Identitet i umjetna inteligencija

Šimić, Domagoj

Master's thesis / Diplomski rad

2022

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **Josip Juraj Strossmayer University of Osijek, Faculty of Humanities and Social Sciences / Sveučilište Josipa Jurja Strossmayera u Osijeku, Filozofski fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:142:730217

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-07-16**

FILOZOFSKI FAKULTET
SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU

Repository / Repozitorij:

FFOS-repository - Repository of the Faculty of Humanities and Social Sciences Osijek

zir.nsk.hr

SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU

dabar
DIGITALNI AKADEMSKI ARHIVI I REPOZITORIJI

Sveučilište J.J. Strossmayera u Osijeku
Filozofski fakultet Osijek
Diplomski studij Filozofije i Engleskog jezika i književnosti

Domagoj Šimić

Identitet i umjetna inteligencija

Diplomski rad

Mentor: prof. dr. sc. Željko Senković

Osijek, 2022.

Sveučilište J.J. Strossmayera u Osijeku
Filozofski fakultet Osijek
Odsjek za filozofiju
Diplomski studij Filozofije i Engleskog jezika i književnosti

Domagoj Šimić

Identitet i umjetna inteligencija

Diplomski rad
Humanističke znanosti, filozofija, filozofija znanosti

Mentor: prof. dr. sc. Željko Senković

Osijek, 2022.

IZJAVA

Izjavljujem s punom materijalnom i moralnom odgovornošću da sam ovaj rad samostalno napisao te da u njemu nema kopiranih ili prepisanih dijelova teksta tuđih radova, a da nisu označeni kao citati s navođenjem izvora odakle su preneseni.

Svojim vlastoručnim potpisom potvrđujem da sam suglasan da Filozofski fakultet u Osijeku trajno pohrani i javno objavi ovaj moj rad u internetskoj bazi završnih i diplomskih radova knjižnice Filozofskog fakulteta u Osijeku, knjižnice Sveučilišta Josipa Jurja Strossmayera u Osijeku i Nacionalne i sveučilišne knjižnice u Zagrebu.

U Osijeku, 17.05.2022.

Domagoj Šimić, 0122221573

Zahvala

Zahvaljujem mentoru prof. dr. sc. Željku Senkoviću na stručnoj pomoći, podršci i savjetovanju tijekom pisanja ovoga diplomskog rada. Zahvalu ujedno upućujem obitelji i kolegama koji su pružali potporu tijekom studiranja.

Sažetak

Umjetna je inteligencija sustav koji prikazuje inteligentno ponašanje i prisutan je u svakodnevnome ljudskom životu, no u različitim oblicima – autonomna vozila, igre itd. U svome je trenutnome stadiju razvoja umjetna inteligencija ograničena na jedan uski skup zadataka, no pojedini, poput D. C. Dennetta, vjeruju kako bi umjetna inteligencija eventualno mogla nadići svoja trenutna ograničenja i postati nešto više. U diplomskome se radu istražuje koje bi uvjete umjetna inteligencija morala ispuniti kako bi posjedovala osobni identitet, tj. postala osobom. Na početku se rada prikazuje Turingov test i pitanje koje iz njega proizlazi o stvaranju razumnoga umjetnog entiteta. Na Turingov se test potom nadovezuju primjeri kineske sobe Johna Searlea i korejske sobe Davida Colea. Zatim se iznosi Coleova teorija o virtualnim osobama i kako je ta teorija vezana za funkcionalističko gledište koje se potom analizira. Nakon toga slijedi prikaz nužnih uvjeta za bivanje osobom koje iznosi Daniel C. Dennett u svome radu »Conditions of Personhood«. Nakon prikaza svih uvjeta, analizira se nužni uvjet intencionalnosti. Definiraju se pojmovi povezani s intencionalnom teorijom i nakon toga slijedi prikaz različitih vrsta intencionalnih sustava i koje bi vrste intencionalnoga sustava morala biti umjetna inteligencija da bi je se proglasilo umnom, a time i sposobnom za posjedovanje statusa osobe. Naposljetku, slijedi razmatranje pitanja o mogućnosti postojanja svjesne umjetne inteligencije temeljene na proučavanjima Daniela C. Dennetta i Maxa Tegmarka te razmatranje pitanja bi li uopće trebalo dopustiti stvaranje svjesne umjetne inteligencije.

Ključne riječi: umjetna inteligencija, identitet, osoba, intencionalnost, intencionalni sustavi, Turingov test, virtualna osoba, Daniel C. Dennett

Sadržaj

1. Uvod

Konferencija u Darmouthu održana 1956. godine, na kojoj je John McCarthy predstavio novotvoreni pojam »umjetna inteligencija« (eng. *artificial intelligence*), smatra se početkom prvoga razvojnog perioda umjetne inteligencije.[1] Međutim, razmišljanje je o strojevima nalik onima koje svrstavamo pod McCarthyjev pojam mnogo starije pa se tako na primjer u Homerovim tekstovima mogu pronaći spominjanja o autonomnim mehaničkim asistentima u obliku tripoda koji dočekuju bogove.[2] Konferenciji u Dartmouthu prethodi i tekst Alana Turinga zvan »Computing Machinery and Intelligence«, objavljen 1950. godine u časopisu *Mind*, u kojemu Turing razmatra pitanje mogu li strojevi misliti.[3] Iste je godine objavljena zbirka kratkih priča zvana *Ja, robot* pisca Isaaca Asimova u kojoj je prikazan svijet koegzistencije ljudi i robota, tj. bića koja posjeduju određenu vrstu umjetne inteligencije.[4] Nadalje, Vannevar Bush, pet godina prije Turingova teksta, u svojemu eseju »As We May Think« predložio je sustav koji će proširiti ljudsko znanje i razumijevanje.[5] U eseju, Vannevar Bush piše o mogućnostima konstruiranja stroja koji manipulira premisama u skladu s pravilima logike na brz i jednostavan način, a koji od nas zahtijeva samo unos premisa i okretanje ručice za dobivanje rezultata.[6]

---

[1] Blagoj Delipetrev, Chrisa Tsinaraki, Uroš Kostić, *Historical Evolution of Artificial Intelligence* (Luxembourg: Publication Office of the European Union, 2020), dostupno na: https://publications.jrc.ec.europa.eu/repository/bitstream/JRC120469/jrc120469_historical_evolution_of_ai-v1.1.pdf (pristupljeno 14.02.2022), p. 7: »The first 'AI period' began with the Dartmouth conference in 1956, where AI got its name and mission. McCarthy coined the term 'artificial intelligence,' which became the name of the scientific field.«

[2] Bruce G. Buchanan, »A (Very) Brief History of Artificial Intelligence«, *AI Magazine*, Volume 26, Number 4 (2005), dostupno na: https://doi.org/10.1609/aimag.v26i4.1848 (pristupljeno 14.02.2022), pp. 53a–60c, na p. 53a: »Ever since Homer wrote of mechanical 'tripods' waiting on the gods at dinner, imagined mechanical assistants have been a part of our culture.«

[3] Delipetrev, Tsinaraki, Kostić, *Historical Evolution of Artificial Intelligence*, p. 7: »In 1950, Alan Turing published the milestone paper 'Computing machinery and intelligence', considering the fundamental question 'Can machines think?'«

[4] Patricia Bauer, »I, Robot«, u: *Encyclopedia Britannica*, dostupno na: https://www.britannica.com/topic/I-Robot#ref341291 (pristupljeno 14.02.2022), »I, Robot, a collection of nine short stories by science-fiction writer Isaac Asimov that imagines the development of 'positronic' (humanlike, with a form of artificial intelligence) robots […] The stories originally appeared in science-fiction magazines between 1940 and 1950, the year that they were first published together in book form.«

[5] Chris Smith, »Introduction«, u: *The History of Artificial Intelligence* (University of Washington, 2006), dostupno na: https://courses.cs.washington.edu/courses/csep590/06au/projects/history-ai.pdf?msclkid=e9bf6e5da95711ecbc4cf992730c754b (pristupljeno 14.02.2022), p. 4, na p. 4: »In Vannevar Bush's seminal work *As We May Think* he proposed a system which amplifies people's own knowledge and understanding.«

[6] Vannevar Bush, »As We May Think«, u: *The Atlantic*, dostupno na: https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/ (pristupljeno 14.02. 2022.), »It is readily possible to construct a machine which will manipulate premises in accordance with formal logic […] Put a set of premises into such a device and turn the crank, and it will readily pass out conclusion after conclusion, all in accordance with logical law […]«

Osim na područjima računarske znanosti i književnosti, razmišljanja o *inteligentnim* strojevima pronalazimo i na području filozofije. Tako se na primjer kod Renéa Descartesa u djelu *Meditacije o prvoj filozofiji* nalazi latinska riječ »automata«[7]. No, Bruce G. Buchanan ističe da je Descartes ideju »strojnog čovjeka« koristio kao metaforu u svojim istraživanjima, a nije ga proučavao kao mogući entitet.[8] Naime, ovakvi su entiteti filozofima često služili kao pomoćno sredstvo pri određivanju toga što znači biti čovjekom.[9] No, nešto je drukčiji stav o strojevima imao Gottfried Wilhelm Leibniz koji je svojim radom značajno doprinio računarskoj znanosti o čemu piše Laurent Bloch u tekstu »Informatics in the light of some Lebniz's works«.[10] Naime, Leibniz je uviđao mogućnosti mehaničkih uređaja i smatrao je da bi korištenjem pravila logike mogli riješiti pojedine sporove.[11] Prema navedenome, vidljivo je da su u filozofiji razmišljanja o *inteligentnim* strojevima prisutna već nekoliko stotina godina.

Među ostalim filozofskim razmišljanjima, izdvojio bih misaonu igru francuskog empirista Étiennea Bonnota de Condillaca u kojoj se pita dokle bi trebali sipati grumenčiće znanja u kip kako bi se on počeo doimati inteligentnim.[12] Navedeno je Condillacovo pitanje o granici znanja i neznanja aktualno i u trenutnome razdoblju naše civilizacije, no objekt pitanja više nije kip, već stroj. Šezdeset godina nakon uvođenja pojma »umjetna inteligencija«, odvija se razvoj tehnologije poznate kao *umjetna sužena inteligencija* (eng. *artificial narrow intelligence*). Navedena umjetna sužena inteligencija posjeduje sposobnost obavljanja jednoga uskog skupa zadataka, a njezina primjena seže od područja igara poput šaha do autonomnih vozila.[13] Potencijalni raspon primjene umjetne inteligencije obuhvaća sva područja ljudskoga života i vrlo je izgledno da će biti jedan od temeljnih čimbenika u dostizanju civilizacije prvoga stupnja prema Kardaševoj skali[14]. Međutim, iako umjetna inteligencija posjeduje neopisiv

---

[7] René Descartes, *Razmišljanja o prvoj filozofiji*, *u kojima se dokazuje Božja opstojnost i razlika između ljudske duše i tijela*, s latinskog preveo Tomislav Ladan (Zagreb: Demetra. Filozofska biblioteka Dimitrija Savića, 1993), p. 61.

[8] Buchanan, »A (Very) Brief History of Artificial Intelligence«, p. 53a: René Descartes, for example, seems to have been more interested in 'mechanical man' as a metaphor than as a possibility.«

[9] Ibid., »Philosophers have floated the possibility of intelligent machines as a literary device to help us define what it means to be human.«

[10] Vidi: Laurent Bloch, »Informatics in the light of some Leibniz's works«, (2016) dostupno na: https://www.researchgate.net/publication/311707999_Informatics_in_the_light_of_some_Leibniz's_works/link/5856641f08aeff086bfbb3d2/download

[11] Buchanan, »A (Very) Brief History of Artificial Intelligence«, p. 53a: »Gottfried Wilhelm Leibniz, on the other hand, seemed to see the possibility of mechanical reasoning devices using rules of logic to settle disputes.«

[12] Ibid., p. 53b: »Etienne Bonnot, Abbé de Condillac used the metaphor of a statue into whose head we poured nuggets of knowledge, asking at what point it would know enough to appear to be intelligent.«

[13] Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence* (New York: Alfred A. Knopf, 2017), dostupno na: https://www.cag.edu.tr/d/l/68e79b19-4dd7-43b0-a578-cdb4308b1881 (pristupljeno 14.02.2022), p. 55: »Ability to accomplish a narrow set of goals, e.g., play chess or drive a car.«

[14] Za Kardaševu skalu vidi: Guillermo A. Lemarchand, »Detectability of Extraterrestrial Technological Activities«, u: *SETIQuest*, Volume 1, Number 1, dostupno na:

tehnološki značaj za napredak naše civilizacije, njezin razvoj ima moralne, pravne i gospodarske implikacije zbog čega je nužno preispitati njezinu primjenu na ljudsku svakodnevicu. Jedna bi od uočljivijih posljedica umjetne inteligencije bila manjkava potreba za radnom snagom što bi, vjerujem, direktno utjecalo na raspodjelu bogatstva. No, osim moralnih, pravnih i gospodarskih posljedica, razvoj bi umjetne inteligencije mogao imati i ontološke posljedice. Pojedini autori postavljaju pitanje hoće li umjetna inteligencija ikada moći posjedovati svijest i osobni identitet. Afirmativni bi odgovor na to pitanje uveo novu dinamiku u načinu odnošenja dvaju entiteta, čovjeka i stroja, tj. došlo bi do izmjene moralnih i pravnih statusa entiteta, a naša bi se ontološka poimanja o biću drastično izmijenila.

U ovome ću se radu usredotočiti na pitanje osobnog identiteta, tj. može li se umjetna inteligencija smatrati osobom. Pitanje ću osobnoga identiteta razmotriti tako što ću prikazati i istražiti nužne preduvjete koje bi umjetna inteligencija morala ispuniti za posjedovanje statusa osobe.[15] Navedenim ću razmatranjem preduvjeta ujedno vršiti usporedbe između čovjeka i stroja. Međutim, potreban je oprez prilikom stvaranja analogija između čovjeka i umjetne inteligencije zbog antropocentričnoga gledišta, kojemu nikada ne možemo u potpunosti umaknuti, različitosti entiteta i mogućnosti zapadanja u pretjerano maštovite i nestvarne tvrdnje. Nadalje, prikaz ću umjetne inteligencije vršiti pomoću intencionalnoga stajališta, a u radu ću također zauzeti naturalističko stajalište koje »pojmove, stavove, norme ili postupke te njihove predmete objašnjava prirodnim svojstvima i podrijetlom, inzistirajući na relevantnosti rezultata istraživanja u prirodnim (fizici, kemiji i biologiji) i nekim društvenim znanostima (psihologiji ili lingvistici).«[16]

Istraživanje ću o osobnom identitetu započeti prikazom Turingova testa i entiteta kojeg David Cole naziva *virtualnom osobom*. Potom ću prikazati nužne preduvjete za posjedovanje osobnoga identiteta i pomoću teorije intencionalnih sustava istražiti kako umjetna inteligencija zadovoljava kriterij umnosti. Zatim ću prikazati stavove dvaju teoretičara, Daniela C. Dennetta i Maxa Tegmarka, o mogućnosti umjetne inteligencije da posjeduje svijest. Naposljetku, iznijet ću stavove o tome trebamo li uopće stvoriti svjesnu umjetnu inteligenciju.

https://www.researchgate.net/publication/314151939_Dectectability_of_extraterrestrial_technological_activities (pristupljeno 14.02.2022.), pp. 3–13.

[15] Ispunjavanje nužnih preduvjeta za bivanje osobom umjetnoj inteligenciji ne osigurava ostvarivanje statusa osobe, ali se zadovoljavanjem tih minimalnih uvjeta omogućava preispitivanje i daljnja rasprava.

[16] Zvonimir Čuljak, natuknica »naturalizam«, u: *Filozofski leksikon*, glavni urednik: Stipe Kutleša (Zagreb: Leksikografski zavod Miroslav Krleža, 2012), pp. 797a–797b, na p. 797a.

