Filozofski problemi umjetne inteligencije

Saftić, Ivan

SVEUČILIŠTE U RIJECI

FILOZOFSKI FAKULTET

Ivan Saftić

Filozofski problemi umjetne inteligencije

Diplomski rad

Rijeka, rujan 2017.

SVEUČILIŠTE U RIJECI

FILOZOFSKI FAKULTET

Odsjek za filozofiju

Studijske grupe: filozofija i informatika

Student: Ivan Saftić

Matični broj: 17299

JMBAG: 0009057711

Filozofski problemi umjetne inteligencije

Diplomski rad

Mentorica: dr. sc. Majda Trobok

Rijeka, rujan 2017.

„We can only see a short distance ahead, but we can see plenty there that needs to be done.“

(Alan Turing, *Computing Machinery and Intelligence*, str. 460.)

Sažetak

Ovaj je diplomski rad koncipiran kao prikaz Turingovog testa te Searleove i Dreyfusove kritike umjetne inteligencije. Rad sam započeo povijesnim prikazom nastanka područja umjetne inteligencije i najavom najvažnijih filozofskih kritika. U drugom sam poglavlju definirao i opisao klasičnu paradigmu UI. Treće je poglavlje posvećeno Alanu Turingu i njegovoj igri oponašanja. U tom mi je poglavlju cilj bio predstaviti Turingov test te ga obraniti od najvažnijih kritika koje su ga pratile, što je poslužilo i kao svojevrsni uvod u četvrto poglavlje i argument kineske sobe Johna Searla. Searlov argument, kao i najvažnije prigovore na njega, detaljno sam izložio te ukazao na filozofske probleme koje on implicira u shvaćanju umjetne inteligencije. Peto poglavlje posvećeno je kritici klasične teorije umjetne inteligencije Huberta Dreyfusa, a s ciljem drugačijeg pristupa istraživanju inteligencije. Ovdje se rad najviše fokusira na prikaz Dreyfusove alternativne teorije inteligencije kroz radikalnu trostruku tezu da je ljudska inteligencija utjelovljena, da su inteligentna tijela situirana te da je svijet u suštini ljudski. Konačno, u zadnjem sam poglavlju iznio završna razmatranja.

Ključne riječi: umjetna inteligencija (UI), Alan Turing, Turingov test, John Searle, argument kineske sobe, razumijevanje, svijest, Hubert Dreyfus, simbolička UI, utjelovljena inteligencija

Sadržaj

Uvod

Umjetna inteligencija (UI) označava posjedovanje inteligencije od strane strojeva poput računala. Filozofski, glavno je pitanje „Postoji li takvo što?“ Ili, kako je Alan Turing rekao, „Mogu li strojevi misliti?“[1] Ono što to čini filozofskim, a ne samo znanstvenim i tehničkim pitanjem, jest znanstveno neslaganje oko pojma inteligencije, razumijevanja i svijesti te njihovih značaja za umjetnu inteligenciju.

Dodatna komplikacija nastaje ako prihvatimo tezu da su ljudi životinje i da su životinje sami strojevi, kao što to pretpostavlja znanstvena biologija. Ipak, iz definicije strojeva želimo isključiti osobe rođene na uobičajeni način[2] kao i osobe rođene medicinski potpomognutom oplodnjom. U slučaju da ne-ljudske životinje misle i njih želimo isključiti iz definicije strojeva. Konkretnije, treba shvatiti da UI smatra da se inteligencija ili svijest mogu proizvesti umjetnim sredstvima; izgraditi, ne uzgojiti. Radi jasnoće, termin "stroj" koristit ću kako bi označio samo one umjetne. Budući da je trenutni interes za strojeve koncentriran na određenu vrstu stroja; na elektroničko ili digitalno računalo, često ću koristiti samo termin "računalo" kako bi označio stroj s tendencijama umjetne inteligencije.

Umjetna inteligencija (UI) nije samo izmišljeni koncept. Gotovo 70 godina istraživanja o mogućnosti izgradnje inteligentnih strojeva rezultiralo je računalima koja su u stanju pobijediti svjetske prvake u šahu, automatski pronaći i dokazati matematičke teoreme ili razumjeti prirodni jezik, a u najnovije vrijeme i humanoidnim robotima koji mogu hodati i komunicirati s nama. Unatoč ranijim tvrdnjama da su inteligentni strojevi nadomak preuzimanju naše civilizacije, napredak umjetne inteligencije bio je spor i težak. Svijest i okoliš i dalje predstavljaju dva najveća problema s kojima se UI susreće. Kako točno treba izgraditi inteligentni stroj? Treba li on funkcionirati kao ljudski um? Treba li raditi kao mozak? Je li mu potrebno tijelo?

Među znanostima, umjetna inteligencija posebno je privlačila filozofe. S obzirom na to da računala mogu naizgled obavljati određene intelektualne zadatke, postavlja se pitanje: „Razmišljaju li računala zaista?“ A ako stvarno razmišljaju, imaju li svijest? Je li um samo kompliciran računalni program? Kako možemo testirati umjetnu inteligenciju? Hubert

[1] Turing, A., 1950. *Computing Machinery and Intelligence,* Mind, volume 59, No.236, , str. 433.

[2] Ibid. str. 435.

Dreyfus jedan je od prvih koji je kritizirao i uzdrmao temelje umjetne inteligencije. Nešto kasnije, drugi filozof, John Searle, raspravljao je s većinom istraživača UI o tome bi li računalo jednoga dana moglo biti ne samo inteligentno, već i svjesno kao što su to ljudi. Za početak razumijevanja filozofskih problema umjetne inteligencije, treba vidjeti kako i na kojim je temeljima nastala.

1. Povijesni prikaz razvoja umjetne inteligencije

Samo desetak godina nakon izuma prvog programabilnog digitalnog računala,[3] 1956. godine, na konferenciji za novinare na Sveučilištu Dartmouth u New Hampshireu, najavljeno je rođenje novog područja istraživanja pod nazivom „umjetna inteligencija“.[4] Umjetna inteligencija ili UI,[5] bila je opisana kao temeljna znanost koja bi sustavno proučavala fenomen „inteligencije“. Najavljeno je da će znanstvenici proučavati taj fenomen pomoću računala za simulaciju inteligentnih procesa. Polazišna točka bila je pretpostavka kako logičke operacije, koje računala izvršavaju, mogu biti strukturirane tako da oponašaju ljudske misaone procese. U principu, istraživači UI pretpostavili su da je, putem pravilnog programiranja, moguće opskrbiti računala izvornom inteligencijom ljudskih bića na isti način na koji su umjetno proizvedeni dijamanti ipak pravi dijamanti.[6] Budući da se rad računala može u potpunosti shvatiti, dok se rad ljudskog mozga ne može, istraživači UI nadali su se da će na taj način doći do znanstvenog razumijevanja fenomena „inteligencije“.

Ideja umjetne inteligencije počela je okupirati maštu računalnih znanstvenika i matematičara 40-ih i 50-ih godina prošloga stoljeća. Britanski matematičar Alan Turing je, 1936. godine, postavio temelje UI razvojem koncepta univerzalnog stroja poznatog i kao Turingov stroj, koji se općenito smatra teorijskim modelom današnjih računala. Turing je pokazao kako je moguće izumiti stroj koji se može koristiti za izračunavanje bilo kojeg komputacijskog

[3] Colossus je ime za skup računala razvijenih od strane britanskih enkriptičara od 1943. do 1945. godine koji su se rabili za razbijanje njemačke šifre Lorenz SZ 40/42 tijekom Drugog svjetskog rata. Colossus računala koristila su vakuumske cijevi i elektroničke sklopove za obavljanje operacija računanja, binarne aritmetike i Booleove logike. Colossus se time smatra prvim svjetskim programabilnim digitalnim računalom, iako je programirano pomoću sklopki, a ne pomoću pohranjenog programa. Unos podataka vršio se pomoću bušene papirnate vrpce. Colossus i razlog njegove izgradnje bili su strogo čuvana tajna sve do 1975 godine. (Vidi: Randell, B. 1980. *The Colossus,* A History of Computing in the Twentieth Century)

[4] Na Dartmouth Artificial Intelligence (AI) konferenciji 1956. godine rodilo se područje umjetne inteligencije. John McCarthy pozvao je mnoge vodeće znanstvenike toga doba u Dartmouthu u New Hampshireu kako bi razgovarali o jednoj posve novoj temi koju su nazvali - umjetna inteligencija. (Vidi: Rajaraman, V. 2014. *John McCarthy – Father of Artificial Intelligence*, u Asia Pacific Mathematics Newsletter, Vol.4 No3.)

