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

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

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

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

Univerzitetna založba Univerze v Mariboru

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

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

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

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

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

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

1 Uvod

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

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

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

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

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

2 Kaj je umetna inteligenca?

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

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

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

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

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

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

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

2.1 Antropocentrizem v razpravi o umetni inteligenci

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

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

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

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

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

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

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

3 Kaj je zavest in kaj inteligenca?

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

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

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

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

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

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

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

3.1 Inteligentna UI brez zavesti ali zavestna in inteligentna UI?

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

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

3.2 Modificiran Turingov test

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4 Fizikalizem in zavestna UI

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

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

4.1 Redukcionizem in zavestna UI

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

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

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

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

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

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

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

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

4.2 Funkcionalizem in zavestna UI

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

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

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

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

5 Zaključek

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

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

Viri in literatura

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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