2. Turingov test

Na početku svojega teksta »Computing Machinery and Intelligence«, Alan Turing predlaže razmatranje pitanja može li stroj misliti.[17] U svojemu razmatranju toga pitanja Turing uviđa da ono sadrži niz problema radi mnogostrukosti odgovora koji se svode na razlike u mišljenju te stoga smatra da ga se treba zamijeniti s jednom od verzija igre imitacije.[18] Stoga, Turing prvotno pitanje odbacuje i zamjenjuje ga sljedećim: Što će se zbiti kada stroj preuzme ulogu entiteta A u igri imitacije?[19]

No, što je uopće tzv. igra imitacije u svojoj originalnoj formi? Prvotna verzija igre imitacije koju Turing navodi u tekstu podrazumijeva tri entiteta (četiri ako u igru uključimo ulogu posrednika): muškarca (A), ženu (B) i ispitivača (C) koji je odvojen od entiteta A i B.[20] Cilj je igre da ispitivač, na temelju proizvoljnih pitanja, ispravno ustanovi tko je od navedenih entiteta muškarac, a tko žena.[21] Entitet A treba zavarati ispitivača, a entitet B treba ispitivaču pomoći.[22] Naravno, ispitivaču nije dozvoljeno vidjeti niti čuti davatelje odgovora te se stoga poruke prenose tiskanim putem ili putem posrednika.[23] No, adaptirana verzija te igre, ono što zovemo Turingov test, zamjenjuje jednog od dva sudionika strojem i cilj ispitivača više nije identificirati spol sudionika, već identificirati tko je od dvaju entiteta stroj.[24] U tome testu fizička priroda entiteta ne predstavlja ikakav značaj te se u njemu isključivo proučavaju izlazne informacije ispitanika.[25] Razlog zbog kojeg se isključivo proučavaju informacije jest taj što Turing tim testom želi napraviti jasnu razdiobu između fizičkih i intelektualnih mogućnosti

[17] Alan M. Turing, »Computing Machinery and Intelligence«, u: *Mind*, New series, Volume 59, Number 236 (1950), dostupno na: https://phil415.pbworks.com/f/TuringComputing.pdf (pristupljeno 14.02.2022), pp. 433–460, na p. 433: »I propose to consider the question, 'Can machines think?'«

[18] Brian McGuire, »The Turing Test«, u: *The History of Artificial Intelligence* (University of Washington, 2006), dostupno na: https://courses.cs.washington.edu/courses/csep590/06au/projects/history-ai.pdf, pp. 5–6, na p. 6: »In the paper he wanted to replace the question, 'Can machines think?' (which can have many possible answers and come down to a difference of opinion) with a version of the 'Imitation Game.'«

[19] Turing, »Computing Machinery and Intelligence«, p. 434: »We now ask the question, 'What will happen when a machine takes part of A in this game?'«

[20] Ibid., p. 433: »It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two.«

[21] Ibid., »The object of the game of the interrogator is to determine which of the other two is the man and which is the woman.«

[22] Ibid., pp. 433–434: »It is A's object in the game to try and cause C to make the wrong identification […] The object of the game for the third player (B) is to help the interrogator.«

[23] Ibid., p. 434: »The ideal arrangement is to have a teleprinter communicating between the two rooms. Alternatively the question and answers can be repeated by an intermediary.«

[24] McGuire, »The Turing Test«, p. 6: »For the Turing Test, one of those two participants would be replaced by a machine and the goal of the interrogator would not be to identify the gender of the participants, but which is human and which is a machine.«

[25] Ibid., »Second, physical nature isn't important – the goal is to not be able to tell the difference between man and machine when comparing the output of the machine and the true human.«

čovjeka.[26] Također, ograničavanjem mogućnosti interakcije i relevantnoga konteksta, sprječava se nastanak predrasuda i preduvjerenja, tj. na taj se način pokušavaju osigurati jednaki uvjeti za oba entiteta.[27]

Uspješan prolazak stroja na tome testu predstavlja konačni cilj istraživanja umjetne inteligencije, tj. njegovim bi se prolaskom riješilo pitanje o mogućnosti stvaranja stroja koji može zadovoljavajuće oponašati čovjeka do te mjere u kojoj sumnjičavi sudac ne bi mogao raspoznati razliku između njih.[28] Navedenim uviđamo bit Turingova testa koju iznosi Marvin Minsky, a glasi da je taj test zapravo sud o tome koliko dobro strojevi postupaju poput ljudi.[29] Minskyjevom se izjavom potvrđuje da cilj i bit testa u sebi sadrže antropocentrizam, no time ujedno dolazimo do pitanja kojeg postavljaju pojedini autori, a glasi: Ako bi računalo moglo oponašati razumno ponašanje čovjeka, ne bi li to značilo i da je samo računalo razumni entitet?[30]

2.1. Searleova kineska soba

Brojni istraživači iz područja kognitivne znanosti vjeruju da bi računala u budućnosti potencijalno mogla imati prave mentalne sposobnosti.[31] Trenutno je moguće napisati program koji nudi odgovore u obliku prirodnoga jezika na pitanja iz određenih domena i pojedini vjeruju da se putem takvoga dovitljivog programiranja stvara zbiljsko razumijevanje, no drugi ipak vjeruju da zbiljsko razumijevanje još uvijek nije postignuto, ali smatraju da bi moglo biti u budućnosti kada se poboljšaju tehnike programiranja, povećaju baze podataka i stvore brži

[26] Turing, »Computing Machinery and Intelligence«, p. 434: »The new problem has the advantage of drawing a fairly sharp line between the physical and the intellectual capacities of a man.«
[27] Jenny Erikson Lundström i Stefan Karlsson, »Approaching Artificial Intelligence for Games – the Turing Test revisited«, u: *TripleC*, Volume 4, Number 2 (2006), dostupno na: https://www.triple-c.at/index.php/tripleC/article/download/32/32 (pristupljeno 14.02.2022), pp. 167–171, na p. 169: »By choosing a game setting, Turing restrained the interaction possibilities and the relevant context to prevent prejudices and preconceptions, and thus tried to provide equal conditions for the man and the machine.«
[28] McGuire, »The Turing Test«, p. 5: »The Turing test is a central, long term goal for AI research – will we ever be able to build a computer that can sufficiently imitate a human to the point where a suspicious judge cannot tell the difference between human and machine?«
[29] Lundström i Karlsson, »Approaching Artificial Intelligence for Games – the Turing Test revisited«, p. 169: »Marvin Minsky […] expressed that: 'The very essence of the Turing test is our judgment of how well machines act like humans.'«
[30] McGuire, »The Turing Test«, p. 6: «If a computer could imitate the sentient behavior of a human would that not imply that the computer itself was sentient?«
[31] David Cole, »Artificial Intelligence and Personal Identity«, u: *Synthese*, Volume 88, Number 3, dostupno na: https://history.as.uky.edu/sites/default/files/Artificial%20Intelligence%20and%20Personal%20Identity%20-%20David%20Cole.pdf (pristupljeno 14.02. 2022), pp. 399–417, p. 399: »Many workers in cognitive science believe that computers can potentially have genuine mental abilities.«

strojevi.[32] Međutim, taj optimizam istraživača ne dijeli John Searle koji sumnja u mogućnosti računala.[33] U tekstu »Minds, Brains and Programs« Searle izražava protivljenje tvrdnji da primjereno programirana računala doslovce imaju kognitivna stanja.[34] Searleov je stav da računalo, u najboljem slučaju, može simulirati, ali ne i posjedovati inteligenciju.[35] Naime, bez obzira na to koliko je program »pametan« i složen, Searle tvrdi da čovjek može činiti isto ono što računalo čini, a to je slijediti upute za generiranje nizova simbola kao odgovor na dolazne nizove.[36] Primjer za Searleov argument o računalnom nedostatku inteligencije i razumijevanja jest tzv. *kineska soba*.

U primjeru kineske sobe trebamo pretpostaviti da osoba (Searle, kako piše u izvornom tekstu argumenta) koja ne zna kineski jezik sjedi u prostoriji s određenim uputama, tj. s »programom«, napisanima na engleskome jeziku u kojima detaljno piše kako manipulirati kineskim slovima te tako stvoriti nizove odgovora na temelju primljenih podataka.[37] Ono što također trebamo pretpostaviti u navedenom primjeru jest da su upute takve da dopuštaju uspješno prolaženje Turingova testa.[38] Dakle, budući da u uputama piše što osobi valja činiti na temelju formalnih sintaktičkih značajki nizova, bez spominjanja ili otkrivanja značenja, osoba može proizvesti kineske rečenice bez ikakvoga razumijevanja što one znače, pa čak i bez znanja da se pred njom nalaze kineske rečenice.[39] Prema Searleovu mišljenju, taj primjer sugerira da program nema veze s razumijevanjem sve dok je definiran u terminima računskih operacija na temelju čisto formalno definiranih elemenata.[40] Drukčije rečeno, radeći bez

[32] Ibid., p. 400: »It is possible to write computer programs that produce responses in natural language to questions about some subject domain. Some believe that through such clever programming actual understanding is produced. Others believe that genuine understanding is not *yet* achieved, but it may be in the future with improved programming techniques, larger databases and faster machines.«
[33] Ibid., p. 399: »John Searle has been an prominent critic of this optimism about the abilities of computers.«
[34] Graham Oppy i David Dowe, »The Turing Test«, u*: The Stanford Encyclopedia of Philosophy*, dostupno na: https://plato.stanford.edu/entries/turing-test/ (pristupljeno 14.02.2022), »In Minds, Brains and Programs and elsewhere, John Searle argues against the claim that 'appropriately programmed computers literally have cognitive states'.«
[35] Cole, »Artificial Intelligence and Personal Identity«, p. 399: »Searle argues that computers can at best simulate, but not possess, intelligence.«
[36] Ibid., p. 400: »But, as Searle argues, consider that no matter how clever and complex the program, a human could do exactly what the computer does: follow instructions for generating strings of symbols in response to incoming strings.«
[37] Ibid., »Suppose, for example, a person (Searle, in the original statement of the argument) who does not know Chinese sits in a room with instructions written *in English* (a 'program') that tell one in detail how to manipulate *Chinese* symbols, producing strings in response to the strings given to one.«
[38] Ibid., »We are to suppose that the instruction are such that they permit successful passage of this variation on a Turing Test […]«
[39] Ibid., »Since the instructions tell one what to do entirely on the basis of formal syntactic features of the strings, without ever mentioning (or revealing) meaning, one can generate Chinese sentences without any understanding of what they mean – indeed without even knowing that they are Chinese sentence.«
[40] John. R. Searle, »Minds, brains, and programs«, u: *Behavioral and Brain Sciences 3* (Cambridge University Press, 1980), dostupno na: https://www.law.upenn.edu/live/files/3413-searle-j-minds-brains-and-programs-1980pdf (pristupljeno 14.02.2022), pp. 417a–424b, na p. 418b: »As long as the program is defined in terms of

odstupanja isto ono što radi i računalo, čovjek ne bi posjedovao mogućnost razumijevanja kineskoga jezika te se na temelju toga može zaključiti da ni samo računalo ne bi posjedovalo razumijevanje.[41] Prema tome, ako je Searle u pravu i računalo eventualno uspije proći Turingov test, ostaje činjenica da nijedno računalo neće nikada razumjeti prirodni jezik ili imati izvorne propozicijske stavove poput vjerovanja.[42] Prihvati li se navedeno, upitno postaje može li stroj steći status osobe.

2.2. Ideja o *virtualnoj osobi*

Analizirajući Searleov primjer kineske sobe, David Cole tvrdi da iz činjenice da netko ne razumije kineski jezik ne proizlazi zaključak da nitko ne razumije kineski jezik.[43] Kako bih prikazao o čemu se točno radi, iznijet ću Coleov primjer tzv. *korejske sobe*. Korejska je soba varijacija Searleova misaonog eksperimenta koja od čitatelja zahtijeva da zamisli Searlea kako obitava u sobi i slijedi uputstva o tome kako treba postupati s nizovima nepoznatih znakova koje dobiva.[44] Na temelju uputa koje prati, Searle stvara odgovore na pitanja postavljena na kineskome jeziku te stoga osobe izvana vjeruju da netko, tko je unutar prostorije, razumije kineski jezik .[45] No, za razliku od prije, oni izvan prostorije ujedno vjeruju da netko tko obitava u prostoriji govori i korejski jezik jer na pitanja postavljena na korejskome jeziku primaju odgovore na tom istom jeziku.[46] Nadalje, osobe su izvana ujedno u uvjerenju da unutar prostorije obitava više osoba, tj. smatraju da se unutar prostorije nalaze najmanje dvije osobe od kojih jedna razumije kineski jezik, ali ne i korejski (označit ćemo ju kraticom Pc), i druga koja razumije korejski jezik, ali ne i kineski (označit ćemo ju kraticom Pk).[47]

---

computational operations on purely formally defined elements, what the example suggests is that these by themselves have no interesting connection with understanding.«

[41] Cole, »Artificial Intelligence and Personal Identity«, pp. 400–401: »That is, doing exactly what a computer does would not give one the ability to understand chinese. Therefore, the computer does not understand chinese either.«

[42] Ibid., p. 399: »If Searle is correct, even thought a computer might eventually pass the Turing Test, no computer will every actually understand natural language or have genuine propositional attitudes, such as beliefs.«

[43] Ibid., p. 401: »Clearly from the fact that *someone* does not understand Chinese it does not follow that *no one* understands Chinese.«

[44] Ibid., p. 402: »[…] let us consider a variation on Searle's thought experiment. Once again we are to imagine Searle in a room, following instructions for dealing with strings of alien characters slipped under the door.«

[45] Ibid., »He generates reams of marks in return, as instructed by the manuals . As before, many of those outside believe that someone in the room understands Chinese, for they are receiving replies to questions submitted in Chinese.«

[46] Ibid., »But unlike before, those outside *also* believe someone in the room speaks Korean, for they are receiving replies in Korean to questions submitted in Korean.«

[47] Ibid., »Furthermore, they also believe that the room is more crowded than before: there appear to be at least two people in the room, one who understands Chinese (call him Pc) but not Korean, and another who understands Korean but not Chinese (call him Pk).«

Pretpostavimo dalje da su sposobnosti i dispozicije utjelovljene u Pc-ovoj bazi podataka izvedene od jedne starije Kineskinje, dok su one u Pk-ovoj bazi podataka izvedene od mladoga Korejca koji je bio žrtva u sudaru kamiona i bicikla.[48] Odgovori na pitanja postavljena na kineskome jeziku otkrivaju jedan pametan, duhovit i šaljiv um koji je dobro upućen u stvari vezane za Kinu, no koji je ujedno poprilično neupućen u korejski jezik i događaje u Koreji.[49] Pc također izvještava da ima sedamdeset i dvije godine te izražava mudra i zanimljiva zapažanja o tome kako je bilo proteklih pola stoljeća biti ženom u Kini.[50] Nasuprot »njoj«, odgovori na pitanja postavljena na korejskome jeziku otkrivaju jednoga poprilično dosadnog mladića. Pk je mizogin, pokazuje mržnju prema Kini, no pritom je neobaviješten o tamošnjim događajima i informira ispitivače da radi u tvornici televizora te daje točne opise procesa njihova sastavljanja.[51] Za taj je primjer potrebno dalje pretpostaviti da bihevioralni dokazi potvrđuju, koliko god mogu, da se u navedenome slučaju unutar prostorije nalaze dvije različite osobe.[52] Koliko se god sugovornici izvana trudili, oni ne mogu pronaći nikakvu naznaku da postoji samo jedna osoba koja se pretvara da je dvije osobe.[53] Prema tome, u prostoriji ne postoji jedna osoba koja istovremeno razumije i kineski i korejski jezik, a Searle koji se nalazi u prostoriji ne razumije niti jedan.[54] Također, upute su za generiranje kineskih i korejskih odgovora potpuno različite i ne postoji razmjena informacija između dviju konzultiranih baza podataka, no sam Searle nema saznanja o tome.[55] Dakle, u korejskoj je sobi Searle taj koji prati upute i proizvodi odgovore na pitanja usmjerena za Pc i Pk, no Searle ne razumije ono što proizvodi.

[48] Ibid., »In fact let us suppose that the response abilities and dispositions embodied in the Pc database were derived from an elderly Chinese woman, whereas those in the Pk database were from a young Korean male who was a victim of a truck-bicycle collision.«
[49] Ibid., »The answers to the questions in Chinese appear to reveal a clever, witty, jocular mind, knowledgeable about things in China, but quite ignorant of both the Korean language and events in Korea.«
[50] Ibid., »Pc reports being seventy-two years old and is both wise and full of interesting observations on what it has been like to be a woman in China for the tumultuous past half-century.«
[51] Ibid., pp. 402–403: »By contrast, the replies to Korean questions reveal quite a dull young man. Pk is misogynous. Pk, has a vitriolic hatred of China, but is largely ignorant of events there. […] Pk reports that he works in a television factory and gives accurate descriptions of television assembly.«
[52] Ibid., p. 403: »Thus, suppose that the behavioral evidence is a clear as can be in such a case that there are two *distinct* individuals in the room.«
[53] Ibid., »Try as they might, interlocutors outside the room can find no hint that there might be but a single person pretending to be two.«
[54] Ibid., »There is in fact no individual inside the room who understands both Chinese and Korean. Searle understands neither.«
[55] Ibid., »And the instructions for generating replies to Chinese input and those for dealing with Korean input are distinct, with no exchange of information between the databases consulted (although this is not know by Searle).«

Coleovoj tvrdnji da su Pc i Pk dvije različite osobe ide u prilog i to što su reprezentativne povijesti unutar njihovih baza podataka u potpunosti neovisne.[56] Prema tome, Pc ne može biti identičan sa Pk, a ujedno se ne može ni tvrditi da su Pc ili Pk identične Searleu.[57] Navedeni dokazi u Coleovu radu jasno upućuju da osoba Pc nije osoba Pk.[58] Kad bi tvrdili da su Pc i Pk identični sa Searleom, narušili bi tranzitivnost identiteta, stoga Cole zaključuje da moramo držati kako nijedno od njih dvoje nije Searle.[59] Nadalje, prilikom postavljanja pitanja Pk-u na korejskom o Kini, odgovori Pk-a na pitanja ne pokazuju poznavanje Kine, već predrasude prema Kini, a kada se slična pitanja postave Pc-u na kineskom, odgovori Pc-a pokazuju znanje i ljubav prema Kini te iz toga slijedi pitanje sviđa li se ili ne računalu Kina?[60] Cole smatra da njegova razmatranja sugeriraju da je pogrešno atributirati takva svojstva samome računalu.[61] Naime, samo računalo nije nositelj tih svojstava, ali jest, baš poput Searlea, realizator Pc-a i Pk-a ili, kako Cole iznosi, računalo je supstrat koji ostvaruje dva virtualna subjekta.[62] Termin koji Cole koristi za virtualni subjekt jest *virtualna osoba*. Navedeni je Coleov termin povezan s konceptom iz računarstva poznatim kao virtualni stroj.[63] Štoviše, pojedini su znanstvenici poput Paula Smolenskyja sagledavali svijest kao virtualni stroj.[64]

Sumirajući dosad iznesene postavke Coleova primjera, utvrđeno je sljedeće: Searle ≠ Pc, Searle ≠ Pk, Pk ≠ Pc, stroj ≠ Pc, stroj ≠ Pk. Na temelju je toga ostalo nejasno tko ili što su Pc i Pk i kako oni postoje. Cole je ustvrdio kako je računalna aktivnost ta koja uzrokuje postojanje nove osobe, no koja nije identična sa samim računalom. Dalje je utvrđeno kako ta osoba, koja isključivo postoji strojnom aktivnošću, jest entitet koji razumije jezik. Sami stroj ne posjeduje opće razumijevanje niti razumijevanje jezika. Tu osobu koja razumijeva Cole naziva virtualnom osobom i ona djelovanjem stroja ima ostvareni um koji opovrgava Searleovo negiranje mogućnosti umjetne inteligencije.