[5] Skraćenica po kojoj je to polje istraživanja ubrzo postalo poznato i kojom ću se u nastavku ovog rada koristiti. (engl. *Artificial Intelligence – AI)*

[6] Sintetički dijamanti stvoreni su u komorama visokog tlaka i temperatura koje reproduciraju uvjete u zemljinoj kori. Rezultat su atomi ugljika poredani u strukturu dijamant-kristala. Ljudi često pogrešno pretpostavljaju da sintetički dijamanti nisu pravi dijamanti. Međutim, to nije točno. Iako, naravno, sintetički dijamanti nisu formirani prirodnim putem, oni imaju isti kemijski sastav i fizikalna svojstva kao i prirodni dijamanti. Dakle, sintetički dijamanti nisu lažni, a njihovo umjetno porijeklo nema nikakve veze s činjenicom da su pravi dijamanti u smislu strukture. (Vidi: Perron, C. 2016. *Can You Tell Which Diamonds Are Lab Grown*)

procesa ili se njime može riješiti bilo koji algoritam. Turingov stroj može se zamisliti kao jednostavan uređaj koji čita i piše simbole na beskonačno dugoj traci, u skladu s unaprijed određenim skupom dobro definiranih pravila. Traka je podijeljena na particije u koje stane po jedan simbol, a početna particija sadrži ulazni simbol. Uređaj se može pomicati lijevo ili desno po traci, čitati, brisati, može upisivati simbole te mijenjati unutarnja stanja slijedeći konačan skup uputa. Turingov je stroj apstraktni prethodnik današnjeg računala. Turingova ideja isprepletala se sa sličnom definicijom koju je iste godine iznio američki matematičar i logičar Alonzo Church, koji radom na lambda računu postavlja tezu da svaki proces za koji postoji odrediva procedura može biti iznijet kroz seriju operacija. Prema ovim idejama nastaje poznata Church-Turingova teza koja pokazuje da se Turingovim strojem može ostvariti svaki odredivi proces formalnih operacija. Ona tvrdi da je bilo koji izračun koji je uopće moguć, moguće napraviti algoritmom koji se izvršava na računalu, uz neograničene vremenske i prostorne resurse. Ova jednostavna apstrakcija dovoljna je za pokretanje bilo kojeg računalnog programa bez obzira na to koliko je program složen. Church-Turingova teza time postaje glavna teza o prirodi računala.

Još jednim ključnim doprinosom UI smatra se članak *A logical calculus of the ideas immanent in nervous activity* iz 1943. godine, u kojem su McCulloch i Pitts prvi puta iznijeli ideju modeliranja neurona pomoću algoritama. Oni su pokušali shvatiti na koji način mozak proizvodi vrlo složene uzorke koristeći međusobno povezane neurone. Povlačenjem paralele između jednostavnih veza neurona i logičkih operatora, pokazali su kako svaka funkcija propozicijskog računa može biti ostvariva pomoću neke neuronske mreže. Time su dali pojednostavljeni model neurona koji je postao važan doprinos razvoju umjetnih neuronskih mreža koje pokušavaju modelirati glavne značajke bioloških neurona.

1950. godine, Alan Turing je predložio igru oponašanja, svojevrsni test koji bi pokazao mogu li strojevi misliti. Test se sastoji od toga da ispitivač postavlja pitanja računalu i čovjeku, a ako ne može razlikovati njihove odgovore, tada se smatra da je računalo inteligentan stroj. Turing je vjerovao da će jednoga dana računala moći proći ovaj test, a njegovom se interesu i optimizmu pridružilo mnogo suvremenih znanstvenika. Time je Turingov test postao temelj za razvoj UI, a Turing otac računalne znanosti i UI.

Moguće je navesti još nekoliko pionirskih radova ove discipline. Iste je godine Claude Shannon opisao kako programirati računalo za igranje šaha, a 1956. godine, Alan Newell, Herbert Simon i Cliff Shaw razvili su prvi UI program (*Logic Theorist*) koji je uspio dokazati

38 od prvih 52 teorema Russellove i Whiteheadove *Principie Mathematice*. Marvin Minsky i John McCarthy su tada pozvali mnoge vodeće matematičare i znanstvenika na ljetnu radionicu UI, gdje je službeno najavljeno rođenje novog područja istraživanja.

Od samog početka, UI je bila polje istraživanja s visokim ciljevima i obećanjima. Najviši cilj bio je ni više ni manje nego izgraditi računalni sustav s inteligencijskim i razumnim sposobnostima odraslog ljudskog bića. Mnogi istraživači UI tvrdili su da će taj cilj biti postignut u roku od samo nekoliko desetljeća, zahvaljujući izumu digitalnog računala i ključnim otkrićima u području teorije informacija i formalne logike. 1965. godine, poznati istraživač UI, Herbert Simon, predvidio je da će do 1985. godine računala biti u mogućnosti izvršiti bilo koji zadatak koji ljudsko biće može izvršiti.[7] Marvin Minsky, jednako poznati istraživač UI, 1967. godine, predvidio je da će se svi važni ciljevi umjetne inteligencije moći ostvariti unutar jedne generacije.[8]

Lako je shvatiti zašto su, u to vrijeme, takva predviđanja uzeta ozbiljno s obzirom na naizgled neograničene mogućnosti koje je računalo moglo ponuditi. Osim toga, niz ranih uspjeha istraživača UI pomoglo je opravdati ove ambiciozne tvrdnje. UI je postigla prvu pobjedu već 1956. godine, u svojoj prvoj službenoj godini postojanja – računalni program koji može igrati šah na početničkoj razini[9] – a šah programi poboljšavani su gotovo svake godine nakon toga. Uskoro su uslijedili i ostali uspjesi. Godine 1964., program pod nazivom STUDENT bio je u stanju protumačiti, razumjeti i riješiti kratke tekstualne odlomke koji sadrže probleme algebre,[10] a dvije godine kasnije, program ELIZA djelovao je poput ljudskog psihologa i bio je u mogućnosti s ljudima provesti skroman terapijski dijalog o njihovim osobnim problemima.[11] Financijske agencije to su uzele u obzir, uključujući i Ministarstvo obrane SAD-a. Novi mladi istraživači pohrlili su u novu znanost. Ovo je prouzročilo ogroman

[7] Simon, H. A. 1965. *The Shape of Automation for Men and Management*, Harper & Row.

[8] Crevier, D. 1993. *AI: The Tumultuous Search for Artificial Intelligence*, New York, NY: BasicBooks, str.109.

[9] 1956. godine MANIAC, razvijen u Los Alamos Scientific Laboratory, postao je prvo računalo koje je pobijedilo čovjeka u šahu. Partija je igrana pojednostavljenim pravilima, a računalo je uspjelo poraziti igrača-početnika u 23 poteza. (Vidi: Douglas, J. R. 1978. *Chess 4.7 versus David Levy*, BYTE. str. 84.)

[10] Norvig, P. 1992. *Paradigms of artificial intelligence programming:case studies in Common Lisp.* San Francisco, California: Morgan Kaufmann Publishers Inc. str. 109–149.

[11] Weizenbaum, J. 1966. *ELIZA—a computer program for the study of natural language communication between man and machine*, Magazine Communications of the ACM, Volume 9 Issue 1, str. 36-45.

napredak za UI, tijekom kojeg se ovo područje afirmiralo kao uzbudljivo i dobro financirano te raširenо među tisućama istraživača UI.

