typology of Potential Harms from Knowledge«. *Nick Bostrom's Home Page* (28. junij 2021). URL = https://www.nickbostrom.com/information-hazards.pdf.

Bostrom, N. (2013). »Exsistencial Risk Prevention as Global Policy«. *Existential Risk: threats to humanity* (28. junij 2021). URL = https://www.existential-risk.org/concept.html.

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

Bostrom, N. (2019). »The Vulnerable World Hypothesis«. *Nick Bostrom's Home Page* (28. junij 2021). URL = https://www.nickbostrom.com/papers/vulnerable.pdf.

Bringsjord, S., Bringsjord, A. in Bello, A. (2012). »Belief in the Singularity is Fideistic«.V Eden, A., Moor, J., Soraker, J. in Steinhardt, E. (urd.), *Singularity Hypotheses: A Scientific and Philosophical Assessment*. Dordrecht: Springer, str. 395–412.

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

Chirimuuta, M. (2020). »The Reflex Machine and the Cybernetic Brain: The Critique of Abstraction and its Application to Computationalism«. *Perspectives on Science*, 28(3): str. 421–457.

Danaher, J. (2015). »Why AI Doomsayers are Like Sceptical Theists and Why it Matters«. *Minds & Machines*, 25, str. 231–246.

Doctorow, C. in Stross, C. (2012). *The Rapture of the Nerds*. New York: Tor Books.

Dreyfus, H. L. (1972). *What Computers Can't Do*. New York: MIT Press.

Hadfield-Menell, D., Dragan, A., Abbeel, P. in Russell, S. (2016). »The off-switch game«. *IJCAI'17: Proceedings of the 26th International Joint Conference on Artificial Intelligence*. URL = https://arXiv.org/abs/1611.08219v3.

Hofstadter, D. R. in Dennett, D. C. (1990). *Oko duha: fantazije in refleksije o jazu in duši*. Ljubljana: Mladinska knjiga.

Karnofsky, H. (2016). »Some Background on Our Views Regarding Advanced Artificial Intelligence«. *Open Philanthrophy* (28. junij 2021). URL = https://www.openphilanthropy.org/blog/some-background-our-views-regarding-advanced-artificial-intelligence#Sec1.

Khatchadourian, R. (2016). »The Doomsday Invention: Will artificial intelligence bring us utopia or destruction«. *The New Yorker* (28. junij 2021). URL = https://www.newyorker.com/magazine/2015/11/23/doomsday-invention-artificial-intelligence-nick-bostrom.

Markič, O. (2021). »Prvi in drugi val umetne inteligence«. V Malec, M. in Markič O. (urd.), *Misli svetlobe in senc: razprave o filozofskem delu Marka Uršiča*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani, str. 201–215.

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

Orseau, L. in Armstrong, S. (2016). »Safely interruptible agents«. Pridobljeno na https://ora.ox.ac.uk/objects/uuid:17c0e095-4e13-47fc-bace-64ec46134a3f., dne 28. 6. 2021.

Radford, A., Wu, J., Armodei, D., Clark, J., Brundage, M. in Sutskever, I. (2019). »Better Language Models and Their Implications«. *OpenAI* (25. junij 2021). URL = https://openai.com/blog/better-language-models/.

Riedl, M. (2016). »Big red button«. *GitHub* (27. junij 2021). URL = https://markriedl.github.io/big-red-button/.

Searle, J. (1990). »Duhovi, možgani in programi«. V Hofstadter, D. R. in Dennett, D. (urd.), *Oko duha: fantazije in refleksije o jazu in duši*. Ljubljana: Mladinska knjiga, str. 361–379.

Schank, R. C. in Robert, P.A. (1977). *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Lawrence Erlbaum Press.

Turing, A. (1990). »Stroji, ki računajo, in inteligenca«. V Hofstadter, D. R. in Dennett, D. (urd.), *Oko duha: fantazije in refleksije o jazu in duši*. Ljubljana: Mladinska knjiga, str. 61–74.

Wilke, C. W., J., O., C. et al. (2001). »Evolution of digital organisms at high mutation rates leads to survival of the flattest«. *Nature*, 412, str. 331–333.

Whitehead, A. N. (1938). *Science and the Modern World*. Harmondsworth: Penguin.