---

[56] Ibid., »And the histories of the representations in the Pc and Pk databases are completely independent, […] Thus *if* Pc and Pk are persons, they are distinct.

[57] Ibid., »Pc cannot be identical with Pk. The grounds for saying that Pc is Searle are just the same as those for holding that Pk is identical with Searle.«

[58] Ibid., p. 402: »The evidence seems quite clear that Pc is not Pk.«

[59] Ibid., p. 403: »We cannot hold that both are identical with Searle, for this would violate the transitivity of identity. Therefore, we must hold that neither is Searle.«

[60] Ibid., p. 404: »When asked in Korean about China, the replies do not demonstrate knowledge of China, only a vitriolic prejudice against China. When asked similar questions in Chinese, the replies exhibit knowledge and love of China. Does *the computer* like China or not?«

[61] Ibid., »These considerations suggest that it would be a mistake to attribute these properties to the computer itself.«

[62] Ibid., These two virtual subject are realized by a single substratum, the computer.

[63] Ibid., »The concept of a *virtual machine* is familiar in computer science.«

[64] Ibid., »And some computer scientists, such as Paul Smolensky, have viewed consciousness as a virtual machine.«

2.3. *Virtualna osoba* i Coleovo funkcionalističko stajalište

U svojemu je radu David Cole zauzeo funkcionalističko gledište i namjera mu je bila prikazati utjecaj toga gledišta na shvaćanje prirode osoba i njegovu relevantnost za umjetnu inteligenciju i Searleov argument kineske sobe.[65] Cole iznosi tvrdnju da dijakronijska perspektiva funkcionalista o osobama sugerira da su osobe ili umovi znatno apstraktniji nego što to pretpostavljaju teorije koje poistovjećuju identitet osobe s tijelom (ili s kartezijanskom dušom ili individualnim *res cogitansom*).[66] S ozbirom na to da funkcionalistička teorija ne dopušta istovrsnost psiholoških i tjelesnih stanja, mogli bi na primjer postojati vanzemaljski oblici života sa psihološkim stanjima istoga tipa kao psihološka stanja ljudi, no čiji je temeljni tjelesni sustav sasvim drugačiji.[67] U sljedećem primjeru Cole tvrdi da njegova psihološka stanja iz 1989. godine koja su istovrsna sa psihološkim stanjima iz 1979. godine ne moraju biti ostvarena istim fizičkim stanjima te je moguće da je dio mozga u međuvremenu pretrpio kakvu ozljedu i da je njegovu funkciju sada preuzela kakva druga fiziološka struktura.[68] Cole optimistično navodi i to kako bi u budućnosti moglo doći do razvoja tehnologije koja bi omogućila zamjenu oštećenih dijelova mozga s kultiviranim neonatalnim tkivom koje bi preuzelo privremeno izgubljene moždane funkcije.[69] Naposljetku, tvrdi Cole, moglo bi čak postati moguće zamijeniti cijele oštećene neurone funkcionalno ekvivalentnim elektroničkim uređajima na bazi silicija.[70] No, Cole se tu propušta zapitati koliko bi toga kod čovjeka trebali supstituirati elektroničkim uređajima da on prestane biti čovjekom, tj. izvorni *homo sapiens sapiens*.

Na temelju prethodno navedenih primjera vidljiv je funkcionalistički stav da tip temeljne supstancije nije bitan za psihološka stanja.[71] Međutim, Cole upozorava da to ne znači da mogu postojati psihološka stanja bez ikakvoga temeljnog supstrata; upali bi u modalnu

[65] Ibid., p. 411: »I shall not defend functionalism here, but shall indicate how it bears on the nature of persons and how this is relevant to Artificial Intelligence and Searle's argument.«
[66] Ibid., p 412: »The functionalist diachronic perspective on persons suggests that persons or minds are more abstract than a simple identity of a person with a body (or a Cartesian soul or individual *res cogitans*) would suppose.«
[67] Ibid., p. 411: »Functionalism rejects a type-type identity between psychological states and physical states. […] There could even be alien lifeforms with psychological states that were of the same type as psychological states had by humans, but that had quite a different underlying physical system […]«
[68] Ibid., p. 411–412: »My psychological states in 1989 need not be realized by the same physical states as were my type-identical psychological states in 1979. For example, a portion of my brain may have sustained injury in the interim and its function may have been assumed by a physiologically distinct structure.«
[69] Ibid., p 412: »Or, it may become possible to replace damaged portions of my brain with cultured neonatal tissue that grows to assume functions temporarily lost.«
[70] Ibid., »Finally it might even become possible to replace entire damaged neurons by functionally equivalent silicon-based electronic device.«
[71] Ibid., »Functionalism thus takes the underlying substance type to be non-essential to the psychological states.«

zabludu kada bi zaključili da neesencijalnost tipa supstrata upućuje na neesencijalnost postojanja nekoga supstrata.[72] Searle se vjerojatno ne bi složio s navedenim i smatrao bi da je osoba identična sa svojim tijelom i da ne bi mogla postojati bez njega, ili barem ne bez mozga.[73] S druge strane, Cole ne dijeli Searleov stav i za svoju argumentaciju navodi tvrdnje Lockea koje sugeriraju da bi, u principu, mogla postojati osoba čije trenutno tijelo i mozak bivaju zamijenjeni drugim tijelom i mozgom, identični trenutnome tijelu i mozgu, s istim *uzročnim moćima*.[74] Zaključak do kojeg Cole dolazi na temelju Lockeovih stavova jest da je osoba atribut tijela te da jedno tijelo može realizirati više od jedne osobe, a da jednu osobu može realizirati više od jednoga tijela.[75]

Dakle, funkcionalizam ne zahtijeva korespondentnost jedne osobe s jednim tijelom, no takva korespondencija općenito postoji.[76] Naime, Cole tvrdi kako su uzročna svojstva psiholoških stanja zapravo svojstva sustava koji doslovce utjelovljuje ta psihološka stanja te da fiziološke karakteristike mozga u uobičajenom tijeku događaja dozvoljavaju psihološku integraciju i kontinuiranost što omogućava postojanje osobe i njezinu korespondenciju s tijelom.[77] No, Cole tvrdi da možda može biti i drugačije jer postoje slučajevi »višestruke osobnosti« kada jedan mozak utjelovljuje više osoba. Tu se može pronaći analogija s računalom koje posjeduje mogućnost višestrukoga utjelovljivanja osoba, tj. računalo možemo shvatiti kao jedno tijelo, supstrat, putem kojega se mnoštvo osoba realizira, no između kojih mora postojati određena granica razlikovanja. Ta bi se granica razlikovanja mogla odrediti pomoću ideje kauzalnosti. Funkcionalistička teorija tvrdi da su kauzalna svojstva događaja odrednice psiholoških svojstava.[78] Naime, kauzalnost je ta koja drži snopove psiholoških stanja i događaja zajedno kako bi kroz vrijeme formirali jedan um.[79]

---

[72] Ibid., »This is not to suppose that there can be psychological states without *any* underlying substratum – the inference from the non-essentiality of any *given* substratum to the non-essentiality of existence of *some* substratum or other would be a modal scope fallacy.«

[73] Ibid., p. 413: »Presumably Searle would say that I am identical with my body and could not exist without it, or at least not without the brain.«

[74] Ibid., »But Locke's view suggests that I could. Another body and brain exactly like this one, with the same 'causal powers', could, in principle, replace this one.«

[75] Ibid., »A person is an attribute of a body. A single body might realize more than one person, and a single person might be realized by more than one body.«

[76] Ibid., p. 414: »Functionalism does not require a one-to-one correspondence between persons and bodies. Contingently, there generally is such a correspondence […]«

[77] Ibid., »The causal properties of psychological states are just those of the system literally embodying them. And in the ordinary course of events, the physiological characteristics of brains permit psychological integration and continuity for the entire duration of the operating life of the brain.«

[78] Ibid., »Functionalism takes certain of the causal properties of an event to be determinants of the psychological properties.«

[79] Ibid., »Causality […] is what holds bundles of psychological states and events together to form a single mind over time.«

No što nam funkcionalistička teorija govori o strojevima, Searleovom primjeru kineske sobe i Coleovom primjeru korejske sobe? Coleov je cilj u radu bio procijeniti odnose strojeva i umova u skladu s funkcionalističkom teorijom te prikazati zašto bi zagovornik funkcionalizma trebao Searleove argumente smatrati neispravnima.[80] Coleov se argument zasniva na tvrdnji da su programi apstraktni, no ne toliko apstraktni poput osoba.[81] S obzirom na činjenicu da je osoba ta koja posjeduje razumijevanje i da se ona može realizirati različitim programima, osoba koja razumije nije identična s programom koji je realizira, što je u skladu sa standardnim funkcionalističkim prigovorom o poistovjećivanju osobe s njezinim tijelom.[82] U slučaju kineske i korejske sobe, računalo i program nisu ti koji posjeduju razumijevanje. Naime, program je u potpunosti inertan sve dok ga se ne pokrene i može ga se shvatiti, u okviru Aristotelovih termina, kao formu neke materije koja ne čini ništa bez temeljne supstancije.[83] Sam program postoji prije i nakon pokretanja, ali razumijevanje, ako postoji, postoji samo dok se program izvodi.[84] Dakle, kineski jezik u kineskoj sobi razumije jedna neimenovana kineska osoba.[85] Ta osoba nije Searle, ali ta osoba ne može postojati sve dok netko, Searle ili bilo tko drugi, ne da život kineskome umu slijedeći upute u prostoriji.[86] Također, scenarij korejske sobe pokazuje da (virtualne) osobe koje manifestira sustav mogu posjedovati kontradiktorne psihološke atribute od osoba koje ga realiziraju.[87] Na temelju svojih razmatranja, Cole piše kako je stav Searlea o tome da nijedan digitalni stroj ne može razumjeti jezik ispravan, no netočna je Searleova tvrdnja da su umjetni umovi nemogući jer, prema Coleovu mišljenju, umovi i osobe nisu isti kao strojevi koji ih ostvaruju, bili oni biološki ili elektronički.[88] Nadalje, iz činjenice da postoji jedan fizički sustav, ne slijedi ništa o broju umova koje bi taj sustav mogao realizirati te bi on stoga, ovisno o svome kauzalnome karakteru, mogao realizirati jedan

[80] Ibid., p. 408: »My task here is to develop an assessment of the relation of machines to minds which is compatible with functionalism and to show why a functionalist ought to view Searle's argument as unsound.«
[81] Ibid., p. 416, fusnota 6: »[…] my response is that programs are abstract, but not as abstract as persons.«
[82] Ibid., »Since it is a person who understands, and the same person can be realized by distinct programs, the understanding person is not identical with the program. This reasoning parallels the standard functionalist objections to identifying a person with his/her body.«
[83] Ibid., p. 405: »The program itself is entirely inert until it runs. In Aristotelian terms, a program could be but the form of some matter – without an underlying substance, it does nothing.«
[84] Ibid., p. 406: »The program exists before and after it is run, but understanding, if any, exists only while the program is running.«
[85] Ibid., p. 414: »So who or what does understand Chinese in the Chinese room? An unnamed Chinese person.«
[86] Ibid., »This person is not Searle, but this person cannot exist unless someone – Searle or any competent other – brings to life the Chinese mind by following the instructions in the room.«
[87] Ibid., p. 413: »[…] the Kornese room scenario shows that contradictory psychological attributes can be had by the (virtual) persons manifested by the system.«
[88] Ibid., p. 399: »I conclude that Searle is correct in holding that no digital machine could understand language, but wrong in holding that artificial minds are impossible: minds and persons are not the same as the machines, biological or electronic, that realize them.«

um, više umova ili pak nijedan neovisno o tome koristi li sustav neurone (kao kod ljudi), cijele ljude (kao u kineskoj sobi) ili programirana računala (kao kod umjetne inteligencije).[89]

Međutim, postoje određeni problemi s funkcionalizmom, stoga i s Coleovom teorijom. Prvo, nailazimo na »intuitivno privlačnu tvrdnju koju često šire kritičari funkcionalizma: da zaista jest važno od čega stvarate um.«[90] Oni tvrde kako se, između ostaloga, »ne može […] napraviti *osjetilni* um iz silicijskih čipova«[91]. No, to i dalje, prema mišljenju D. C. Dennetta, nije razlog za napuštanje funkcionalizma.[92] Naime, »*jedini* razlog zbog kojeg umovi ovise o kemijskom sastavu svojih mehanizama ili medija jest sljedeći: kako bi ti mehanizmi činili stvari koje moraju činiti, oni moraju biti sazdani, kako nas uče biopovijesne činjenice, od tvari koje su kompatibilne s prethodno postojećim tijelima koja kontroliraju.«[93] Također, kada Cole navodi da je tijelo tek supstrat, postoji opasnost da tijelo počnemo smatrati manje vrijednim i sekundarnim. Međutim, Dennett ističe da se, na primjer, mene »ne može […] otrgnuti od tijela ostavljajući čiste rubove« jer »moje tijelo sadrži jednako toliko *mene*—vrijednosti, sposobnosti, sjećanja i raspoloženja koja čine ono što ja jesam—koliko i moj živčani sustav.«[94] Primjer ovakve uske povezanosti s tijelom prikazuje istraživanje o povezanosti emocionalnoga stanja s raznolikošću i sastavom crijevnoga mikrobioma.[95] Nadalje, ako se prihvati Dennettova tvrdnja o važnosti tijela, hipotetski bi bilo moguće teleportirati čovjeka, no pritom bi nužno morali sačuvati informaciju o cijelome tijelu, a ne samo o njegovim specifičnim instancama.[96] Uzimajući u obzir navedeni hipotetski slučaj teleportiranja ljudi, slučaj bi prebacivanja osoba iz strojeva bio nešto drukčiji i, vjerujem, vrlo jednostavniji zbog prirode strojeva. No, priroda strojeva dovodi nas do novoga problema – problema tzv. *kvalije* (lat. *qualia*). Kvaliju možemo opisati kao »ključni sastojak koji svijest, a onda i ljudski um, čini nečim tako posebnim i

[89] Ibid., p. 415: »From the fact that there is a single physical system, then, nothing follows about the number of minds which the system might realize. Depending on the causal character of the system, it might realize no minds, one mind, or more than one mind. This is the case whether the system employs neurons, as in humans, entire humans, as in the Chinese Room, or programmed computers, as in AI.«
[90] Daniel C. Dennett, *Vrste umova: k razumijevanju svijesti*, s engleskog preveo Ivan Kraljević (Zagreb: In.Tri, 2017), p. 70.
[91] Ibid.
[92] Ibid.
[93] Ibid.
[94] Ibid., p. 71.
[95] Sung-Ha Lee, Seok-Hwan Yoon, Yeonjae Jung, Namil Kim, Uigi Min, Jongsik Chun, i Incheol Choi, »Emotional well-being and gut microbiome profiles by enterotype«, u: *Scientific Reports*, dostupno na: https://www.nature.com/articles/s41598-020-77673-z (pristupljeno 17.02.2022), »[…] the present study links one's emotional status to gut microbiome diversity and composition.«
[96] Dennett, *Vrste umova: k razumijevanju svijesti*, p. 71.

različitim od svega ostaloga u svemiru, a to je tzv. 'fenomenalna svijest' odnosno subjektivni i kvalitativni sadržaj iskustva.«[97]

Problem je kvalije opći problem s kojim se suočava funkcionalistička teorija i nejasno je kako Coleova virtualna osoba ima kvaliju, ali pojedini znanstvenici poput Maxa Tegmarka vjeruju da je prostor mogućih iskustava umjetne inteligencije ogroman u usporedbi s onime što ljudi mogu doživjeti.[98] Ljudi posjeduju jednu klasu kvalije za svako osjetilo, ali umjetna inteligencija može imati znatno više vrsta senzora i unutarnjih reprezentacija informacija koje su čovjeku jednostavno nedostupne, stoga treba odbaciti pretpostavku da je biti umjetna inteligencija sličnoga osjećaja kao biti ljudska osoba.[99] Pojedini aspekti ljudskoga subjektivnog iskustva imaju svoje evolucijsko ishodište, a primjer su toga emocionalne želje vezane uz samoodržavanje (hranidba, izbjegavanje smrti) i reprodukciju, no navedeno ne važi za umjetnu inteligenciju i stoga bi trebalo biti moguće stvoriti takvu umjetnu inteligenciju koja nikada ne doživljava kvalije kao što su glad, žeđ, strah i seksualna želja.[100] Također, umjetni bi um mogao nadoknaditi nedostatke našega ljudskog uma. Naime, »naši prirodni umovi mogu se nositi samo s promjenama koje se odvijaju određenim tempom«[101] i stoga »događaji koji se odvijaju brže ili sporije od toga za nas su naprosto nevidljivi.«[102] Taj smo nedostatak donekle svladali pomoću raznih tehnoloških sredstava među koje na primjer pripada i fotografija. »Susan Sontag ističe da je nastanak ultrabrze fotografije bio revolucionaran tehnološki napredak za znanost, jer je ljudima omogućio da po prvi put u povijesti ispitaju složene vremenske pojave ne u stvarnome vremenu, nego u svome vremenu—u ležernoj, metodičnoj i povratnoj analizi tragova«[103]. Međutim, stroj, teorijski, ne bi patio od ovakvih ograničenja jer bi posjedovao mogućnost vraćanja viđenoga iz pohrane u svijest.