1960-ih godina, kada je UI još uvijek bila novo polje, mladi filozof po imenu Hubert Dreyfus, na neposredan je način uveden je u ovo područje. Dreyfus je u to vrijeme bio docent na Massachusetts Institute of Technology (MIT), gdje je predavao kolegije o filozofskim teorijama znanja i percepcija. Njegovi su mu studenti često govorili kako su teorije koje poučava zastarjele izumom računala te su jednog dana informirali Dreyfusa kako su MIT znanstvenici, pod vodstvom njegovog kolege Marvina Minskog iz MIT Artificial Intelligence Laboratory, na putu prema stvaranju stroja koji bi mogao nezavisno znati i percipirati svijet oko sebe.[12]

Potaknut ovom viješću, Dreyfus je, s bratom Stuartom, koji je tada radio kao računalni specijalist za RAND korporaciju, istaknutu neprofitnu organizaciju za istraživanje tehnologije, započeo raspravljati o računalima i o njihovim mogućnostima. Uz bratovu pomoć, RAND ga je zaposlio kao filozofskog savjetnika za procjenu njihovog novog programa umjetne inteligencije. Program su predvodili Allen Newell i Herbert Simon, koji će kasnije postati poznati po svojim istraživanjima u području umjetne inteligencije. No, u procjeni njihovog istraživanja, Dreyfus je došao do zaključka da, iako je uspješno pokazana sposobnost računala u rješavanju određenih vrsta problema, to nije pružilo nikakav dokaz o fenomenu same inteligencije te da je istraživanje bilo na potpuno pogrešnom putu tražeći da računala simuliraju ljudsku inteligenciju.[13] Njegovo pesimistično izvješće napisano 1964. godine pod nazivom *Alkemija i umjetna inteligencija*[14] Newell i Simon su snažno kritizirali, ali je ono iduće godine svejedno izdano od strane RAND korporacije, usprkos njihovim primjedbama.

Dreyfusov je izvještaj bio prva detaljna kritika umjetne inteligencije koja je objavljena, a gotovo je odmah zauzela središnje mjesto u oštroj raspravi računalnih znanstvenika diljem svijeta. Bila je to njegova prva utjecajna publikacija na tu temu i prva u nizu filozofskih kritika umjetne inteligencije u obliku knjiga i članaka. Dreyfusova najvažnija objava u tom području je svakako knjiga *What Computers Can't Do*, koja mu je donijela svjetsku slavu kao

[12] Dreyfus, H. L. 2007. *Why Heideggerian AI Failed and How Fixing it Would Require Making it More Heideggerian*, Oxford University Press. str. 249.

[13] Ibid.

[14] Dreyfus, H. L. 1965. *Alchemy and Artificial Intelligence*, RAND Corporation.

kritičaru UI. Prvi put objavljena 1972. godine, kasnije je znatno revidirana i objavljena 1992. godine kao *What Computers Still Can't Do*.[15] Još jedan važan doprinos raspravi o umjetnoj inteligenciji svakako je i knjiga *Mind Over Machine* koju je napisao 1986. godine, s bratom Stuartom.

Drugu detaljnu i važnu filozofsku kritiku UI uputio je, 1980. godine, američki filozof John Searle u članku *Minds, Brains and Programs* u kojem je predstavio argument kineske sobe, misaoni eksperiment čiji je cilj da pokaže da računala ne mogu razmišljati. Krajem 1970-ih, kada su računala postala brža i jeftinija, neki istraživači na području UI tvrdili su da njihovi programi mogu razumjeti engleske rečenice koristeći velike baze podataka. Rad jednog od njih, istraživača Rogera Schanka s Yale-a privukao je Searleovu pozornost. Ideja je bila dati računalu priču i postavljati mu pitanja o toj priči. Na primjer, računalo bi moglo dobiti priču o posjetu restoranu, a zatim ispravno odgovoriti na pitanja o tome što se tamo dogodilo, gdje odgovori na pitanja zahtijevaju zaključke i nisu eksplicitno uključeni u priču.[16] Tipična Schankova priča ide ovako: Muškarac ulazi u restoran i naručuje hamburger. Konobarica mu donese prepečeni hamburger. Muškarac odjuri iz restorana bez plaćanja računa. Pitanje: je li pojeo hamburger? Ili drugi scenariji: Muškarac ulazi u restoran i naručuje hamburger. Pri odlasku plati račun i ostavi veliku napojnicu. Pitanje: je li pojeo hamburger? Searleov argument kineske sobe je izvorno predstavljen kao odgovor na tvrdnju da UI programi kao što je Schankov, doslovce razumiju rečenice na koje reagiraju. Searle je smatrao da bi najbolji način testiranja takve tvrdnje bilo simulirati rad računala, odnosno pokrenuti program i onda vidjeti omogućuje li to ljudima razumijevanje nekog jezika koji nisu znali prije pokretanja programa.[17]

Searle je, 1984. godine, objavio i knjigu *Minds, Brains and Science* u kojoj argument kineske sobe zauzima centralni dio, a u siječnju 1990. godine, njegov je argument objavljen u časopisu *Scientific American* te je time postao još popularniji u znanstvenoj zajednici. Argument kineske sobe je najfrekventniji filozofski argument u kognitivnoj znanosti i filozofiji umjetne inteligencije u posljednjih 30-ak godina. Iako je od prve objave postao predmetom mnogih komentara, rasprava i prigovora, Searle ga je nastavio braniti i precizirati u mnogim radovima, časopisima, knjigama i predavanjima sve do danas.

[15] Revizija iz 1992. godine uključuje opsežan uvod u kojemu Dreyfus rezimira zbivanja u području UI unutar dvadeset godina nakon što mu je knjiga *What Computers Can't Do* prvi put objavljena.

[16] Cole, D. 2014. The Chinese Room Argument. *The Stanford Encyclopedia of Philosophy*. Edward N. Zalta (ur.)

[17] Searle, J., 1980. *Minds, Brains and Programs*, Behavioral and Brain Sciences, str. 417-418.

2. Klasična paradigma umjetne inteligencije

Od samog početka postojalo je mnogo različitih vrsta istraživanja UI s različitim ciljevima, metodama i formalizmima. Ipak, od početka istraživanja umjetne inteligencije 1950-ih do početka 1980-ih godina, različite vrste istraživanja UI imale su toliko toga zajedničkog da možemo reći da su stvorile paradigmu, u smislu artikuliranom od strane filozofa znanosti Thomasa Kuhna; zbirku metoda, tehnika, ciljeva, pretpostavki i uzornih primjeraka uspješnog istraživanja koje dijele znanstvenici, a koji zajedno određuju istraživački program. Ova paradigma, koja i dalje karakterizira velik dio istraživanja UI, poznata je pod različitim imenima, ali ja ću je nazivati „simboličkom UI“[18] jer je njezin središnji princip da je inteligencija manipulacija simbolima.

U prvom desetljeću svog postojanja, simbolička UI imala je za cilj izgradnju inteligentnih računalnih sustava. Cilj je bio stvoriti sustav koji posjeduje univerzalnu inteligenciju, to jest, univerzalnu sposobnost razmišljanja, rješavanja problema, razumijevanja jezika i obavljanja drugih inteligentnih zadataka koje inteligentna ljudska odrasla osoba može obavljati. Ovo istraživanje u početku nije bilo usmjereno na razvoj tehničkih aplikacija te je promovirano prije svega kao znanost – nova znanost inteligencije. Neki klasični istraživači UI, uključujući Newella i Simona, postavili su eksplicitan cilj svojeg istraživanja kao modeliranje kognitivnih (misaonih) procesa ljudskih bića; UI s tim ciljem se ponekad naziva „kognitivna simulacija“. U okviru ovog pristupa, smatra se da programi UI simuliraju i objašnjavaju inteligentno ljudsko ponašanje.[19]

Drugi istraživači unutar simboličke UI, uključujući Minskog, nisu se pretvarali da njihovi računalni programi simuliraju ljudske misaone procese, već da njihov rad pruža teorijski doprinos razumijevanju fenomena „inteligencije“ prikazujući opća svojstva inteligentnih procesa. Oni su tvrdili da, iako njihovo istraživanje ne dopušta izravan uvid u psihološku izvedbu inteligentnih zadataka, ono dopušta uvid u izvedbu mjerodavnog inteligentnog ljudskog ponašanja, to jest, pruža opći uvid u kognitivne sposobnosti koje ljudska bića moraju posjedovati za inteligentno ponašanje.[20] Ali razlike između pristupa kognitivne simulacije i

[18] "Simbolička UI" ili ponekad "klasična paradigma umjetne inteligencije"

[19] Newell, A. and Simon, H. A. 1961. *Computer Simulation of Human Thinking*, *Science*, New Series, Vol. 134, No. 3495. str. 2011-2017.