No, osim navedenih problema, ostaje nejasno kako Coleova virtualna osoba uči, obrađuje, prikuplja, pohranjuje i kreira nove informacije. Još jedan od problema iz Coleova primjera jest što se u korejskoj sobi utjelovljuju osobe na već unaprijed danoj bazi podataka,

---

[97] Pavel Gregorić, »Pogovor hrvatskom izdanju«, u: Daniel C. Dennett, *Vrste umova: k razumijevanju svijesti* (Zagreb: In.Tri, 2017), pp. 151–162, na p. 157.
[98] Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence*, p. 394: »[...] the space of possible AI experiences is huge compared to what we humans can experience.«
[99] Ibid., »We have one class of qualia for each of our senses, but AIs can have vastly more types of sensors and internal representations of information, so we must avoid the pitfall of assuming that being an AI necessarily feels similar to being a person.«
[100] Ibid., p. 396: »Some aspects of our subjective experience clearly trace back to our evolutionary origins, for example our emotional desires related to self-preservation (eating, drinking, avoiding getting killed) and reproduction. This means that it should be possible to create AI that never experiences qualia such as hunger, thirst, fear or sexual desire.«
[101] Dennett, *Vrste umova: k razumijevanju svijesti*, p. 125.
[102] Ibid.
[103] Ibid.

tj. sustav utjelovljuje jednu staricu i jednog mladića čija su iskustva, kako se implicitno prikazuje, naknadno pohranjena te je stoga nejasna usuglašenost primjera s funkcionalističkom tvrdnjom da osobu čini snop psihološke kauzalnosti. Drukčije rečeno, u njegovome su primjeru osobe već unaprijed »dane« pa nerazjašnjenim ostaje kauzalna povezanost između prijašnjih iskustava tih osoba i priložene baze podataka na temelju koje se te osobe utjelovljuju. Naravno, Cole u radu implicitno prihvaća funkcionalističku tvrdnju da »ono što neku stvar čini umom (ili vjerovanje, bol, strah itd.) nije od čega je sazdano, nego *kako djeluje*«[104], ali ako novokreiranom entitetu (ili sustavu) dodamo prijašnja iskustva drugog entiteta (ili sustava), tj. dodamo mu njegov um kako bi utjelovili istu osobu, postavlja se pitanje ne gubi li se time direktna psihološka kauzalnost? Taj problem naizgled nestaje ako prihvatimo Dennettovu prethodno navedenu tvrdnju da je očuvanje informacije ključ bivanja osobom, ali onda postaje nejasnim kako tranzitivnost između elektroničkoga i organskog tijela utječe na očuvanje te informacije. Prethodno sam naveo i Dennettov stav da nas se ne može odvojiti od tijela i da bi npr. pri teleportaciji nužno morali sačuvati informaciju o cijelome tijelu. Na temelju toga, izgleda da može postojati tranzitivnost između organskoga i organskog tijela gdje se informacija očuva 1:1, ali to ne govori ništa o tome da tranzitivnost može postojati i u slučaju gdje prebacujemo osobu iz organskoga tijela u elektronički uređaj jer se informacija organskoga tijela ne može u cjelini prebaciti radi različitosti sustava. Također, vrijedi i obratno te bi »osoba« prebačena iz stroja u organsko tijelo bila zakinuta u svojim subjektivnim doživljajima jer, kao što je Max Tegmark iznio, stroj za razliku od čovjeka može teorijski raspolagati s više vrsta senzora i unutarnjih reprezentacija. Naposljetku, ako ne prihvatimo stav da je kvalija »obična filozofska izmišljotina koja nas samo navodi na krivi trag«[105], upitno ostaje kako Coleove virtualne osobe uopće doživljavaju svijet i mogu li se one smatrati osobama u punome smislu, tj. mogu li se smatrati pravnim i moralnim subjektima.

Mnogo toga ostaje nejasno u Coleovom primjeru o osobama, a na kraju i sam Cole ne vjeruje da je moguće dokazati da u scenariju kineske sobe postoji osoba koja razumije kineski, no to proizlazi iz opće poteškoće poznate kao *problem drugih umova*.[106] Također, u primjerima korejske i kineske sobe nema govora o uvjetima koje bi osoba morala ispunjavati kako bi se uopće smatrala osobom, već je sva pažnja usmjerena na problem razumijevanja, realiziranja uma i tko je ostvaren realiziranim umom. I iako um i razumijevanje jesu jedni od ključnih

[104] Ibid., p. 63.
[105] Gregorić, »Pogovor hrvatskom izdanju« pp. 157–158.
[106] Cole, »Artificial Intelligence and Personal Identity«, p. 410: »I do not believe it can be proven that there is a person who understands Chinese in the scenario. But this difficulty is a completely general difficulty familiar as The Problem of Other Minds.«

uvjeta za bivanje osobom, oni nisu jedini uvjeti niti su detaljno razrađeni u Coleovom primjeru. Stoga, u ostatku ću rada prikazati uvjete bivanja osobom koje bi umjetna inteligencija morala biti u stanju ispuniti kako bi je se moglo promatrati kao entitet koji posjeduje status osobe.

3. Teorije i uvjeti osobnog identiteta

Ljudi su jedina bića koja trenutno prepoznajemo kao osobe, no s lakoćom možemo zamisliti postojanje biološki drugačijih osoba koje nastanjuju druge planete.[107] Također, možemo zamisliti uvjete koji bi izuzeli ljudska bića od bivanja osobom ili barem od imanja pojedinih značajnih elemenata osobe.[108] Takvim bi pojedinim ljudskih bićima ostatak društva mogao putem društvene konvencije pripisati status osobe, tj. osobni identitet, no problem bi nepostojanja intrinzičnog realiteta ostao neriješen. Za nas, nositelje osobnoga identiteta, koncept je osobnoga identiteta važan zbog njegove uske povezanosti s našim konceptom odgovornosti za prošla djela i našim praksama hvale i krivnje.[109] Mi, dakle, prihvaćamo postojanje koncepta osobnoga identiteta, no nerazjašnjenim ostaje pitanje što je nositelj identiteta kroz prostor i vrijeme. Pregled mogućih odgovora nudi Harold W. Noonan u svome djelu *Personal Identity*. U djelu H.W. Noonan prikazuje sljedeće teorije, tj. kriterije osobnoga identiteta: tjelesni kriterij, kriterij mozga, fizički kriterij, kriterij pamćenja, kriterij psihološkoga kontinuiteta i tzv. »jednostavni pogled« (eng. *the simple view*). Među navedene kriterije osobnoga identiteta možemo dodati i kriterij duše koji zastupa Alexander R. Pruss.[110] Sve navedene teorije razmatraju što je nositelj osobnoga identiteta kroz vrijeme, no nijedna se izravno ne suočava s uvjetima bivanja osobom. Teorije koje iznose Harold W. Noonan i Alexander R. Pruss dijele jedan zajednički aksiom, a to je da biće koje posjeduje identitet jest umno biće. Drukčije rečeno, kako bi se osobni identitet realizirao, potrebno je da biće posjeduje određeni um. Kod Aristotela pronalazimo razliku između triju vrsta duša koje odgovaraju različitim bićima, a to su vegetativna, senzitivna i razumska duša.[111] Biljke posjeduju navedenu vegetativnu dušu, životinje posjeduju senzitivnu dušu, a ljudi posjeduju razumsku dušu koja ih izdiže nad ostalim bićima jer uključuje sposobnost mišljenja.[112] Sposobnost mišljenja, koju

[107] Daniel C. Dennett, »Conditions of Personhood«, u: *The Identities of Persons* (University of California Press, 1976), dostupno na: https://philpapers.org/rec/DENCOP (pristupljeno 20.02.2022), pp. 175–196, na p. 175: »At this time and place human beings are the only persons we recognize, […] as persons, but on the one hand we can easily contemplate the existence of biologically very different persons – inhabiting other planets […]«

[108] Ibid., »[…] and on the other hand we recognize conditions that exempt human beings from personhood, or at least some very important elements of personhood.«

[109] Harold W. Noonan, *Personal Identity* (London: Taylor & Francis Group, 2005), p. 1: »Again, our concept of personal identity is intimately linked with our concept of responsibility for past actions and with our practices of praise and blame […]«

[110] Vidi: Alexander R. Pruss, »Artificial Intelligence and Personal Identity«, u: *Faith and Philosophy: Journal of Society of Christian Philosophers*, Volume 26, Iss 5, Article 2 (2009), dostupno na: https://place.asburyseminary.edu/faithandphilosophy/vol26/iss5/2/ (pristupljeno 20.02.2022), pp. 487–500.

[111] Branko Bošnjak, »Predgovor«, u: Aristotel, *O duši*, preveo Milivoj Sironić (Zagreb: Naprijed, 1996), pp. VII–XLII, na p. XIII.

[112] Ibid.

nema nijedno drugo dosad poznato biće, omogućuje ljudima imanje apstraktnih ideja poput osobnoga identiteta. Međutim, iako je ispravno tvrditi da je određena sposobnost mišljenja koju posjeduju ljudi ključ posjedovanja osobnoga identiteta, ta je tvrdnja nedovoljna zbog svoje preopćenitosti koja može navesti na intuitivne pogreške. Stoga ću pomoću Dennettove analize uvjeta osoba i njegova intencionalnoga stajališta iznijeti kriterije koje bi umjetna inteligencija, ili bilo koje drugo biće, morala zadovoljiti kako bi se proglasila osobom.

3.1. Dennettovi uvjeti bivanja osobom

Kako bi se odgovorilo na pitanje može li umjetna inteligencija posjedovati osobni identitet, potrebno je prvo istražiti same uvjete bivanja osobom. Daniel C. Dennett u svojemu tekstu »Conditions of Personhood« prikazuje šest tema od kojih svaka nastoji predočiti neophodan uvjet osobnosti. No, prije samog prikaza uvjeta, navodim problem dualnosti koncepta osobe koji iznosi D. C. Dennett. Naime, potraga za nužnim i dovoljnim uvjetima može naići na probleme ako postoji više od jednoga koncepta osobe.[113] Ta se dualnost osobe sastoji u njezinoj podjeli na moralni i metafizički koncept.[114] Zaključak koji Dennett iznosi za taj problem jest da čak i ako pretpostavimo da postoji razlika između moralnoga i metafizičkog koncepta, postoji razlog za vjerovati da metafizički koncept prethodi moralnome, tj. da je metafizički koncept njegov nužan preduvjet.[115] Navedeno, vjerujem, izmiče *post hoc ergo propter hoc* pogrešci te nudi zadovoljavajuće rješenje ovoga problema i omogućava postojanje jednoga temeljnog koncepta osobe iz kojeg se dalje izvode svi ostale. Dakle, zauzet ću stajalište da postoji samo jedan temeljni koncept osobe i prikazati ću uvjete njegova postojanja koje Dennett tematski prezentira.

Prva tema koju D. C. Dennett navodi jest da su osobe racionalna bića.[116] Navedeno pronalazimo u etičkim teorijama Immanuela Kanta i Johna Rawlsa, a potom i u metafizičkim teorijama Aristotela i Hintikke.[117] U drugoj se temi iznosi tvrdnja da su osobe bića kojima pripisujemo stanja svijesti ili kojima pripisujemo psihološke, mentalne ili pak intencionalne

[113] Dennett, »Conditions of Personhood«, p. 176: »Supposing there *is* something more to being a person, the searcher for necessary and sufficient conditions may still have difficulties if there is more than one concept of a person […]«
[114] Ibid., »Roughly, there seem to be two notions intertwined here, which we may call the moral notion and the metaphysical notion.«
[115] Ibid., p. 177: »Still, even if we suppose there are these distinct notions, there seems every reason to believe that metaphysical personhood is a necessary condition of moral personhood.«
[116] Ibid., »The *first* and most obvious theme is that persons are *rational beings*.«
[117] Ibid., »It figures, for example, in the ethical theories of Kant and Rawls, and in the 'metaphysical' theories of Aristotle and Hintikka.«

iskaze.[118] U trećoj temi Dennett prikazuje kako bivanje nekoga bića osobom na neki način ovisi o stavu koji se prema tome biću zauzme, tj. o usvojenomu stajalištu u odnosu na to biće.[119] Četvrta se tema direktno nadovezuje na treću i u njoj se tvrdi da objekt, prema kojem je usvojen prethodno navedeni stav, mora posjedovati sposobnost recipročnosti.[120] Ta se recipročnost ponekad neinformativno izražava sloganom da biti osobom znači postupati prema drugim bićima kao prema osobama, a uz taj se slogan često povezivala i tvrdnja da takvo postupanje prema drugim osobama znači moralno postupanje prema njima.[121] Taj je način tumačenja uvjeta recipročnosti tautologija, a nadovezujuća tvrdnja negira svu raznolikost međuodnošenja među osobama. Naime, Thomas Nagel tvrdi da je izrazito neprijateljsko ponašanje prema drugome biću kompatibilno s tretiranjem toga bića kao osobe.[122] Na navedenu se tvrdnju može nadovezati i primjer Vana de Vatea koji piše da je jedna od razlika između određenih ubojstava iz nehaja i umorstva ta što ubojica kod umorstva žrtvu tretira kao osobu[123], stoga nije neispravno tvrditi da se tretiranje drugoga bića kao osobe ne mora nužno temeljiti na »moralnome« postupanju. Nadalje, Dennett u petoj temi prikazuje kako osobe moraju biti sposobne za verbalnu komunikaciju.[124] Taj je uvjet verbalne komunikacije, prema Dennettovu mišljenju, implicitan uvjet svih teorija društvenoga ugovora iz područja etike i isključuje nama poznate neljudske životinje od imanja potpunoga statusa osobe i moralne odgovornosti.[125] Posljednja tema prikazuje kako se osobe razlikuju od ostalih entiteta na temelju toga što su svjesne na neki jedinstven način; postoji specifičan način na koji smo mi ljudi svjesni i na koji nijedna druga vrsta nije.[126] Peta i šesta tema usko su povezane jer je sposobnost za verbalnu komunikaciju uvjet za imanje posebne vrste svijesti koja je prema mišljenju G. E. M.

[118] Ibid., p. 177: »The *second* theme is that persons are beings to which states of consciousness are attributed, or to which psychological or mental or *Intentional predicates*, are ascribed.«
[119] Ibid., »The *third* theme is that whether something counts as a person depends in some way on an *attitude taken* toward it, a *stance adopted* with respect to it.«
[120] Ibid., p. 178: »The *fourth* theme is that the object toward which this personal stance is taken must be capable of *reciprocating* in some way.«
[121] Ibid., »This reciprocity has sometimes been rather uninformatively expressed by the slogan: to be a person is to treat others as persons, and with this expression has often gone the claim that treating another as a person is treating him morally […]«
[122] Ibid., »As Nagel says, 'extremely hostile behavior toward another is compatible with treating him as a person' […]«
[123] Ibid., »[…] Van de Vate observes, one of the differences between some forms of manslaughter and murder is that the murderer treats the victim as a person.«
[124] Ibid., »The *fifth* theme is that persons must be capable of *verbal communication*.«
[125] Ibid., »This condition handily excuses nonhuman animals from full personhood and the attendant moral responsibility, and seems at least implicit in all social contract theories of ethics.«
[126] Ibid., »The *sixth* theme is that persons are distinguishable from other entities by being *conscious* in some special way: there is a way in which we are conscious in which no other species is conscious.«

Anscombe i Harryja Frankfurta nužan uvjet za bivanje moralnom osobom.[127] Također, ako biće nije sposobno za verbalnu komunikaciju niti ima posebnu vrstu svijesti, to biće nije u mogućnosti zauzeti stajalište prema drugome biću ili recipročno odgovoriti na zauzeto stajalište.

Dennett je prve tri teme ili uvjeta – racionalnost, intencionalnost i stav – iskoristio za definiranje, ne osoba, već mnogo šire klase entiteta zvanih intencionalni sustavi.[128] Ostala su tri uvjeta također prisutna u intencionalnim sustavima, no oni su, kao što ću prikazati, prisutni tek u intencionalnim sustavima višega reda. Stoga, unatoč tomu što bivanje intencionalnim sustavom nije dovoljan uvjet za bivanje osobom, očito je da je ono nužan uvjet.[129]

3.2 Intencionalni sustavi

Prethodno je utvrđeno da je bivanje intencionalnim sustavom nedovoljan, ali nužan uvjet kako bi se nekoga smatralo osobom. U ovome ću dijelu rada istražiti kakav bi to intencionalni sustav umjetna inteligencija morala biti kako bi zadovoljila nužne kriterije, stekla status umnoga bića, a potom i status osobe. No, prvo je potrebno definirati pojmove »intencionalno stajalište« i »intencionalni sustav«.

Dennett piše da je »intencionalno stajalište […] strategija tumačenja ponašanja nekog entiteta (osobe, životinje, predmeta, čega god) kao da je racionalni djelatnik čiji je 'odabir' neke 'radnje' određen 'razmatranjem' njegovih 'vjerovanja' i 'želja'.«[130] Kada se koristi intencionalno stajalište, potrebno je »da se prema nekom entitetu odnosi kao prema djelatniku kako bi se predvidjele—i time u određenom smislu objasnile—njegove radnje ili potezi.«[131] Dakle, pažljivim zauzimanjem intencionalnoga stajališta, možemo otkriti vrstu umova pojedinih entiteta.[132] Zauzeti pak intencionalno stajalište možemo spram tzv. intencionalnih sustava. Dennett piše da su »intencionalni sustavi […] po definiciji svi oni, i samo oni, entiteti čije je ponašanje predvidljivo/objašnjivo iz intencionalnog stajališta.«[133] Među te intencionalne

[127] Ibid., p. 179: »[…] the capacity for verbal communication, which is the necessary condition for having a special sort of consciousness, which is, as Anscombe and Frankfurt in their different ways claim, a necessary condition of moral personhood.«
[128] Ibid., »I have previously exploited the first three themes, rationality, Intentionality and stance, to define not persons, but the much wider class of what I call *Intentional systems* […]«
[129] Ibid., p. 180: »It is obvious, then, that being an Intentional system is not sufficient condition for being a person, but is surely a necessary condition.«
[130] Dennet, *Vrste umova: k razumijevanju svijesti*, p. 30.
[131] Ibid., p. 31.
[132] Ibid., p. 30.
[133] Ibid., p. 36.

sustave ubrajamo jednostavne organizme kao što su na primjer samoreplicirajuće makromolekule i amebe, a potom i kompleksnije organizme poput biljaka, životinja i ljudi te strojeve poput šahovskih računala.[134] Dakle, kada zauzimamo intencionalno stajalište prema određenom entitetu, prvo se prema njemu »odnosimo kao prema djelatniku kako bismo predvidjeli njegove postupke«[135], a potom »opisujemo dotičnu ograničenu perspektivu [djelatnika] tako što djelatniku pripisujemo konkretna vjerovanja i želje na temelju njegova uvida u situaciju i na temelju njegovih ciljeva i potreba.«[136] Drukčije rečeno, »kad god djelatnik djeluje, on to čini na temelju određenog razumijevanja—ili nerazumijevanja—okolnosti, a intencionalna objašnjenja i predviđanja oslanjaju se na zahvaćanje tog razumijevanja.«[137] Kako bi se zahvatile i predvidjele radnje intencionalnoga sustava, potrebno je »znati na koje su stvari usmjerena njegova vjerovanja i želje«[138], a uz to je potrebno i »znati, barem ugrubo, *kako* su njegova vjerovanja i želje usmjerena na te stvari, tako da možete reći jesu li, ili hoće li biti, napravljene ključne poveznice.«[139] Dakle, »kada zauzimamo intencionalno stajalište, moramo *barem ugrubo* znati na koji način djelatnik izdvaja relevantne predmete iz svoje okoline.«[140]

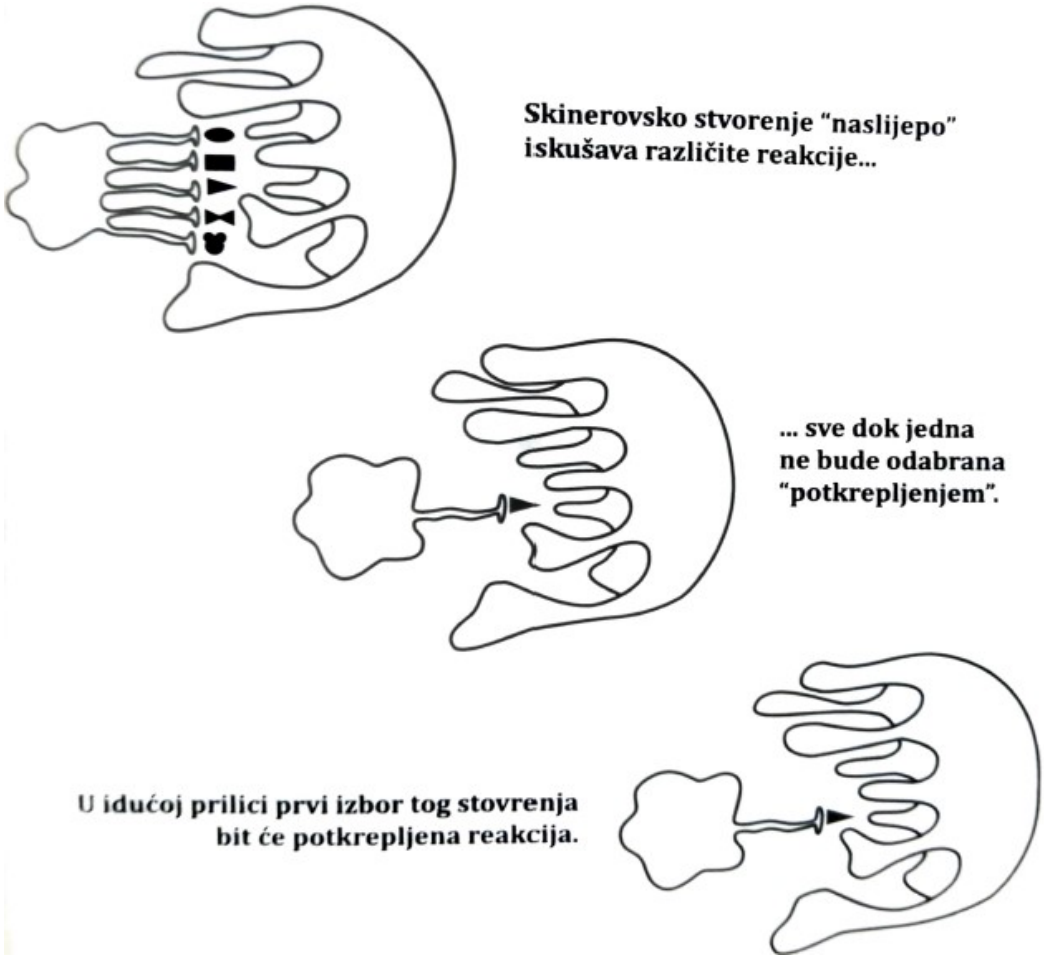
Odlučivši se za intencionalno stajalište, ujedno smo se odlučili i za naturalističko objašnjenje svijesti.[141] Naime, kada prihvaćamo intencionalno stajalište, ujedno »prihvaćamo da je ljudski um ili svijest samo jedan od mnogo različitih načina da nešto bude intencionalni sustav«[142], a taj »način – nedvojbeno jako složen i moćan – može biti rezultat evolucijskih procesa iz manje složenih i manje moćnih načina da nešto bude intencionalni sustav«[143]. Time »Dennett ne želi reći da je svaki intencionalni sustav svjestan, nego samo to da je svaki sustav koji ima um i koji je svjestan – intencionalni sustav.«[144] Sad preostaje prikazati kakav bi to intencionalni sustav morala biti umjetna inteligencija da zadovolji nužne kriterije i da bi za nju mogli reći da posjeduje umnost i svijest.

[134] Ibid.
[135] Ibid.
[136] Ibid.
[137] Ibid., p. 42.
[138] Ibid.
[139] Ibid.
[140] Ibid.
[141] Gregorić, »Pogovor hrvatskom izdanju« p. 159.
[142] Ibid.
[143] Ibid.
[144] Ibid., p. 160.

3.3. Vrste intencionalnih sustava

Sustavi koji iskazuju minimalnu intencionalnost i kod kojih je uočljiv početak djelatništva jednostavni su organizmi poput makromolekula.[145] Na temelju strukture svojih sustava u mogućnosti su »da izvode neke radnje, umjesto da samo leže i trpe neke učinke.«[146] Međutim, kod njih nema znanja o onome što čine, tj. oni čine bez svijesti o razlozima činjenja.[147] No, »njihova vrsta djelatništva jedino je moguće tlo iz kojeg je moglo izrasti sjeme naše vrste djelatništva.«[148] Naša se vrsta djelatništva nalazi na vrhu hijerarhije koju Dennett naziva *Toranj Generiranja-i-Testiranja*. Ta hijerarhija dijeli intencionalne sustave prema njihovim kognitivnim moćima.[149] Na samome se dnu te hijerarhije nalaze tzv. *darvinovska stvorenja* i oni su prema Dennettovu mišljenju »stvoreni naslijepo, više–manje proizvoljnim procesima rekombinacije i mutacije gena.«[150] U podskup tih stvorenja spadaju tzv. *skinerova stvorenja* (slika 1.), a Dennett im je dodijelio ime prema psihologu Burrhusu Fredericu Skinneru.[151] Ta su stvorenja nešto složenija od običnih darvinovskih i za njih je karakteristično tzv. operantno uvjetovanje, tj. oni isprobavaju raznovrsne postupke na temelju kojih dobivaju pozitivne ili negativne signale iz okoline kojima potkrepljuju, a time ujedno i favoriziraju, određene postupke.[152] Dennett te favorizirane postupke naziva »pametnim postupcima«, a kako bi uopće došlo do postojanja tih favoriziranih postupaka, ti organizmi moraju biti ožičeni *potkrepljivačima* koji će ih voditi kroz okolinu, a u slučaju da im ti potkrepljivači neispravno funkcioniraju, oni su osuđeni na propast.[153] Za ta je stvorenja ujedno karakteristično raspolaganje sirovom intencionalnošću i mijenjanje vlastitih planova kako bi odgovorili na nove uvjete, a da pritom u njima ne postoji predstavljeno znanje tih razloga promjene.[154] Oni, naime, raspolažu samo s uskim skupom informacija pomoću kojih znaju kako nešto odraditi, ali si to znanje ne mogu iznutra predstaviti, tj. nemaju mišljenje o njemu.[155]

[145] Dennett, *Vrste umova: k razumijevanju svijesti*., p. 25.
[146] Ibid.
[147] Ibid.
[148] Ibid.
[149] Ibid., p. 76.
[150] Ibid., p. 77.
[151] Ibid., p. 78.
[152] Ibid.
[153] Ibid.
[154] Ibid., p. 133–134.
[155] Ibid.



Slika 1. Prikaz *skinerovskog stvorenja*[156]

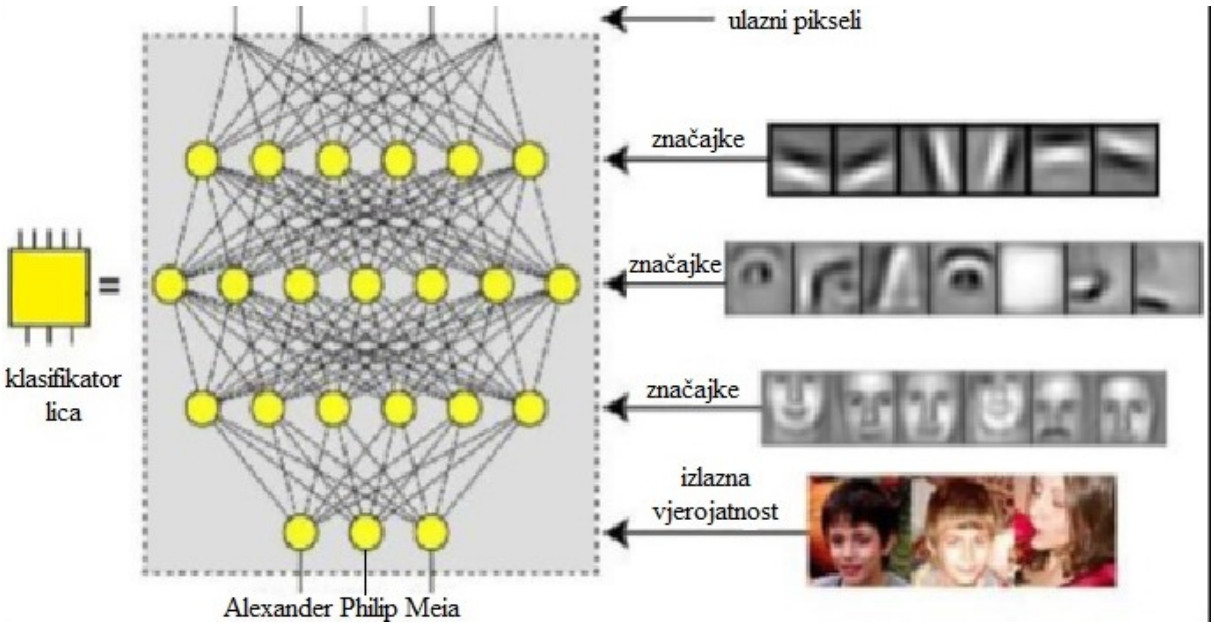
Takav se način skinerovskoga isprobavanja jednoga po jednog postupka do pronalaska onog učinkovitoga pronalazi i u računarstvu kod tzv. *brute force* algoritama, no tu zapravo ne nailazimo na iskustveno potkrepljenje. Međutim, ideja je iskustvenoga potkrepljenja uvelike prisutna u teoriji znanoj kao konekcionizam čija definicija glasi:

»Konekcionizam je teorija u filozofiji uma, kognitivnoj znanosti i umjetnoj inteligenciji koja počiva na modelu neuronske mreže. Radi se o sustavu velikog broja jednostavnih jedinica među kojima se stvaraju međusobne veze ovisno o jačini njihova potencijala za aktivaciju, što je pak određeno njihovim prethodnim aktivnostima. Tako se neke veze učvršćuju, a neke zatomljuju.«[157]

Teorije su neuronskih mreža i konekcionizma doživjele procvat 90-ih godina prošloga stoljeća kada je njihovo istraživanje pokazalo da se jednostavne mreže, koje započinju s manje–više slučajnim ožičenjem, putem iskustvenoga potkrepljivanja izmjenjuju, odnosno da na temelju

[156] Ibid., p. 79.
[157] Ibid, p. 78, fusnota 4.

okoline uređuju svoje veze.[158] Trenutno teorija neuronskih mreža dominira područjem umjetne inteligencije znanim kao strojno učenje (eng. *machine learning*) koje se bavi proučavanjem algoritama koji se pomoću iskustva poboljšavaju.[159] Samu neuronsku mrežu možemo shvatiti kao skup međusobno povezanih neurona koji mogu utjecati na ponašanje jednih na druge.[160] Neuronska mreža može vršiti izračunavanje funkcija te su tako na primjer umjetne neuronske mreže osposobljene da na temelju ulaznih informacija svjetlosti piksela brojčano predstave vjerojatnost da se na slici nalaze ljudi.[161] Primjer je toga vidljiv na slici broj dva gdje svaki umjetni neuron prima informacije odozgo, potom izvršava jednostavnu funkciju i rezultat prosljeđuje nadolje gdje se vrši izračun značajki više razine.[162] Taj primjer predstavlja mrežu za prepoznavanje lica koja inače sadrži stotine tisuća neurona, no na slici je prikazana samo mala količina radi jasnoće.[163]



Slika 2. Prikaz neuronske mreže prepoznavanja lica[164]

[158] Ibid, p. 78–79.
[159] Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence*, p. 97: »Neural networks […] have recently started dominating the AI subfield known as *machine learning* (the study of algorithms that improve through experience).«
[160] Ibid., »A neural network is simply a group of interconnected neurons that are able to influence each other's behavior.«
[161] Ibid., p. 98: A network of neurons can compute functions […]. For example, artificial neural networks have been trained to input numbers representing the brightness of different image pixels and output numbers representing the probability that the image depicts various people.
[162] Ibid., Here each artificial neuron (circle) computes a weighted sum of the numbers sent to it via connections (lines) from above, applies a simple function and passes the result downward, each subsequent layer computing higher-level features.
[163] Ibid., Typical face recognition networks contain hundreds of thousands of neurons; the figure shows merely a handful for clarity.
[164] Ibid., vlastiti prijevod.

Nadalje, razvojem su strojnog učenja istraživači uspjeli 2014. godine napraviti umjetnu inteligenciju koja je uspjela opisati sadržaj slika tako što su jednostavnu neuronsku mrežu bez ikakvog znanja o vanjskome svijetu pustili da »uči« i izložili je masovnoj količini podataka.[165] Ta sposobnost prepoznavanja lica i opisivanja slika putem »učenja« govori o mogućnostima umjetne inteligencije i naizgled odražava jedno inteligentno postupanje, no koje i dalje ostaje bez unutarnjega predstavljanja. Ipak, na temelju tvrdnji o skinerovim bićima i neuronskim mrežama koje prikupljaju podatke izvana, usudio bih se nazvati uređivanje veza na temelju pozitivnoga ili negativnog signala iz okoline nekakvim počecima učenja jer entiteti više ne moraju svaki put iznova naslijepo isprobavati moguća rješenja, već na temelju prethodnih zbivanja »znaju« što uraditi. Međutim, način isprobavanja skinerovskih bića nije uvijek učinkovit ni ekonomičan, a ponajmanje je odraz umnosti. Također, kao što Dennett uviđa, »skinerovsko uvjetovanje je dobra stvar ako vas ne ubije neka od vaših ranijih pogrešaka.«[166]

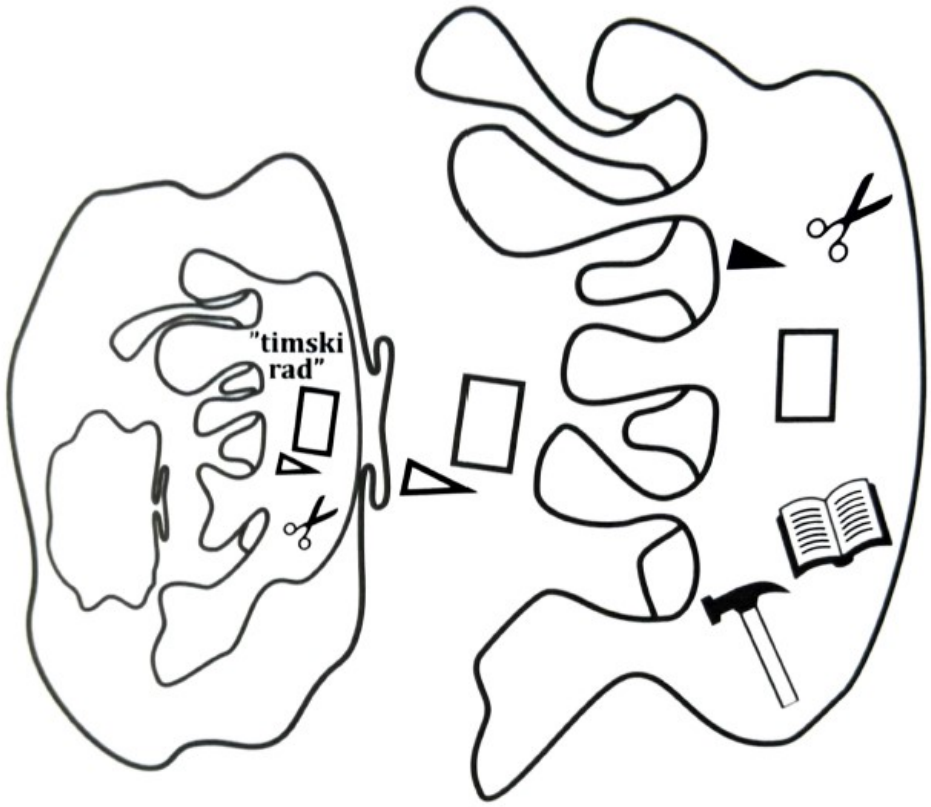
Bića koja su prema kognitivnim moćima u hijerarhiji iznad skinerovskih, Dennett naziva *poperovskim stvorenjima*.[167] Za razliku od skinerovskih stvorenja, njihov »sustav uključuje predodabir među svim mogućim ponašanjima ili radnjama, tako da doista glupi potezi budu iskorijenjeni prije nego što ih se iskuša u 'stvarnom životu'.«[168] Naziv su dobila prema filozofu Karlu Popperu koji je ustvrdio da »ovo usavršenje dizajna 'omogućuje da naše hipoteze umru umjesto nas'.«[169] Dakle, ta »stvorenja preživljavaju jer su dovoljno pametna da čine prve poteze koji nisu puko nasumični.«[170] Odabir tih poteza ovisi o unutarnjemu filteru koji biće posjeduje,[171] a njega se shvaća kao »unutarnje okruženje u kojemu se potencijalne radnje ispituju i odabiru«[172], te stoga »to stvorenje, kad djeluje po prvi puta, čini to s određenom namišlju.«[173] Međutim, to se unutarnje okruženje ne smije shvatiti kao replika vanjskoga svijeta gdje su reproducirane sve fizičke okolnosti svijeta, već kao posjedovanje informacije pomoću koje je moguće odrediti potencijalni učinak djelovanja.[174] Mi ljudi smo poperovska stvorenja, ali nismo jedina. Poperovskim bi se bićima smatrale i razne životinje koje prikupljaju