[20] Minsky, M. 1961. *Steps Toward Artificial Intelligence*, Proceedings of the IRE 49(1):8-30, February '61.

ovog uobičajenog pristupa unutar simboličke UI su od manjeg značaja od njihovih točaka podudarnosti; oba pristupa imaju za cilj razumjeti fenomen inteligencije te dijele važne teorijske točke polaska, metode i formalizme.

Simbolička UI pretpostavlja, te kao glavno polazište uzima, da je inteligencija stvar manipulacije simbolima slijeđenjem fiksnih i formalnih pravila. Da bi se došlo do ove ideje potrebno je napraviti niz pretpostavki. Prvo, potrebno je pretpostaviti da su svi inteligentni procesi, uključujući opažanje, zaključivanje, računanje i uporaba jezika, oblici obrade informacija, odnosno unos informacija iz okoline, obrada i manipuliranje tih informacija te pružanje odgovora. Tako kada netko zbraja brojeve, prvo definira brojeve (informacije) koje zbraja, zatim obavlja određenu operaciju nad tim informacijama (u ovom slučaju zbrajanje) te u konačnici pokazuje rješenje. Šah, iako nešto sofisticiraniji, ima istu strukturu: igrač promatra figure i njihov položaj, analizira stanje na šahovnici i određuje koji će potez odigrati. Implikacija takvih primjera je da naizgled inteligentni organizmi i sustavi imaju nešto zajedničko, a to je da su oni sustavi obrade informacija.

Nakon što se napravi ova pretpostavka, prirodno je postaviti sljedeća dva pitanja: koji je karakter ove informacije te kako je ona „obrađena“? U ovom trenutku simbolička UI radi dvije ključne pretpostavke. U odgovoru na prvo pitanje pretpostavlja da informacija mora prvo biti reprezentirana da bi se obradila pomoću sustava obrade informacija. Kako bi upravljao informacijom, sustav mora prvo raditi s medijem u kojem se ta informacija može pohraniti. Takav medij koji pruža informacije o vanjskoj stvarnosti zove se reprezentacija. Poznati primjeri reprezentacija uključuju fotografije, slike, izgovorene i pisane rečenice, ali to nisu vrste reprezentacija koje se mogu koristiti u sustavima obrade informacija. Pretpostavlja se da sustavi za obradu informacija koriste interne reprezentacije dane u obliku koji je prilagođen onome što oni mogu procesuirati. Tako bi i ljudsko razmišljanje trebalo raditi putem sustava unutarnjih mentalnih reprezentacija u kojima su upisane naše misli, percepcije i sjećanja.

Najvažnija pretpostavka simboličke UI je da su unutarnje reprezentacije inteligentnih sustava simbolične naravi. Alternativna mogućnost je da su unutarnje reprezentacije sličnije fotografijama i slikama, da su ikonične, odnosno da sadrže informacije fizičkih sličnosti onome na što se odnose, na način na koji portret prenosi informacije o objektu kojeg reprezentira na temelju svoje sličnosti. No, simbolička UI kao svoju početnu točku uzima pretpostavku da su unutarnje reprezentacije poput riječi u prirodnom jeziku. Jezik je

simbolički; njegovi tokeni su proizvoljni u smislu da oni nemaju ni sličnosti ni inherentnu referencu na ono što predstavljaju. Riječ 'mačka' ne izgleda kao mačka niti ima bilo kakvu intrinzičnu vezu s istom. Pretpostavka da je sve što nosi informaciju u inteligentnim sustavima simbolično uvjerljivija je od pretpostavke da je ikonično. Teško je, na primjer, zamisliti ikonični prikaz apstraktne stvari bez strukture koja se može opaziti. Simbolički prikazi su također mnogo jednostavniji za kombiniranje od onih ikoničnih jer se konačan broj simbola može koristiti u različitim kombinacijama da bi predstavio beskonačnu količinu sadržaja. Prirodan jezik predstavlja takav slučaj; konačan broj riječi može se kombinirati za stvaranje, u načelu, beskonačnog broja rečenica. Također, simboli imaju jasno istaknutu ulogu u vrstama kognitivnih zadataka, kao što su matematički izračuni i logičko zaključivanje, za koje se često vjeruje da predstavljaju najviši oblik inteligencije.

Iz ranije pretpostavke da se inteligencija sastoji u sposobnosti obrade informacija, zajedno s pretpostavkom da se proces obrade informacija sastoji u manipulaciji simbolima, slijedi da su inteligentni sustavi zapravo sustavi obrade simbola. Do sada ništa nije rečeno o tome kako su ti simboli obrađeni. Daljnja je pretpostavka simboličke UI da su simboli obrađeni samo na temelju njihovih formalnih svojstava, odnosno, oblik simbola je u suprotnosti od svog sadržaja ili značenja. Značenje simbola, dakle, ne igra izravnu ulogu u procesu njegove obrade. Dakle, kada računalo obrađuje simbol 'mačka' na određeni način, to čini zbog onoga što prepoznaje u obliku simbola, a ne zato što ima bilo kakav uvid u njegovo značenje.

No kako se određuje koji se procesi provode u sustavu na temelju tih formalnih svojstava? Ovdje susrećemo još jednu ključnu pretpostavku; sustav obrade informacija uključuje pravila prema kojima se ti simboli tumače i obrađuju. To su nužno formalna pravila, jer su ona isključivo povezana s formalnim svojstvima tih simbola. Sustavi rade automatski; kada sustav kao ulaz dobije simbol ili niz simbola, on izvršava određeni proces koji rezultira novim simbolom ili skupom simbola koji se tada još jednom automatski povezuju s drugim pravilom i tako dalje. U nedostatku takvih pravila, inteligencija bi bila misterij, barem iz perspektive simboličke UI, jer ne bi bilo jednostavnog načina za obradu simbola od strane inteligentnih sustava.

Čini se kako podrška pretpostavci da se inteligentna obrada informacija sastoji od primjene pravila, dolazi od uloge koju pravila imaju u zadacima koji zahtijevaju inteligenciju. Tako se čini da razumijevanje jezika uključuje znanje o pravilima gramatike, logičko zaključivanje uključuje primjenu analitičkih pravila, a rješenje problema iz matematike i prirodnih znanosti

uključuje primjenu matematičkih načela i prirodnih zakona. Treba uzeti u obzir pretpostavku da je znanje potrebno za inteligentno ponašanje teoretsko; što znači da znati i razumjeti nešto znači posjedovati apstraktnu simboličku teoriju izraženu pravilima pomoću kojih se razumijeva određeni fenomen.

Ovako skicirana teorija inteligencije može se formulirati i razraditi bez pozivanja na prirodu i na mogućnosti digitalnog računala. Međutim, jasno je da izrada računala čini ovu teoriju inteligencije znatno atraktivnijom. Digitalno računalo je koncipirano kao sustav za obradu informacija koji omogućuje korištenje simboličkih prikaza (nizova nuli i jedinica) i koji procesuira ove simbole prema formalnim pravilima (pomoću definiranih programa). Stoga se čini da postojanje računala nudi priliku za testiranje i istraživanje pretpostavke o simboličkom i pravilima-definiranom karakteru inteligencije te priliku za izgradnju modela inteligentnih procesa na znanstveni način. Gore spomenute pretpostavke o inteligenciji tako nude mogućnost potencijalno plodnog znanstvenog programa istraživanja koji može dovesti do zanimljivih tehničkih primjena. Početni uspjesi simboličke UI u dizajnu inteligentnih računalnih programa dali su ovoj teoriji dodatan legitimitet.