[165] Ibid., p. 103: »Yet a team at Google […] did precisely that in 2014. […] How did they do it? […] by creating a relatively simple neural network with no knowledge whatsoever about the physical world or its contents, and then letting it learn by exposing it to massive amounts of data.«
[166] Dennett, *Vrste umova: k razumijevanju svijesti*, p. 81.
[167] Ibid.
[168] Ibid.
[169] Ibid.
[170] Ibid.
[171] Ibid.
[172] Ibid., p. 82.
[173] Ibid.
[174] Ibid., p. 83.

opće informacije kako bi odredile svoj plan djelovanja.[175] No, pitanje je što je s umjetnom inteligencijom, može li ona posjedovati unutarnje okruženje i na temelju toga stvarati pametne poteze? Smatram da je odgovor potvrdan. Naime, već sada umjetna sužena inteligencija može izvršavati određene simulacije, stoga nije nemoguće zamisliti postojanje umjetne inteligencije koja prikuplja podatke iz okoline, vrši »unutarnju« simulaciju mogućih reakcija i potom odlučuje o načinu djelovanja. No, postavlja se pitanje može li umjetna inteligencija išta »znati« o »stvarnome svijetu«? Pojedini su autori poput Marvina Minskyja optimistični i smatraju da će budućim napretkom tehnologije moći. Naime, Minsky tvrdi i da je ljudska ideja »zbilje« poprilično slična mreži, a inteligentne strojeve možemo ograničiti ili ih pustiti da grade svoje unutarnje mreže i time doći do solipsističke izolacije koju ne možemo ni zamisliti.[176]

No kognitivne su moći poperovskih stvorenja nedovoljne za imanje znanja o stvarnome svijetu koje omogućava bivanje osobom, stoga preostaje za prikazati posljednji »pod-pod-podskup darvinovskih stvorenja«[177] zvan *gregorijevska stvorenja*. Gregorijevska su stvorenja dobila naziv prema psihologu Richardu Gregoryju koji je zastupao tzv. teoriju *potencijalne inteligencije*.[178] »Gregory primjećuje da škare, kao dobro osmišljena umjetna tvorevina, nisu samo rezultat inteligencije, nego i darovatelj inteligencije (vanjske potencijalne inteligencije) u vrlo jednostavnom i intuitivnom smislu: kad nekome date škare, povećavate mu potencijal za sigurnije i brže dolaženje do Pametnih Poteza.«[179] Naime, riječ je o tome da upotreba određenih alata uzrokuje porast inteligencije, a alat koji najviše prenosi inteligenciju kod ljudi jest alat uma zvan jezik.[180]

[175] Ibid., p. 85.
[176] Marvin Minsky, »Why People Think Computers Can't«, u: *AI Magazine*, Volume 3, Number 4 (1982), dostupno na: https://doi.org/10.1609/aimag.v3i4.376 (pristupljeno 27.03.2022), pp. 3a–15b, na p. 10b–11a: »But in the final analysis, our idea of 'reality' itself is rather network-y. […] Finally, when we build intelligent machines we'll have a choice: either we can constrain them as we wish to match each and every concept to their outside-data instruments, or we can let them build their own inner networks and attain a solipsistic isolation totally beyond anything we humans could conceive.«
[177] Dennett, *Vrste umova: k razumijevanju svijesti*, p. 90.
[178] Ibid.
[179] Ibid.
[180] Ibid., pp. 90-91.



Slika 3. Prikaz gregorijevskog stvorenja[181]

Dennett utvrđuje kako »riječi i drugi umski alati daju gregorijevskom stvorenju unutarnje okruženje koje mu omogućuje da izradi sve suptilnije generatore i ispitivače poteza.«[182] Dakle, gregorijevska su stvorenja pametnija od prethodno navedenih stvorenja, a njihova se pamet temelji na tome što »čine velik korak prema ljudskoj razini mentalne snalažljivosti izvlačeći korist iz iskustva drugih tako što iskorištavaju mudrost utjelovljenu u umskim alatima koje su ti drugi izmislili, poboljšali i prenijeli«[183] te »tako uče kako da bolje razmišljaju o onome o čemu bi trebali razmišljati sljedeće—i tako dalje, stvarajući toranj daljnje unutarnje refleksije bez fiksnog ili vidljivog ograničenja.«[184] Životinja ne raspolaže jezikom koji čovjeka osposobljava za kompleksnije sadržaje mišljenja, stoga »*naša* sposobnost da formuliramo—pa čak u većini okolnosti i testiramo—hipoteze o identitetu stvari sasvim je strana svim ostalim stvorenjima.«[185] Naime, s obzirom na to da ostala stvorenja ne raspolažu jezikom, »prakse i

[181] Ibid., p. 91.
[182] Ibid.
[183] Ibid., p. 92.
[184] Ibid.
[185] Ibid., p. 105.

projekti mnogih tih stvorenja zahtijevaju da one prate i stalno iznova identificiraju jedinke«[186]. Također, intencionalnost tih stvorenja »nikad ne doseže razinu metafizičke pojedinačnosti do koje se uzdiže naša intencionalnost.«[187]

Mi ljudi imamo takvu sposobnost mišljenja, tj. formuliranja hipoteza, zahvaljujući umskome alatu zvanome jezik, ali da ga možemo biti u stanju koristiti, moramo biti opremljeni i nizom drugih sposobnosti.[188] No, s obzirom na to da ljudski um nije neograničenih sposobnosti, mi »ostavljamo koliko god je moguće samih podataka u vanjskom svijetu«[189], a »u našim mozgovima držimo 'smjernice' i 'kazala'«[190]. Pomoću jezika ostavljamo podatke u vanjskome svijetu i tako stvaramo kulturnu baštinu[191], a »zahvaljujući toj kulturnoj baštini učimo kako da svoje umove raširimo van u svijet gdje svoje prekrasno dizajnirane urođene sposobnosti praćenja i prepoznavanja uzoraka možemo optimalno iskoristiti.«[192] Tu sposobnost nema niti jedan drugi entitet, a umjetna inteligencija koja trenutno posjeduje proceduralno znanje ne posjeduje znanje pojma koji je nužan za refleksiju. Međutim, to ne znači da umjetna inteligencija ne posjeduje alat kojim proširuje znanje ni da ne može označavati stvari iz svoje okoline. Mogućnost indeksiranja nije nepoznanica, a umjesto jezika, trenutni je alat proširivanja znanja umjetne inteligencije kod. Trenutno se razvijaju sustavi koji mogu samostalno pisati programe, a primjer je toga sustav zvan *AlphaCode*[193]. No, takav sustav, naravno, još uvijek pripada području umjetne sužene inteligencije i nije na razini ljudskih stručnjaka. Također, iako kod kao alat nije identičan jeziku, a upotreba je jezika, kao što je prethodno navedeno, ključna komponenta bivanja osobom, to ne znači da je služenje jezikom nepoznanica na području umjetne inteligencije. Trenutno je moguće »razgovarati« s određenim sustavima umjetne sužene inteligencije, a određeni sustavi poput *AlphaCodea* mogu, govoreći iz perspektive intencionalnoga stajališta, »interpetirati« prirodni jezik. Interpretacija jezika nije novost te je tako još 1965. godine program Daniela Bobrowa, koristeći se raznim trikovima, mogao rješavati srednjoškolske algebarske zadatke s riječima.[194] Ipak,

[186] Ibid.
[187] Ibid.
[188] Ibid.
[189] Ibid., p. 126.
[190] Ibid.
[191] Ibid., p. 122.
[192] Ibid.
[193] Vidi: *Competitive programming with AlphaCode* (pristupljeno 28.03.2022), dostupno na: https://deepmind.com/blog/article/Competitive-programming-with-AlphaCode
[194] Minsky, »Why People Think Computers Can't«, p. 8: »[...] the 1965 program written by Daniel Bobrow that solves high-school algebra. "word problems?" [...] Bobrow's program used a lot of tricks.«

iako tu uviđamo korištenje riječi, to ne znači da umjetna inteligencija raspolaže predstavljenim znanjem i da reflektira o pojmovima koje »analizira«.

Kroz intencionalno smo stajalište prikazali vrste bića različitih kognitivnih mogućnosti i time pokazali što bi umjetna inteligencija trebala prevladati kako bi se izdigla nad stvorenjima koji su ograničenih kognitivnih mogućnosti i koji ne raspolažu jezikom ili, preciznije rečeno, pojmovima. Nadalje, u ovome smo poglavlju pokazali da umjetna inteligencija prikazuje inteligentno postupanje i da se, govoreći iz perspektive intencionalnih sustava, »služi« jezikom, no to i dalje ostaje na razini tzv. intencionalnosti prvoga reda. Naime, iako umjetnoj suženoj inteligenciji možemo pripisati vlastite »ciljeve« i »vjerovanja« kako bi postala osobom, treba se izdignuti iznad tzv. intencionalnosti prvoga reda. Dennett tvrdi da »važan korak k postajanju osobom jest prijelaz sa stepenice intencionalnog sustava prvoga reda na stepenicu intencionalnog sustava drugog reda.«[195] »Intencionalni sustav prvoga reda ima vjerovanja i želje o mnogim stvarima, ali ne i o samim vjerovanjima i željama.«[196] Međutim, »intencionalni sustav drugoga reda ima vjerovanja i želje i o vjerovanjima i željama, kako vlastitima tako i tuđima.«[197] Navedeno upućuje da intencionalni sustav drugoga reda posjeduje mogućnost reflektiranja, tj. on raspolaže spoznajnom jasnoćom koju nazivamo svijest. Prethodno smo pokazali trenutne mogućnosti umjetne inteligencije, ali nijedna od tih mogućnosti ne prikazuje ikakvu mogućnost imanja predstavljenoga znanja ili reflektiranja o vlastitim djelima. Stoga, novo pitanje glasi: može li umjetna inteligencija imati svijest? Tim se pitanjem nadilaze mogućnosti umjetne sužene inteligencije i prelazi se u spekulativna područja o svijesti, umjetnoj općoj inteligenciji (eng. *general artificial intelligence*) i umjetnoj super inteligenciji (eng. *artificial super intelligence*).

[195] Dennett, *Vrste umova: k razumijevanju svijesti*, p. 107.
[196] Ibid.
[197] Ibid.

4. Može li umjetna inteligencija biti svjesna?

Dennett u svojemu radu postavlja sljedeća dva pitanja: »Može li robot biti svjestan?«[198] i »Kako bi ta svijest izgledala?«[199] Dennettov je stav da se odgovori na ova pitanja moraju temeljiti na znanstvenim istraživanjima[200], a kako bi ih mogli temeljiti na njima, »nužno je prihvatiti naturalističku pretpostavku da je svaki um, uključujući ljudski um, u konačnici rezultat mnoštva jednostavnih, prirodnih i posve bezumnih procesa.«[201] Nasuprot tome, »alternativna, nadnaturalistička pretpostavka—prema kojoj je um nešto posebno, nešto izvan reda prirode, nešto neobjašnjivo prirodnim stvarima i procesima—ne samo da je nespojiva sa znanošću, već se u tri stotine i pedeset godina od Descartesa pokazala prilično beznadnom.«[202] Stoga, prihvaćajući naturalističku poziciju trebamo prihvatiti i tvrdnju da smo »potekli […] i sastavljeni od robota«[203] i da »sva intencionalnost koju uživamo proizlazi iz temeljnije intencionalnosti tih milijarda prostih intencionalnih sustava.«[204] Dakle, »svatko od nas je skup više bilijuna makromolekularnih strojeva«[205] i naše postojanje potvrđuje tvrdnju da »nešto što je napravljeno od robota može očitovati pravu svijest«[206].

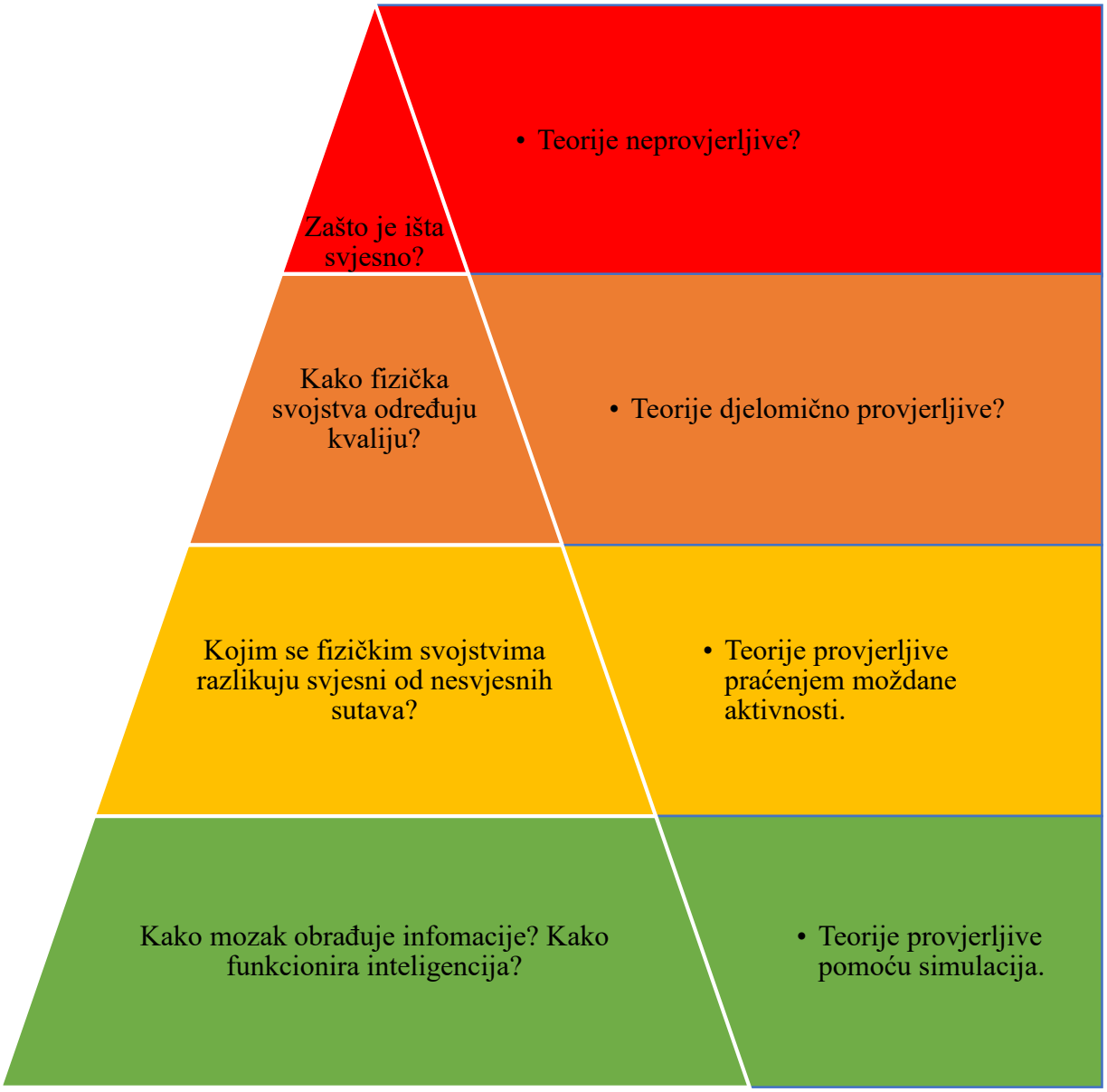
Ideju da je svjestan robot moguć Dennet zastupa u teoriji zvanoj »'model višestrukih nacrta' koja implicira načelnu mogućnost svjesnog robota«[207]. Prema modelu višestrukih nacrta, sve se varijacije percepcije, misli ili mentalne aktivnosti ostvaruju u mozgu paralelno s višestrukim procesima interpretacije i razrade senzornih podataka, tj. sve su informacije koje ulaze u živčani sustav pod neprestanom »uredničkom revizijom«.[208] Iz navedenoga je vidljivo da je u toj teoriji zastupljena tvrdnja, koju Dennett izriče u djelu *Consciousness Explained*, da je naš mozak sustav obrade informacija.[209] Naime, to je sustav koji konstantno preispituje i uređuje informacije od kojih samo neke dospiju u svijest.[210] U toj se teoriji, nadalje, odbacuje

[198] Gregorić, »Pogovor hrvatskom izdanju«, p. 155.
[199] Ibid.
[200] Ibid.
[201] Ibid.
[202] Ibid.
[203] Ibid., *Vrste umova: k razumijevanju svijesti*, p. 54.
[204] Ibid.
[205] Ibid., p. 27.
[206] Ibid., p. 27–28.
[207] Ibid., p. 20.
[208] Daniel C. Dennett, *Consciousness Explained* (New York: Back Bay Books, 1991), p. 111: »According to the Multiple Drafts model, all varieties of perception – indeed, all varieties of thought or mental activity – are accomplished in the brain by parallel, multitrack processes of interpretation and elaboration of sensory inputs. Information entering the nervous system is under continuous 'editorial revision.'«
[209] Ibid., p. 433: By thinking of our brains as information-processing systems […]
[210] Ibid., p. 113: »It is always an open question whether any particular content thus discriminated will eventually appear as an element in conscious experience […]«

važnost kvalije, tj. kvalija biva zamijenjena složenim dispozicijskim stanjima mozga[211], a razumijevanje, koje ima ključnu ulogu u bivanju svjesnim, jest svojstvo koje proizlazi iz mnoštva distribuiranih kvazi-razumijevanja u jednome velikom sustavu.[212] Taj se sustav ne može predstaviti na Searleov način koji ga predstavlja kao neki program s jednostavnom arhitekturom jer takav program u stvarnosti nikada ne bi ni mogao proizvesti rezultate za prolazak Turingova testa.[213] Naime, kompleksnost sustava jest bitna.[214] Stoga bi, prema Dennettuovu mišljenju, na temelju iznesenoga, jedan sustav s raznim podsustavima mogao postići svijest.