3. Mogu li strojevi misliti? Turingov test i njegove kritike

Alan Turing (1912. - 1954.) bio je britanski matematičar, logičar, kriptograf, teoretski biolog, filozof i maratonac. Izmislio je Turingov stroj, hipotetski stroj koji unatoč svojoj jednostavnosti može simulirati logiku bilo kojeg računalnog algoritma. Tijekom Drugog svjetskog rata radio je u Bletchley Parku gdje je izgradio stroj pomoću kojeg su saveznici mogli odgonetnuti njemačke poruke šifrirane putem *Enigme* i koji je imao veliki utjecaj na završetak rata. Nakon rata, sagradio je prvo računalo i bavio se problemima umjetne inteligencije te se smatra utemeljiteljem modernog računalstva. Poznat po ekscentričnom životu i homoseksualnosti, uhićen je 1952. godine zbog kršenja javnog morala i osuđen na obveznu hormonsku terapiju. Dvije godine kasnije počinio je samoubojstvo. 2014. godine snimljen je i Holivudski film o njemu pod nazivom *The Imitation Game.*

Alan Turing je 1950. godine, u časopisu *Mind*, objavio članak pod nazivom *Computing Machinery and Intelligence* u kojem razmatra pitanje "Mogu li strojevi misliti?" S s obzirom na to da odgovor na postavljeno pitanje iziskuje definicije značenja riječi 'stroj' i 'misliti', umjesto definiranja, Turing je prvotno pitanje zamijenio svojevrsnom igrom koju je nazvao "Igra oponašanja". Igru čine tri osobe, jedan muškarac (A), jedna žena (B) i jedan ispitivač (C) koji može biti bilo kojeg spola. Cilj igre za ispitivača je da utvrdi koja je od dvije osobe muškarac, a koja žena. Ispitivaču (C) nije dopušteno vidjeti A i B te je fizički odvojen od njih kako ne bi mogao donijeti odluku na temelju njihovog fizičkog izgleda. On ih poznaje po oznakama X i Y, a na kraju igre ispitivač zaključuje kako je X A, a Y je B ili kako je X B, a Y je A. Ispitivaču je dopušteno postavljati proizvoljna pitanja. Kako zvukovi glasa ne bi olakšali odluku ispitivaču, odgovori bi trebali biti napisani, ili još bolje, tipkani. Idealni raspored je da postoji komunikacijski uređaj između dvije sobe. Sada se Turing pita što će se dogoditi kada stroj preuzme dio od A u ovoj igri? U ovoj verziji, cilj ispitivača (C) je odrediti koji je od dva entiteta s kojima razgovara čovjek. Hoće li ispitivač griješiti istim postotkom kao što bi griješio i u slučajevima odabira između muškarca i žene? Ta pitanja zamjenjuju izvorno pitanje "Mogu li strojevi misliti?".[21]

[21] (Turing 1950. str. 433-434.)

3.1. Turingov test

"Igra oponašanja" u kojoj stroj ili, u današnjim terminima bolje rečeno, računalo preuzima ulogu jednog od kandidata je takozvani Turingov test[22] te se smatra kako ovaj test predstavlja konačni cilj UI i služi kao sredstvo za odgovaranje na pitanje "Mogu li strojevi misliti?". U TT-u ljudski ispitivač razgovara s dva odvojena entiteta u zaključanim prostorijama: s računalom (A) i s čovjekom (B). Cilj ispitivača (C) je utvrditi koji je entitet čovjek, dok računalo i čovjek pokušavaju uvjeriti ispitivača da su ljudi. Trebali bismo pretpostaviti da je ispitivač dobro pripremljen za provođenje ovog testa. Ako ispitivač (C) ne može prepoznati koji je kandidat čovjek, a koji je računalo nakon niza pitanja, onda se smatra da je računalo uspješno prošlo TT. Dakle, TT može se tumačiti kao test kojim se procjenjuje sposobnost računala da zavara da je čovjek. Turing je smatrao da cilj UI mora biti stvaranje računala koja mogu proći ovaj test, tj. računala koja se jezično ne razlikuju od ljudi.

Turingov je test postao standardni način ispitivanja sadrži li računalo umjetnu inteligenciju. U članku *Computing Machinery and Intelligence* Turing je napisao kako vjeruje da bi do 2000. godine računala mogla biti programirana s kapacitetom pohranjivanja od oko 10^9 koja će moći igrati igru oponašanja toliko dobro da prosječni ispitivač nakon pet minuta ispitivanja neće imati više od 70% šanse da ispravno identificira radi li se o čovjeku ili o računalu.[23] Ovo se i danas smatra osnovnim pravilima TT-a. Iako je Turing napravio prilično točna predviđanja o veličini, memoriji i brzini računala, do sada niti jedno računalo nije uspjelo proći TT, ali je postojalo nekoliko lažnih tvrdnji da jest (poput one o chatbotu Eugenu Goostmanu iz 2014. godine).

Od kada je Turing prvi put opisao svoj test, isti se pokazao vrlo utjecajnim te je postao vrlo važnim i široko kritiziranim konceptom u filozofiji umjetne inteligencije. U nastavku ću opisati niz kritika protiv korištenja TT-a kao testa za umjetnu inteligenciju.

3.2. Kritike Turingovog testa

Neke je kritike jednostavno opovrgnuti dok su druge mnogo složenije i primoravaju nas da

[22] Dalje u tekstu koristit ću kraticu “TT”.
[23] (Turing 1950. str. 442.)

ispitamo suptilnosti TT-a kao i našu inteligenciju. Dok mnoge od tih kritika otvaraju vrlo duboke i zanimljive argumente, mislim da je većina kritika u ovom trenutku zanemariva. Iako ne vjerujem da će TT zauvijek biti najbolji test za umjetnu inteligenciju, moja glavna teza je sljedeća: prije nego što pokušamo promijeniti TT, moramo razumjeti mnogo više ne samo o umjetnoj inteligenciji, već i o ljudskoj inteligenciji, a jedan od najboljih načina da u tome uspijemo jest da razvijamo računala koja bi mogla proći TT kakav je danas.

3.2.1. Argument iz svijesti

Neki ljudi vjeruju da računala moraju biti svjesna kako bi imala um (na primjer da su svjesna svojih postignuća, da osjećaju užitak u uspjehu i uzrujavaju se zbog neuspjeha i tako dalje). Na ekstremu takve tvrdnje nalazi se solipsizam. Jedini način da stvarno znamo misli li računalo jest da smo mi to računalo. Međutim, prema tom stajalištu, jedini način da znamo da drugo ljudsko biće misli (ili je svjesno, sretno i slično) jest da smo mi to ljudsko biće. To se obično naziva problem drugih umova i pojavit će se nekoliko puta u raspravama TT-a i u daljnjem radu. Turing smatra da je to najlogičniji prigovor, ali otežava komunikaciju ideja pa umjesto da se stalno raspravlja o njemu, uobičajena je konvencija da svatko misli.[24] Turing smatra da bi većina onih koji podržavaju argument iz svijesti radije napustila argument nego da bude prisiljena na solipsističku poziciju. Tada bi vjerojatno prihvatili TT.[25] Njegov odgovor na argument iz svijesti je jednostavan, ali moćan: alternativa igri oponašanja (ili sličnim procjenama ponašanja) bila bi solipsizam, ali s obzirom na to da ga ne prakticiramo kod procjene inteligencije drugih ljudi bilo bi fer da ga ne prakticiramo ni kod procjene umjetne inteligencije računala. Turing je vjerovao da se igrom oponašanja može utvrditi je li netko zaista razumio nešto ili je to naučio napamet bez razumijevanja, što se očituje u primjeru razgovora koji daje.[26] Treba također napomenuti kako on ne uzima svijest kao trivijalan ili nepostojeći problem, već samo vjeruje da ne moramo nužno riješiti misterij ljudske svijesti prije nego što možemo odgovoriti na pitanja o razmišljanju, a osobito o mogućnosti razmišljanja u računalima.[27]

3.2.2. Turingov test potiče pogreške i superartikulaciju

Neki ljudi smatraju da TT omogućava "varanje". Na nekim mjestima u članku, Turing opisuje kako računala mogu biti "opremljena" kako bi prevladala određene prepreke. Budući da

[24] Ibid. str. 446.
[25] Ibid. str. 447.
[26] Ibid. str. 446.
[27] Ibid. str. 447.

računalo treba biti praktički nerazlučivo od čovjeka kako bi prošlo TT, mora, poput ljudi, praviti pogreške. Turing je upravo na to upozorio u svom radu gdje je napisao da stroj koji je programiran za TT ne bi uvijek davao prave odgovore na aritmetičke probleme već bi namjerno uvodio pogreške kako bi uvjerio ispitivača da je čovjek.[28] Kada se računalo suoči s aritmetičkom operacijom, kako ne bi izdalo svoj identitet dajući brz i točan odgovor, može se zaustaviti oko 30 sekundi prije nego što reagira i može povremeno dati pogrešan odgovor. Sposobnost brzog i preciznog izvršavanja aritmetičkih izračuna inače se smatra inteligentnim ponašanjem. Ipak, Turing žrtvuje preciznost i brzinu kako bi računalo djelovalo "čovječnije". Neki kritičari smatraju da je to "varanje" i da računalo koristi određene "trikove" u svom radu. Međutim, kada računalo ne bi pravilo pogreške, ispitivač bi mogao postaviti teški aritmetički problem kao svoje prvo pitanje i odmah znati koji je entitet računalo. Vjerujem da je najbolji način gledanja na to kao "obmana", a ne kao "varanje". Uostalom, na neki način, cijeli smisao testa i igre oponašanja je obmana ispitivača.

Turing primarno govori o aritmetičkim problemima, ali njegovu logiku možemo proširiti i na druge trivijalne vrste pogrešaka. Druga vrsta "pogreške", koju Turing vjerojatno nije predvidio, je ta da računalo često daje preartikulirane odgovore na određena pitanja, pogotovo kad uzmemo u obzir da je komunikacija između ispitivača i računala vrsta dopisivanja. Taj problem navodi Donald Michie u svom radu *Turing's Test and Conscious Thought* i naziva ga "superartikulacijom".[29] Michie smatra kako TT potiče računala na obje vrste pogrešaka i da bi test trebalo proglasiti neispravnim ako obvezuje kandidate da prave pogreške kako bi dokazali inteligenciju. Ovu kritiku možemo riješiti na dva načina. Prvo, moguće je da se pogreške pojavljuju kao posljedica strojnog algoritma za inteligenciju. Uostalom, ljudi obično ne čine pogreške namjerno; to se događa kao rezultat unutarnjeg djelovanja našeg mozga (zaboravljamo neke činjenice i tako dalje). Drugo, ako jednog dana budemo u mogućnosti stvoriti računala koja mogu proći TT, ali zbog činjenice da čine premalo grešaka ne prođu test, mislim da bi programerima bilo lako ugraditi nekakav "modul grešaka" sličan ljudima. Izmjena programa na ovaj način ne može naštetiti testu: ako računalo prođe test, može se ponovno programirati kako ne bi činilo pogreške. Svakako, prvo moramo doći do trenutka skorog prolaska TT-a prije no što ovo postane problem.

Pogreška superartikulacije donosi zanimljivu ideju koja se obično ne veže uz kritike TT-a:

[28] Ibid. str. 448.

[29] Michie D., 1999. *Turing's Test and Conscious Thought*, u Millican P. and Clark A. (ur.) *Machines and Thought: The Legacy of Alan Turing, Volume I*, Oxford: Clarendon Press., str. 42.

korištenje računala u testu daje povratne informacije o ljudima. Na putu razvoja inteligentnih računala, superartikulacija se vjerojatno neće pojaviti kao grom iz vedra neba. Vjerojatnije je da ćemo prvo razviti računalo koje je samo malo artikuliranije od nas. Kao rezultat toga, takvo računalo će nam dati uvid u naše intuitivne vještine, a mi ćemo ih moći artikulirati poput računala. Stvar je u tome da će nam gradnja inteligentnih računala pomoći da sami postanemo inteligentniji. Ovo je izvrsna motivacija za provođenje TT-a kao sredstva za učenje o inteligenciji.

3.2.3. Turingov test ne dopušta stupnjevanje inteligencije

Sljedeću kritiku TT-a iznose mnogi filozofi poput R. M. Frencha, L. Hausera i P. H. Millara. Budući da TT traži binarni rezultat, stroj ili je ili nije inteligentan, on ne dopušta ispitivaču određivanje stupnja inteligencije (na primjer tvrdnju: "Ovo računalo je inteligentno kao šestogodišnjak"). Postoje dva rješenja ove kritike, a ovisna su o "načinu rada" umjetne inteligencije. Ako se ispostavi da računala uče na sličan način i sličnom brzinom kao i ljudi, to jest da postoji pojam računala koji je inteligentan kao šestogodišnjak, ali ne i inteligentan kao odrasla osoba, onda je rješenje jednostavno; računalo treba suprotstaviti šestogodišnjaku u TT-u. To ne mijenja prirodu samog testa. Međutim, ako se ispostavi da računala uče drugačije od ljudi, ili barem brže, možda nam takvo stupnjevanje inteligencije nije potrebno. Na primjer, veliku razliku između šestogodišnjaka i odrasle osobe čini njihov vokabular. Dok su šestogodišnjaku potrebne godine kako bi razvio rječnik koji se može usporediti s odraslom osobom, za računalo to možda ne mora biti slučaj. Jednom kad računalo ima osnovne vještine za stjecanje jezika i može naučiti koristiti riječi, može biti nevjerojatno brzo u stjecanju velikog vokabulara, baš kao što je računalu trivijalno pomnožiti velike brojeve nakon što ima sposobnost množenja. Još jednom, nije jasno otkriva li ova kritika bilo kakav problem s TT-om, a mi nemamo načina da to saznamo sve dok ne izgradimo računala mnogo inteligentnija od ovih koje imamo danas.

3.2.4. Turingov test je test za isključivo ljudsku inteligenciju

Sljedeća vrsta kritika TT-a o kojoj ću ovdje raspravljati smatra da je TT test za ljudsku inteligenciju, a ne za opću inteligenciju s obzirom na to da je ispitivač čovjek i da se računalo natječe protiv drugog čovjeka. Drugim riječima, kada bi TT bio suočen s nekom vrstom inteligencije koja uopće ne nalikuje našoj ne bi ju uspio prepoznati, smatraju kritičari. No, možemo li tvrditi da je ljudska inteligencija opća inteligencija? Bi li mi kao ljudi mogli

prepoznati neljudsku inteligenciju, u slučaju da ona postoji?

U članku *Subcognition and the Limits of the Turing Test*, R. M. French tvrdi da je TT isključivo test za ljudsku inteligenciju. Kako bi ilustrirao ovu tvrdnju, on uvodi test galebova. U ovom testu, dva filozofa jednog nordijskog otoka pokušavaju pronaći definiciju pojma leta. Znaju da je let nešto više od lebdenja u zraku (na primjer baloni lebde u zraku) ili od imanja krila i perja (znamo da ih pingvini imaju, a ne lete). Međutim, njihovi jedini primjeri letećih objekata su nordijski galebovi na njihovom otoku. Oni su stoga razvili test galebova koji je inspiriran TT-om: pomoću dva trodimenzionalna radarska ekrana, na jednom promatraju galeba, a na drugom potencijalni leteći stroj. Ako ne mogu otkriti koji je od njih galeb, zaključuju da potencijalni leteći stroj zapravo može letjeti. Možemo odmah vidjeti da je vrlo vjerojatno da su jedini objekti koji mogu proći ovaj test nordijski galebovi ili nekakve njihove replike. Kao rezultat toga, test galebova nije test za opći let već za let nordijskih galebova. French ovo uspoređuje s TT-om i zaključuje da je TT isključivo test za ljudsku inteligenciju, a ne za opću inteligenciju.[30]

No, pitanje je da li postoje drugi oblici inteligencije? Ili još bolje, čak i da postoje, bi li ih mogli prepoznati? Smatram da vjerojatno ne bi mogli prepoznati neljudsku inteligenciju, a kada bi i mogli, prepoznali bi ju pomoću jezika, kao što uostalom prepoznajemo i ljudsku inteligenciju (o važnosti jezika u prepoznavanju ljudske inteligencije ću raspravljati u odjeljku 3.2.6.). Kao primjer mogućeg prepoznavanja drugog oblika inteligencije spomenuo bi recentni SF film *Arrival* u kojem tim predvođen lingvisticom pokušava odgonetnuti jezik inteligentnih vanzemaljaca koji su došli na Zemlju te otkriti razlog njihova dolaska.