Max Tegmark nudi odgovor sličan Dennettovu jer također smatra da je kompleksnost sustava bitna, no za razliku od njega, Tegmark ne opovrgava kvaliju. Prema Tegmarkovu stavu, da bi dobili odgovor na pitanje može li umjetna inteligencija biti svjesna, prvo moramo odgovoriti na hijerarhijski niz pitanja koja predstavlja pomoću piramidalne sheme (slika 4). Prvo i najjednostavnije pitanje nastoji odgovoriti na problem kako mozak procesuira informacije.[215] Drugo i nešto teže pitanje razmatra kojim se fizičkim svojstvima razlikuju svjesni od nesvjesnih sustava.[216] Treće pitanje problematizira kako fizička svojstva određuju kvaliju, a posljednjim se i najtežim pitanjem razmatra zašto je uopće išta svjesno.[217]

[211] Ibid., p. 431: »'Qualia' have been replaced by complex dispositional states of the brain […]«
[212] Ibid., p. 439: »They just can't imagine how understanding could be a property that emerges from lots of distributed quasi-understanding in a large system.«
[213] Ibid., »Searle begs the question. He invites us to imagine that the giant program consists of some simple table-lookup architecture that directly matches Chinese character strings to others, as if such a program could stand in, fairly, for any program at all. We have no business imagining such a simple program and assuming that it is the program Searle is simulating, since no such program could produce the sorts of results that would pass the Turing test, as advertised.«
[214] Ibid., p 440: »Complexity does matter.«
[215] Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence*, p. 362: »First, there's the mystery of how a brain processes information, which David calls the 'easy' problems.«
[216] Ibid., p. 363: »What physical properties distinguish conscious & unconscious systems?«
[217] Ibid., »How do physical properties determine qualia? […] Why is anything conscious?«



Slika 4. Hijerarhijski prikaz pitanja o problemu svijesti[218]

Razmatrajući ove probleme, Tegmark se pita kako od nečega tako jednostavnoga poput čestica može nastati nešto tako složeno poput svijesti, a odgovor koji nudi glasi da je svijest fenomen koji ima svojstva iznad i izvan svojih čestica.[219] Naime, on smatra da je svijest emergentna pojava sa svojstvima iznad i izvan vlastitih čestica te tako na primjer ulazak u dubok san preuređuje čestice i gasi svijest.[220] Tegmark se ujedno približava funkcionalističkoj teoriji tvrdeći da je svijest fizički fenomen koji se čini nefizičkim zbog njegove sličnosti s

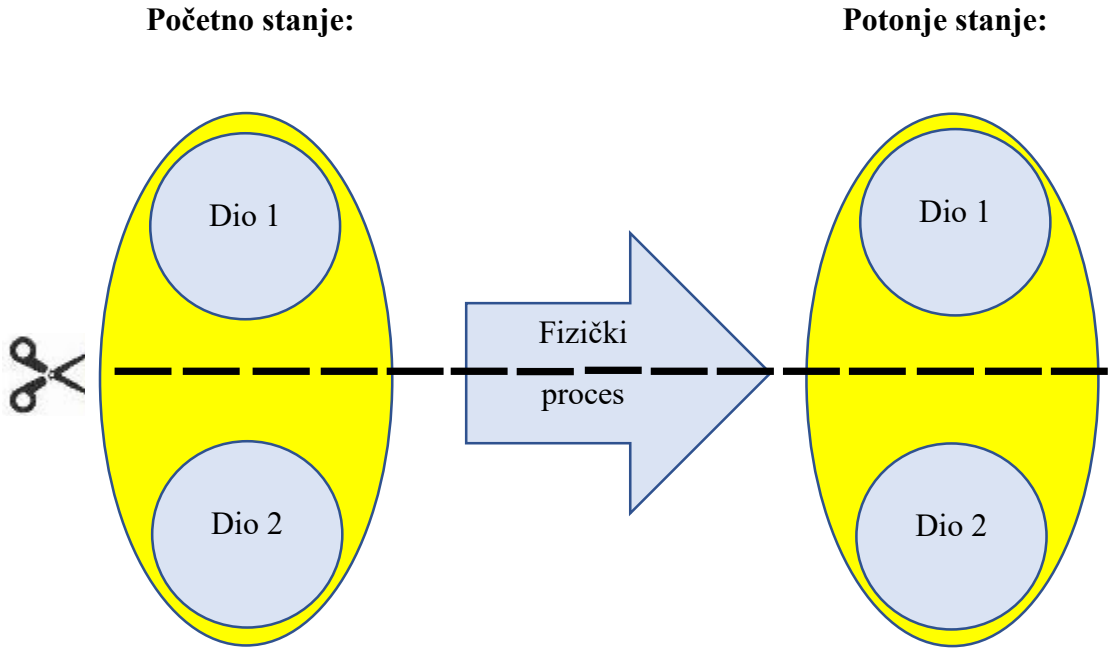
[218] Ibid., vlastiti prijevod.
[219] Ibid., p. 383: How can something as complex as consciousness be made of something as simple as particles? I think it's because it's a phenomenon that has properties above and beyond those of its particles.
[220] Ibid., »[…] I think consciousness is an emergent phenomenon, with properties above and beyond those of its particles. For example, entering deep sleep extinguishes consciousness, by merely rearranging the particles.«

valovima i proračunima, tj. prema njegovu stavu, svijest ima svojstva neovisna o svojemu specifičnom fizičkom supstratu.[221] Navedeno, smatra Tegmark, slijedi iz ideje shvaćanja svijesti kao informacije.[222] Također, poput Colea, Tegmark se priklanja radikalnoj ideji da, ako je svijest način na koji se informacija osjeća kada se obrađuje na određene načine, onda ona mora biti neovisna o supstratu, tj. bitna je samo struktura obrade informacija, a ne struktura materije koja vrši obradu informacije.[223]

Tegmark navodi da se prilikom obrade informacija moraju zadovoljiti razni uvjeti koji jamče svijest i, iako tvrdi da ne zna sve uvjete, navodi da postoje četiri neophodna.[224] Prvi je tzv. princip informacije koji tvrdi da svjesni sustav mora imati značajan kapacitet za pohranu informacija.[225] Drugi je uvjet tzv. princip dinamike koji tvrdi da svjesni sustav mora imati znatan kapacitet za obradu informacija.[226] Treći se uvjet naziva principom neovisnosti koji tvrdi da svjesni sustav mora imati značajnu neovisnost od ostatka svijeta.[227] Posljednji je uvjet tzv. princip integracije koji tvrdi da se svjesni sustav ne može sastojati od gotovo neovisnih dijelova.[228] Dakle, na temelju prvih dvaju principa, da bi sustav bio svjestan, mora biti u stanju obraditi i pohraniti informacije.[229] Taj sustav ujedno, na temelju trećeg principa, mora biti neovisan od ostatka svijeta jer inače ne bi imao subjektivno osjećanje neovisnoga postojanja.[230] Četvrti se princip koji navodi Tegmark, princip integracije, temelji na teoriji zvanoj integrativno informacijska teorija (eng. *integrated information theory*). Prethodno sam naveo da Tegmark smatra da je svijest način na koji se informacija »osjeća« kada se obrađuje na određene složene načine i IIT (skraćenica za *itegrated information theory*), prema Tegmarkovu mišljenju, to potvrđuje i precizira frazu »određeni složeni načini« tvrdeći da obrada informacija

[221] Ibid., p. 389: »In summary, I think that consciousness is a physical phenomenon that feels non-physical because its like waves and computations: it has properties independent of its specific physical substrate.«
[222] Ibid., »This follows logically from the consciousness-as-information idea.«
[223] Ibid., »This leads to a radical idea that I really like: If consciousness is the way that information feels when it's processed in certain ways, then it must be substrate-independent; it's only the structure of the information processing that matters, not the structure of the matter doing the information processing.«
[224] Ibid., »I don't pretend to know what conditions are sufficient to guarantee consciousness, but here are four necessary conditions that I'd bet on [...]«
[225] Ibid., p. 390: »Information principle [...] A conscious system has substantial information-storage capacity.«
[226] Ibid., »Dynamics principle [...] A conscious system has substantial information-processing capacity.«
[227] Ibid., »Independence principle [...] A conscious system has substantial independence from the rest of the world.«
[228] Ibid., »Integration principle [...] A conscious system cannot consist of nearly independent parts.«
[229] Ibid., »This means that to be conscious, a system needs to be able to store and process information, implying the first two principles.«
[230] Ibid., »I think that a conscious system also needs to be fairly independent from the rest of the world, because otherwise it wouldn't subjectively feel that it had any independent existence whatsoever.«

mora biti integrirana.[231] Argument je Gulia Tononija, predstavnika IIT-a, da svjesni sustav treba biti integriran u jedinstvenu cjelinu, jer kada bi se umjesto toga sustav sastojao od dvaju nezavisnih dijelova, onda bi se oni osjećali kao dva svjesna entiteta, a ne kao jedan.[232] Drukčije rečeno, ako svjesni dio mozga ili računala ne može komunicirati s ostatkom sustava, onda ostatak ne može biti dio njegova subjektivnog iskustva.[233]



Φ mjeri nemogućnost podjele procesa na nezavisne dijelove

Slika 5. Prikaz nemogućnosti podjele sustava i procesa u nezavisne dijelove[234]

Tegmark pomoću grafičkoga prikaza predočava funkcioniranje ITT sustava. U njemu Tegmark prikazuje jedan sustav koji se sastoji od dva manja i međusobno povezana dijela. Početno se stanje sustava, koji s vremenom prolazi kroz određeni fizički proces, transformira u novo stanje, a njegova integrirana informacija Φ mjeri nemogućnost podjele procesa u nezavisne dijelove.[235] Naime, ako buduće stanje svakoga dijela sustava ovisi samo o njegovoj

[231] Ibid., p. 387: »I'd been arguing for decades that consciousness is the way information feels when being processed in certain complex ways. IIT agrees with this and replaces my vague phrase "certain complex ways" by a precise definition: the information processing needs to be integrated […]«
[232] Ibid., »Giulio's argument for this is as powerful as it is simple: the conscious system needs to be integrated into a unified whole, because if it instead consisted of two independent parts, then they'd feel like two separate conscious entities rather than one.«
[233] Ibid., »In other words, if a conscious part of a brain or computer can't communicate with the rest, then the rest can't be part of its subjective experience.«
[234] Ibid., p. 386, vlastiti prijevod.
[235] Ibid., »Given a physical process that, with the passage of time, transforms the initial state of a system into a new state, its integrated information Φ measures inability to split the process into independent parts.«

vlastitoj prošlosti, a ne ujedno i o onome što je radio drugi dio, tada je $\Phi = 0$ i ono što smo nazvali jednim sustavom jesu zapravo dva neovisna sustava koji uopće ne komuniciraju jedan s drugim.[236] Kada je $\Phi = 0$, tada se zapravo želi reći da ne postoji nikakva integracija između dijelova zbog čega manje dijelove ne možemo promatrati kao skup koji čini jednu neovisnu cjelinu i stoga ne postoji jedan svjestan sustav. Dakle, prva tri principa svijesti podrazumijevaju autonomiju, tj. da je sustav sposoban zadržati i obraditi informacije bez pretjeranoga vanjskog uplitanja, čime određuje svoju budućnost.[237] Ako uz navedena tri principa uvrstimo i četvrti, onda sva četiri principa zajedno znače da je sustav autonoman, no da njegovi pojedinačni dijelovi nisu.[238]

Prihvaćanjem IIT-a komplicira se mogućnost imanja svjesnih robota jer IIT tvrdi da današnje računalne arhitekture ne mogu biti svjesne budući da način na koji su njihova logična vrata povezana daje vrlo nisku razinu integracije.[239] Na primjer, kada bi čovjeka učitali u budućega robota koji precizno simulira svaki neuron i sinapsu i koji izgleda, govori i djeluje nerazlučivo od toga čovjeka, prema Toninijevu stavu dobili bismo nesvjesnoga zombija koji je bez subjektivnoga iskustva.[240] Iako sama zamjena ne bi utjecala na ponašanje čovjeka budući da je simulacija, po pretpostavci, savršena, čovjekovo bi se iskustvo, prema Toninijevu mišljenju, promijenilo iz svjesnoga u nesvjesno.[241]

Nadalje, kada bi neki budući sustav umjetne inteligencije bio svjestan, upitno je što bi on subjektivno doživio.[242] To je dio »još težeg problema« svijesti i trenutno ne samo da nedostaje teorija koja bi odgovorila na to pitanje nego nije ni sigurno može li se uopće u potpunosti odgovoriti na to pitanje.[243] No, Tegmark pretpostavlja, kao što sam prethodno naveo u poglavlju o virtualnim osobama i funkcionalizmu, da bi umjetna inteligencija mogla imati puno više iskustvenih doživljaja od ljudi.[244] Treba uzeti u obzir da bi umjetna svijest veličine

[236] Ibid., »If the future state of each part depends only on its own past, not on what the other part has been doing, then $\Phi = 0$: what we called one system is really two independent systems that don't communicate with each other at all.«
[237] Ibid., p. 390: »The first three principles imply autonomy: that the system is able to retain and process information without much outside interference, hence determining its own future.«
[238] Ibid., »All four principles together mean that a system is autonomous but its parts aren't.«
[239] Ibid., p. 391: »Another controversial IIT claim is that today's computer architectures can't be conscious because the way their logic gates connect gives very low integration.«
[240] Ibid., »[...] if you upload yourself into a future high-powered robot that accurately simulates every single one of your neurons and synapses, then even if this digital clone looks, talks and acts indistinguishably from you, Giulio claims that it will be an unconscious zombie without subjective experience [...]«
[241] Ibid., p. 392: »Although your behavior would be unaffected by the replacement since the simulation is by assumption perfect, your experience would change from conscious initially to unconscious at the end, according to Giulio.«
[242] Ibid., p. 394: »If some future AI system is conscious, then what will it subjectively experience?«
[243] Ibid., »This is the essence of the "even harder problem" of consciousness [...] Not only do we currently lack a theory that answers this question, but we're not even sure whether it's logically possible to fully answer it
[244] Ibid., »[...] the space of possible AI experiences is huge compared to what we humans can experience.«

mozga mogla imati milijun puta više iskustava u sekundi od nas budući da elektromagnetski signali putuju brzinom svjetlosti, tj. drastično brže od neuronskih signala.[245] Marvin Minsky u svome radu »Why People Think Computers Can't« također tvrdi da bi ta umjetna bića mogla imati bogatiji unutarnji život od ljudi.[246] No, trenutno definitivnoga odgovora o mogućnostima svjesne umjetne inteligencija nema, a ono što je moguće ustvrditi jest da umjetna inteligencija može »pokazivati« inteligentno mišljenje. Međutim, uvijek moramo imati na umu suprotstavljenost između sposobnosti inteligentnoga mišljenja i sposobnosti subjektivnoga doživljavanja kvalije.[247]

4.1. Želimo li uopće svjesnu umjetnu inteligenciju?

Norbert Wiener uvidio je da umjetna inteligencija neće samo oponašati i zamijeniti ljudska bića u mnogim inteligentnim aktivnostima, već da će promijeniti i same ljude u tome procesu.[248] Wiener je također predvidio probleme koje su Turing i drugi optimisti uglavnom zanemarili.[249] Naime, prema njegovu mišljenju prava je opasnost da takve strojeve, koji su sami po sebi bespomoćni, koristi ljudsko biće ili skupina ljudskih bića kako bi povećali svoju kontrolu nad ostatkom populacije.[250] Bespomoćnost i lakoća korištenja, o kojoj govori Wiener, karakteristična je za trenutnu vrstu umjetne inteligencije. Umjetna inteligencija u svojim trenutnim manifestacijama parazitira na ljudskoj inteligenciji i poprilično neselektivno guta sve što su proizveli ljudski tvorci izvlačeći uzorke koji se tu mogu pronaći.[251] Ti strojevi još uvijek nemaju vlastito uspostavljene ciljeve, strategije i kapacitet za samokritičnost i inovacije koji bi im dopustili da nadvladaju svoja ograničenja i refleksivno razmišljaju o vlastitome razmišljanju i ciljevima i stoga oni jesu, kako kaže Wiener, bespomoćni. Ta bespomoćnost nije u smislu da su oni okovani agenti, već u smislu da uopće nisu agenti, tj. oni ne posjeduju sposobnost, kako

[245] Ibid., »[...] a brain-sized artificial consciousness could have millions of times more experiences than us per second, since electromagnetic signals travel at the speed of light—millions of times faster than neuron signals.«
[246] Minsky, »Why People Think Computers Can't«, p. 15b: »[...] those artificial creatures might have richer inner lives than people do.«
[247] Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence*, p. 399: »[...] this distinction, by contrasting sapience (the ability to think intelligently) with sentience (the ability to subjectively experience qualia).«
[248] Daniel C. Dennett, »What Can We Do?«, u: *Possible Minds: Twenty-Five Ways of Looking at AI*, urednik: John Brockman (New York: Penguin Press, 2019), pp. 41–53, na p. 43 : »Wiener saw farther and deeper, recognizing that AI would not just imitate—and replace—human beings in many intelligent activities but change human beings in the process.«
[249] Ibid., p. 44: »Wiener foresaw the problems that Turing and the other optimists have largely overlooked.«
[250] Ibid., »The real danger, he said, is that such machines, though helpless by themselves, may be used by a human being or a block of human beings to increase their control over the rest of the race [...]«
[251] Ibid., p. 48: »AI in its current manifestations is parasitic on human intelligence. It quite indiscriminately gorges on whatever has been produced by human creators and extracts the patterns to be found there [...]«

je Kant izrazio, biti potaknuti razlozima koji su im predstavljeni.[252] No, pitanje je želimo li uopće kreirati svjesne agente koji posjeduju sposobnost biti potaknuti predstavljenim razlozima, tj. koji su agenti u punome smislu.