No, ako mi kao ljudi ne možemo prepoznati neljudsku inteligenciju, ne možemo niti očekivati da će neki test koji razvijemo to moći. Slično tome, iako se TT često smatra testom za opću inteligenciju, Turing ga je gotovo sigurno osmislio kao test za inteligenciju računala proizvedenih od strane ljudi. Hoćemo li ikada moći stvoriti neljudsku inteligenciju u računalima bez poznavanja ikakvog primjera takve inteligencije?

Vratimo se testu galebova. Kako bi još nordijski filozofi mogli testirati let ako imaju samo jedan primjer leta na svom otoku? Činjenica je da je test galebova iznimno pasivan. Možda bi nordijski filozofi mogli aktivno pristupiti testu, na primjer, kada bi od pilota zatražili da izvodi

[30] French R. M., 1999. *Subcognition and the Limits of the Turing Test*, u Millican P. and Clark A. (ur.) *Machines and Thought: The Legacy of Alan Turing, Volume I*, Oxford: Clarendon Press., str. 13-15.

određene manevre koje su vidjeli od nordijskih galebova. Imajmo na umu da TT to ostvaruje pomoću ispitivača kao aktivnog sudionika koji je u stalnoj interakciji s računalom.

French dalje tvrdi da, ne samo da je TT isključivo test za ljudsku inteligenciju, nego da je potrebno iskusiti svijet kao što to ljudi rade kako bi ga prošli. On daje mnogo primjera "subkognitivnih pitanja", za koje tvrdi da na njih može odgovoriti samo čovjek koji je iskusio svijet. Na primjer; "Na skali od 0 do 10, molimo Vas da ocijenite 'Flugblogs' kao ime koje bi Kellogg's[31] dao svom novom proizvodu žitarica ili suho lišće kao mjesto skrivanja." Bilo koji čovjek, ako dobro poznaje engleski jezik, može vidjeti da je "Flugblogs" grozno ime za žitarice jer podsjeća na riječi "ugly" (*engl.* ružno) i "blob" (engl. grudica). Također, iako definicija "suho lišće" neće nikada uključivati činjenicu da hrpa suhog jesenskog lišća može djeci biti odlično mjesto za skrivanje, većina ljudi bi uočila asocijaciju između ta dva koncepta.[32]

Mislim da je Frenchova tvrdnja kako računalo ne bi moglo odgovoriti na ovakva pitanja nerazumna. Računala mogu usvojiti morfološka, fonološka i sintaktička pravila određenog jezika (vidi odjeljak 3.2.6.), što im omogućava odgovor na pitanje "Flugblogs", a mogu i učiti o različitim ljudskim iskustvima i "čitati" priče, na primjer s Interneta, što im omogućuje odgovor na pitanje o suhom lišću. Također, treba imati na umu da računalo ne mora nužno pasti TT samo zato što nije imao određeno ljudsko iskustvo. Nećemo za čovjeka koji se nikad nije skrivao u hrpi suhog lišća reći da je neinteligentan, već samo da je neiskusan. Ako je računalo dovoljno inteligentno, ali nije imalo puno ljudskih iskustava, bi li ispitivač odmah pomislio da se radi o računalu ili samo o vrlo neiskusnom čovjeku?

Bez obzira na to, čini se da jedini primjer inteligencije kojeg trenutno imamo je ljudska inteligencija vezana za ljudsko iskustvo. Vratimo se na test galebova. Što će se dogoditi ako nordijski filozofi vide zrakoplov na radarskom ekranu? French nam ostavlja dojam da će klasificirati zrakoplov kao ne-leteći i krenuti dalje. To mi se ne čini vjerojatnim. Premda bi ga mogli klasificirati kao ne-leteći, vjerojatno je da će ga klasificirati i kao nešto što nikada prije nisu vidjeli i zabilježiti da na primjer stoji u zraku različito od nordijskog galeba ili balona. Tada bi vjerojatno otišli ispitati avion izbliza i, nakon toga, usavršili svoj test. To je upravo

[31] Tvrtka Kellogg's ili Kellogg je američka multinacionalna tvrtka za proizvodnju hrane sa sjedištem u Battle Creeku, Michigan, Sjedinjene Države. Kellogg's proizvodi žitarice i razne namirnice uključujući kolače, kekse, krekere, zamrznute vafle i vegetarijansku hranu. Najpoznatiji proizvod je Corn flakes.

[32] (French 1999. str. 18-21.)

ono što mislim da bi se dogodilo kada bi se u TT-u pojavio drugačiji oblik inteligencije. Mogli bi prepoznati da nova inteligencija nema istu inteligenciju kao normalna ljudska odrasla osoba; u nekom smislu, vjerojatno ne bi prošla TT. No, zbog načina na kojem je ispitivač u interakciji s računalom putem jezika, on bi je također prepoznao kao nešto potpuno novo. Primjerice kada u filmu *The Machine* ispitivač TT-a postavlja pitanje: "Marija je vidjela psića s prozora i poželjela ga. Što Marija želi?", a računalo odgovori: "Prozor." Ispitivač: "Zašto?" Računalo: "Prozor gleda na svijet, oni su lijepi i pomažu da se ne osjećaš usamljeno". Iako se računalo time "otkrilo" i palo TT, ispitivač prepoznaje da se radi o novom načinu implementacije strojnog učenja u računalu.

Čini se da, ako želimo testirati inteligenciju koja ne zahtijeva ljudsko iskustvo ili je sasvim drugačija od ljudske, o tim vrstama inteligencije moramo naučiti što je više moguće. A najbolji način za učenje o neljudskoj inteligenciji, bez mogućnosti da nam se pojavi iz vedra neba, je pokušati ju stvoriti. Čak i ako se ispostavi da nije moguće stvoriti neljudsku inteligenciju ili ljudsku inteligenciju bez ljudskog iskustva, zasigurno ćemo, pokušavajući to, bolje razumjeti vlastitu.

3.2.5. Nužni i dovoljni uvjeti inteligencije

Što bi točno značilo da računalo prođe TT? Je li prolazak TT nužan uvjet za inteligenciju, dovoljan uvjet ili možda samo daje neku vjerojatnost da je računalo inteligentno? Kritike kojima sam se do sada bavio smatraju da prolazak TT-a nije dovoljan uvjet za inteligenciju (čak i ako računalo prođe TT ne znači da posjeduje inteligenciju) ili da TT ne može testirati sve vrste inteligencija. Međutim, ima i onih koje smatraju da je TT prezahtjevan, odnosno da njegov prolazak nije nužan uvjet za inteligenciju. Ako tvrdimo da prolazak TT-a nije nužan uvjet za inteligenciju, to bi značilo da čak i ako računalo ne prođe TT ne znači nužno da ne posjeduje inteligenciju. Ovu kritiku možemo jednostavno odbaciti ukazujući kako ni sam Turing nije namjeravao da prolazak TT-a bude nužan uvjet za inteligenciju. On je smatrao kako je prigovor da TT nije nužan uvjet za inteligenciju vrlo snažan, ali ako ipak možemo konstruirati stroj koji može zadovoljavajuće igrati igru oponašanja, ne treba nas zabrinjavati.[33] Međutim, prihvaćanje TT-a samo kao dovoljan uvjet za inteligenciju, na neki način umanjuje njegovu važnost budući da bi svatko mogao opravdati pad na testu govoreći da je to ionako samo dovoljan, a ne i nužan uvjet za inteligenciju. Kao rezultat toga, u nastavku, tvrdit ću da bi prolazak TT zapravo mogao biti nužan uvjet za inteligenciju, u smislu da test doista ispituje

[33] (Turing 1950. str. 435.)

opću ljudsku inteligenciju koju obuhvaća putem jezika, i iako nije savršen, i dalje je najbolji test za umjetnu inteligenciju kojim raspolažemo.

3.2.6. Uloga jezika u inteligenciji

Podsjetimo se da u TT-u, ispitivač interagira s računalom isključivo pomoću jezika, nije dopušten nikakav vizualni ili fizički kontakt. S jedne strane, ova metoda interakcije čini se fer; ne bi trebali zaključiti da računala nisu inteligentna samo zato što ne izgledaju kao ljudi. S druge strane, je li jezik dovoljan da obuhvati sve vrste inteligencije koje ljudi posjeduju?