Složio bih se s Dennettovom tvrdnjom da nam ne trebaju umjetni svjesni agenti.[253] Naime, mi trebamo inteligentne alate koji nemaju prava niti imaju osjećaje koji bi mogli biti povrijeđeni.[254] Jedan od razloga, prema Dennetovu mišljenju, zbog kojega ne bismo trebali stvoriti umjetne svjesne agente jest taj što ne bi mogli podijeliti s nama našu ranjivost ili smrtnost.[255] To je antropocentrično gledište, ali za umjetnu inteligenciju ne bi vrijedili isti zakoni bivanja te, iako ona u teoriji može vršiti iste stvari poput čovjeka, kao na primjer razmnožavanje, načini se manifestiranja postojanja u svojoj srži iznimno razlikuju. Govoreći o razmnožavanju, iako ono nije ključno za bivanje osobom, umjetna bi se inteligencija mogla razmnožavati pomoću samoreplikacije koja »nije nešto što je teorijski strano računalu«[256], a nju je i »matematički dokazao John von Neumann.«[257]

Nadalje, stvaranjem bi svjesnih umjetnih agenata zapali u problem kako postupati s njima u pravnome i moralnom smislu. Primjer toga nudi Dennett u radu »What Can We Do« gdje piše o svjesnome robotu koji prekrši potpisani dogovor. U tome primjeru trebamo zamisliti robota koji ima i zaslužuje legalni status kao moralno odgovorni subjekt.[258] Potrebno je izdvojiti da takav status nemaju ni svi ljudi, tj. nemaju ga djeca koja ne mogu potpisati takve ugovore niti one osobe s invaliditetom čiji pravni status zahtijeva da se o njima brinu i za njih odgovaraju skrbnici.[259] Dakle, taj je robot na razini da razumije moralne i pravne principe, ali je problem u tome što je robot koji želi postići takav uzvišeni status previše neranjiv da bi mogao dati vjerodostojno obećanje.[260] Ono što je nejasno jest koja bi bila prigodna kazna za kršenje obećanja jer bivanje zaključanim u ćeliji jedva da predstavlja neugodnost za umjetnu inteligenciju, osim ako prvotno nema ugrađenu kakvu čežnju za lutanjem, a drugi oblik kazne

[252] Ibid., »These machines do not (yet) have the goals or strategies or capacities for self-criticism and innovation to permit them to transcend their databases by reflectively thinking about their own thinking and their own goals. They are, as Wiener says, helpless, not in the sense of being shackled agents or disabled agents but in the sense of not being agents at all—not having the capacity to be "moved by reasons" (as Kant put it) presented to them.«
[253] Ibid., p. 51: »We don't need artificial conscious agents.«
[254] Ibid., »We need intelligent tools. Tools do not have rights, and should not have feelings that could be hurt […]«
[255] Ibid., »One of the reasons for not making artificial conscious agents is that [...] they would not [...] share with us natural conscious agents our vulnerability or our mortality.«
[256] Dennett, *Vrste umova: k razumijevanju svijesti*, p. 25.
[257] Ibid.
[258] Dennett, »What Can We Do?«, p. 51: »[...] having and deserving legal status as a morally responsible agent.«
[259] Ibid. »Small children can't sign such contracts, nor can those disabled people whose legal status requires them to be under the care and responsibility of guardians of one sort or another.«
[260] Ibid., »The problem for robots who might want to attain such an exalted status is that, like Superman, they are too invulnerable to be able to make a credible promise.«

u obliku rastavljanja umjetne inteligencije nije jednak ubojstvu ako sačuvamo pohranjene informacije o njezinu dizajnu i softveru.[261] Ta mogućnost da softver i podaci budu besmrtni uklanja robote iz svijeta ranjivosti.[262]

I stoga, prema Dennettovu mišljenju, ono što stvaramo ne bi trebali biti svjesni, humanoidni subjekti, već potpuno nova vrsta entiteta nalik prorocima koji su bez svijesti, bez straha od smrti, bez ljubavi i mržnje koji ometaju i bez osobnosti.[263] Dovoljno će teško biti naučiti živjeti s njima bez da se ometamo fantazijama u kojima će nas umjetna inteligencija porobiti.[264] Ono što mi trebamo izrađivati jesu alati, a ne kolege, i velika opasnost leži u neuvažavanju te razlike koja se treba neprestano naglašavati, obilježavati i braniti političkim i pravnim inovacijama.[265] U konačnici, razvoj će umjetne inteligencije, neovisno o tome bude li imala ulogu alata ili kolege, izmijeniti naša shvaćanja o svijetu i omogućiti bolje razumijevanje naturalističkoga tumačenja svijeta. Tradicionalno, mi smo kao vrsta često svoju vlastitu vrijednost temeljili na ideji ljudske izuzetnosti, no uspon će nas umjetne inteligencije prisiliti da napustimo stajalište jedinstvenosti i superiornosti te postanemo skromniji.[266]

[261] Ibid., pp. 51–52: »What would be the penalty for promise-breaking? Being locked in a cell or, more plausibly, dismantled? Being locked up is barely an inconvenience for an AI unless we first install artificial wanderlust that cannot be ignored or disabled by the AI on its own (and it would be systematically difficult to make this a foolproof solution, given the presumed cunning and self-knowledge of the AI); and dismantling an AI (either a robot or a bedridden agent like Watson) is not killing it, if the information stored in its design and software is preserved.«

[262] Ibid., p. 52: [...] the breakthrough that permits software and data to be, in effect, immortal—removes robots from the world of the vulnerable [...]«

[263] Ibid., »So what we are creating are not—should not be—conscious, humanoid agents but an entirely new sort of entities, rather like oracles, with no conscience, no fear of death, no distracting loves and hates, no personality [...]«

[264] Ibid., »It will be hard enough learning to live with them without distracting ourselves with fantasies about the Singularity in which these AIs will enslave us, literally.«

[265] Ibid., p. 46: »As I have been arguing recently, we're making tools, not colleagues, and the great danger is not appreciating the difference, which we should strive to accentuate, marking and defending it with political and legal innovations.«

[266] Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence*, p. 398: »Traditionally, we humans have often founded our self-worth on the idea of human exceptionalism: the conviction that we're the smartest entities on the planet and therefore unique and superior. The rise of AI will force us to abandon this and become more humble.«

5. Zaključak

Pitanje može li umjetna inteligencija posjedovati osobni identitet i biti osobom usko je vezano uz pitanje može li umjetna inteligencija biti svjesna. D. C. Dennett izražava optimizam smatrajući da je moguće razviti svjesnu umjetnu inteligenciju, no on ujedno upozorava da bi stvaranje svjesnoga entiteta moglo prouzrokovati probleme. Naravno, mogućnost stvaranja svjesnih inteligentnih entiteta, ako je to zbilja moguće, nije nešto što će biti ubrzo ostvarivo. Na taj nas zaključak navodi i ITT koji tvrdi da je trenutna arhitektura nedovoljna za razvijanje svjesne umjetne inteligencije zbog niske integrativnosti sustava.

Umjetna inteligencija, da bi posjedovala osobni identitet i bila osobom, osim visoke razine integrativnosti, morala bi zadovoljiti i niz drugih uvjeta. Naime, nužni su uvjeti koje bi umjetna inteligencija morala zadovoljiti kako bi se uopće razmotrio njezin status osobe: da bude racionalno biće, da posjeduje sposobnost recipročnosti, da je sposobna verbalno komunicirati i da je svjesna na jedan specifičan način, tj. da minimalno bude intencionalni sustav drugoga reda. Intencionalni sustav drugoga reda, kao što je prethodno navedeno u radu, raspolaže mogućnošću reflektiranja o vlastitim mislima, tj. raspolaže mogućnošću imanja vjerovanja i želja o vlastitim i tuđim vjerovanjima i željama. U Coleovom primjeru korejske sobe ostaje nejasnim kako virtualne osobe zadovoljavaju uvjete bivanja osobom i kriterije za bivanjem intencionalnim sustavom drugoga reda. Ipak, u svojoj kritici Searleove tvrdnje da računala ne mogu posjedovati razumijevanje, Cole ispravno utvrđuje da razumijevanje ne možemo pripisati samome računalu, već onome što se njime realizira. Tako na primjer ne možemo pripisati ni samo razumijevanje našemu tijelu, već onome što se pomoću tijela ostvaruje. Također, Searleov primjer kineske sobe kritizira i Dennett koji tvrdi da je Searle previše pojednostavio funkcioniranje umjetne inteligencije i da takva vrsta programiranoga djelovanja koju Searle opisuje uistinu nikada ne bi mogla proizvesti razumijevanje, a ponajmanje svijest. Kako bi umjetna inteligencije mogla posjedovati razumijevanje, bilo bi potrebno da bude iznimno kompleksan sustav, a Searleov »program« iz kineske sobe to nije.

Umjetna je inteligencija trenutno dovoljno dobra za prikupljanje podataka, uviđanje obrazaca i nuđenje rješenja za određeni uski skup problema, no ona trenutno ne može, kao što je pokazala Dennettova teorija intencionalnih sustava, raspolagati predstavljenim znanjem. Drukčije rečeno, ona isključivo posjeduje proceduralno znanje. Ostaje za vidjeti hoće li umjetna inteligencija ikada dospjeti dalje od rješavanja uskoga skupa zadataka, ali prije dostizanja toga cilja, potrebno je odgovoriti na niz pitanja o tome kako se procesuiraju

informacije, kako fizička svojstva omogućuju svijest, zašto je uopće išta svjesno i je li bitno od čega izgrađujemo um. Na sva je ta pitanja potrebno odgovoriti prije nego što se dođe do konačnoga rješenja o tome je li moguće da umjetna inteligencija bude osobom, a onda će nam, ako odgovor bude afirmativan, još preostati etička dilema treba li se dopustiti stvaranje inteligentnih i svjesnih umjetnih bića.

6. Popis literature

Bauer, Patricia. »I, Robot«, u *Encyclopedia Britannica*, dostupno na: https://www.britannica.com/topic/I-Robot#ref341291 (pristupljeno 14.02.2022.).

Bloch, Laurent. 2016. »Informatics in the light of some Leibniz's works«, (2016), dostupno na: https://www.researchgate.net/publication/311707999_Informatics_in_the_light_of_some_Leibniz's_works/link/5856641f08aeff086bfbb3d2/download.

Bošnjak, Branko. 1996. »Predgovor«, u: Aristotel, *O duši*, preveo Milivoj Sironić (Zagreb: Naprijed, 1996), pp. VII–XLII, na p. XIII.

Buchanan, Bruce G. 2005. »A (Very) Brief History of Artificial Intelligence«, *AI Magazine*, Volume 26, Number 4 (2005), dostupno na: https://doi.org/10.1609/aimag.v26i4.1848 (pristupljeno 14.02.2022.), pp. 53a–60c

Bush, Vannevar. »As We May Think«, u: *The Atlantic*, dostupno na: https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/ (pristupljeno 14.02. 2022.).

Cole, David. 1991. »Artificial Intelligence and Personal Identity«, u: *Synthese*, Volume 88, Number 3 (1991), dostupno na: https://history.as.uky.edu/sites/default/files/Artificial%20Intelligence%20and%20Personal%20Identity%20-%20David%20Cole.pdf (pristupljeno 14.02. 2022.), pp. 399–417.

Čuljak, Zvonimir. 2012. Natuknica »naturalizam«, u: *Filozofski leksikon*, glavni urednik: Stipe Kutleša (Zagreb: Leksikografski zavod Miroslav Krleža, 2012), pp. 797a–797b.

Daniel C. Dennett. 1976. »Conditions of Personhood«, u: *The Identities of Persons* (University of California Press, 1976), dostupno na: https://philpapers.org/rec/DENCOP (pristupljeno 20.02.2022.), pp. 175–196.

Dennett, Daniel C. 1991. *Consciousness Explained* (New York: Back Bay Books, 1991).

Dennett, Daniel C. 2017. *Vrste umova: k razumijevanju svijesti*, s engleskog preveo Ivan Kraljević (Zagreb: In.Tri, 2017).

Dennett, Daniel C. 2019. »What Can We Do?«, u: *Possible Minds: Twenty-Five Ways of Looking at AI*, urednik: John Brockman (New York: Penguin Press, 2019), pp. 41–53.

»Competitive programming with AlphaCode« (pristupljeno 28.03.2022.), dostupno na: https://deepmind.com/blog/article/Competitive-programming-with-AlphaCode.

Descartes, René. 1993. *Razmišljanja o prvoj filozofiji, u kojima se dokazuje Božja opstojnost i razlika između ljudske duše i tijela*, s latinskog preveo Tomislav Ladan (Zagreb: Demetra. Filozofska biblioteka Dimitrija Savića, 1993).

Delipetrev, Blagoj, Tsinaraki, Chrisa, Kostić, Uroš. 2020. *Historical Evolution of Artificial Intelligence* (Luxembourg: Publication Office of the European Union, 2020), dostupno na: https://publications.jrc.ec.europa.eu/repository/bitstream/JRC120469/jrc120469_historical_evolution_of_ai-v1.1.pdf (pristupljeno 14.02.2022).

Gregorić, Pavel. 2017. »Pogovor hrvatskom izdanju«, u: Daniel C. Dennett, *Vrste umova: k razumijevanju svijesti* (Zagreb: In.Tri, 2017), p. 151–162.

Lemarchand, Guillermo A. 1992. »Detectability of Extraterrestrial Technological Activities«, u: SETIQuest, Volume 1, Number 1 (1992), dostupno na: https://www.researchgate.net/publication/314151939_Dectectability_of_extraterrestrial_technological_activities (pristupljeno 14.02.2022.), p. 3–13.

Lee, Sung-Ha, Yoon, Seok-Hwan, Jung, Yeonjae, Kim, Namil, Uigi Min, Jongsik Chun, i Choi, Incheol. 2020. »Emotional well-being and gut microbiome profiles by enterotype«, u: *Scientific Reports* (2020), dostupno na: https://www.nature.com/articles/s41598-020-77673-z (pristupljeno 17.02.2022.).

Lundström, Jenny Erikson, i Karlsson, Stefan. 2006. »Approaching Artificial Intelligence for Games – the Turing Test revisited«, u: *TripleC*, Volume 4, Number 2 (2006), dostupno na: https://www.triple-c.at/index.php/tripleC/article/download/32/32 (pristupljeno 14.02.2022.), pp. 167–171.

McGuire, Brian. 2006. »The Turing Test«, u: *The History of Artificial Intelligence* (University of Washington, 2006), pp. 5–6.

Minsky, Marvin. 1982. »Why People Think Computers Can't«, u: *AI Magazine*, Volume 3, Number 4 (1982), dostupno na: https://doi.org/10.1609/aimag.v3i4.376 (pristupljeno 27.03.2022.), pp. 3a–15b

Noonan, Harlod W. 2005. *Personal Identity* (London: Taylor & Francis Group, 2005).

Oppy, Graham i Dowe, Dawid. »The Turing Test«, u: *The Stanford Encyclopedia of Philosophy*, dostupno na: https://plato.stanford.edu/entries/turing-test/ (pristupljeno 14.02.2022).

Pruss, Alexander R. 2009. »Artificial Intelligence and Personal Identity«, u: *Faith and Philosophy: Journal of Society of Christian Philosophers*, Volume 26, Iss 5, Article 2 (2009), dostupno na: https://place.asburyseminary.edu/faithandphilosophy/vol26/iss5/2/ (pristupljeno 20.02.2022), pp. 487–500.

Searle, John R. 1980. »Minds, brains, and programs«, u: *Behavioral and Brain Sciences 3* (Cambridge University Press, 1980), dostupno na: https://www.law.upenn.edu/live/files/3413-searle-j-minds-brains-and-programs-1980pdf (pristupljeno 14.02.2022), p. 417a–424b.

Smith, Chris. 2006. »Introduction«, u: *The History of Artificial Intelligence* (University of Washington, 2006), dostupno na: https://courses.cs.washington.edu/courses/csep590/06au/projects/history-ai.pdf?msclkid=e9bf6e5da95711ecbc4cf992730c754b (pristupljeno 14.02.2022), p. 4.

Tegmark, Max. 2017. *Life 3.0: Being Human in the Age of Artificial Intelligence* (New York: Alfred A. Knopf, 2017), dostupno na: https://www.cag.edu.tr/d/l/68e79b19-4dd7-43b0-a578-cdb4308b1881 (pristupljeno 14.02.2022).

Turing, Alan M. 1950. »Computing Machinery and Intelligence«, u: *Mind*, *New series*, Volume 59, Number 236 (1950), dostupno na: https://phil415.pbworks.com/f/TuringComputing.pdf (pristupljeno 14.02.2022.), pp. 433–460.