Jedan od prvih filozofa koji su uputili kritiku da jezik ne obuhvaća sve aspekte inteligencije bio je Keith Gunderson. Gunderson daje zanimljivu analogiju između TT-a i "igre gaženja nožnih prstiju". U ovoj igri, ispitivaču nožne prste gazi ili čovjek ili mehanizam koji uključuje kamen i polugu (koja ispušta kamen na prste ispitivača, a zatim ga brzo uklanja, kao što bi to učinila osoba koja gazi nožne prste). Zadaća ispitivača je da odluči je li mu stopalo nagazila ljudska noga ili je na nju ispušten kamen. Gunderson tvrdi kako bi kamen lako mogao proći ovaj test pa bi se iz toga pogrešno moglo smatrati da je kamen čovjek. Zatim naglašava da je ova igra manjkava jer uzima u obzir samo jedan aspekt sposobnosti kamena koji je, zapravo jedina "sposobnost" kamena. Njegova usporedba s TT je da, baš kao što igra gaženja nožnih prstiju testira samo jednu sposobnost kamena, TT testira samo jednu od sposobnosti računala, sposobnost interakcije putem jezika.[34]

Smatram da Gundersonov argument ne uspijeva na jednom kritičnom mjestu. U osnovi on tvrdi da je igra gaženja nožnih prstiju analogna TT-u u smislu da ukazuje samo na sposobnost gaženja nožnih prstiju kao što TT ukazuje samo na jezičnu sposobnost računala, dok se pojam "razmišljanje" referira na više od jedne sposobnosti. S ovim se nikako ne bih složio. Znamo da sposobnost gaženja prstiju ne ukazuje na gotovo nikakvu drugu sposobnost (osim, možda, sposobnosti gaženja drugih dijelova tijela), dok jezična sposobnost ukazuje na mnogo širi raspon sposobnosti. Mislim da bi se s time složio i John G. Stevenson koji u svom radu *On the Imitation Game* iznosi slične argumente protiv Gundersona. Stevenson tvrdi da računalo koje je dobro u igri oponašanja ima različite sposobnosti; čak i ako ne uključuje sve sposobnosti ljudskog razmišljanja, vjeruje da bi popis stvari koje računalo može bilo dosta impresivno. Također navodi da Gunderson ignorira specifičan karakter igre oponašanja i da

[34] Gunderson K., 1964. *The Imitation Game*. Mind, vol. 73, str. 234–245.

predlaže neispravne argumente.[35] Smatram kako je najbolji način da shvatimo sposobnost jezika osvještavanje činjenice da koristimo jezik za podučavanje drugih ljudi gotovo svemu što znamo. Ovdje ne tvrdim da je jezik najučinkovitiji način podučavanja određenih zadataka, već da je jezik opće sredstvo za podučavanje. Drugi način da shvatimo jezičnu sposobnost je da pomno razmislimo o tome kako prepoznajemo inteligenciju kod ljudi.

Netko može tvrditi da prepoznajemo ljudsku inteligenciju pomoću sljedeće logike: "Znam da sam inteligentan i znam da sam čovjek. Stoga, drugi su ljudi inteligentni i ja ih prepoznajem pomoću izgleda." Ovakav test za inteligenciju je nedvojbeno manje striktan od TT-a; prepoznajemo ljudsku inteligenciju jednostavno gledanjem drugog bića. No moglo bi se reći da je ta logika manjkava i da ne smatramo drugu osobu inteligentnom dok nemamo barem neku interakciju s njom. Zato ću proširiti svoj argument na sljedeći način: "Očekujem da su drugi ljudi inteligentni, vizualno ih prepoznajem, a nakon kratke interakcije s njima, znam jesu li inteligentni." Kako možemo interagirati s ljudima kako bi saznali jesu li inteligentni? Pomoću jezika. Ovo priznaju čak i kritičari TT-a. Ovakva logika za procjenu inteligencije kod ljudi funkcionira jer sami možemo pokrenuti proces: ja sam čovjek i stoga mogu pripisati inteligenciju drugim ljudima.[36] Da bi ovaj argument funkcionirao i za računala, moramo tvrditi da je jezik taj koji obuhvaća cijelu ljudsku inteligenciju.

Subartikulirana misao

Čini se da u ovom trenutku postoji velik broj dokaza da jezik obuhvaća većinu vrsta inteligencije, ako ne i svu inteligenciju. Koji su kandidati za vrste inteligencije koje možda ne obuhvaća? U svom radu *Turing's Test and Conscious Thought,* Michie tvrdi da TT ne obuhvaća subartikularnu misao – vrstu kognitivnih operacija kojih uglavnom nismo svjesni (slično Frenchovim subkognitivnim testovima u članku *Subcognition and the Limits of the Turing Test* koji se koriste u prilog tvrdnji da je TT isključivo test za ljudsku inteligenciju; vidi odjeljak 3.2.4.). Michie u svom radu zapravo tvrdi dvije stvari: da TT ne može obuhvatiti subartikularnu misao, što znači da postoji dio inteligencije koju TT ne obuhvaća i kasnije tvrdi da računala ne mogu postići subartikularnu inteligenciju.[37] Treba imati na umu da bi, kada bi druga tvrdnja bila istinita, odmah mogli razlikovati čovjeka od računala, bez obzira

[35] Stevenson J. G., 1976. *On the Imitation Game*. Philosophia, vol. 6, str. 131–133.

[36] Vidi prigovor drugih umova

[37] (Michie 1999. str. 35-41.)

koliko je računalo inteligentno u testu. Pomoću primjera koji Michie daje, pokazat ču da je prva tvrdnja lažna kao što je to vjerojatno i druga.

Michie daje primjer subartikularne inteligencije kroz pitanje "Kako izgovaramo množinu imaginarne engleske riječi poput "platch", "snorp" i "brell"?". Odgovori, očiti bilo kojem čovjeku koji dobro zna engleski, su "platchez", "snorpss" i "brellz".[38] Pomoću ovog primjera možemo odmah razriješiti Michiev prvi argument. Upravo nam je dao precizan način testiranja subartikularne misli. Što ako upitamo računalo koja je množina riječi "platch"? Je li doista tako da računalo ne bi nikada moglo postići ovu vrstu inteligencije? Michieva glavna obrana je da ne možemo programirati računalo s ispravnim morfološkim i fonološkim pravilima pluralizacije. Iako on zapravo za riječi "platch", "snorp" i "brell" navodi algoritam (prema Allen, i sur.)[39], također tvrdi da postoje "pravila" koja programeri teško mogu ugraditi u računalo, dijelom zato što ljudi nisu formulirali sva svoja nesvjesna morfološka pravila tvorbe riječi.

Zanimljivo je da Michie ovdje koristi riječ "pravila", jer zaista se čini da način na koji pluraliziramo riječi slijedi određeni skup pravila, iako ga možda nikada nismo eksplicitno zapisali (ovu hipotezu podupire činjenica da Michie navodi pravilo za riječ "platch"). Može li to spriječiti da računalo usvoji pravila? Ne čini mi se vjerojatnim. Ukoliko opskrbimo računalo velikim jezičnim korpusom, možda internetom i dobrim algoritmom strojnog učenja, bilo bi moguće da računalo otkrije bezbroj primjera riječi i njihovih množina te prepozna pravila tvorbe riječi. Ne mislim trivijalizirati taj proces, već želim naglasiti da je on moguć. Vjerujem da je izgradnja inteligentnih računala mnogo više ostvariva u svijetu u kojem je strojno učenje uobičajena tehnika i gdje postoje veliki skupovi podataka kao što je Internet. Danas programeri ne moraju nužno programirati svako pravilu o jeziku već računala mogu sama učiti iz velikih skupova podataka. Stoga, smatram da postoji mogućnost dostizanja subartikularne misli u umjetnoj inteligenciji.

3.2.7. Alternative Turingovom testu

Prije nego krenem dalje, želio bih nakratko raspraviti o drugim ljudskim sposobnostima za koje određeni autori tvrde kako bi bile bolje u procjeni inteligencije nego što je to jezik. To nisu argumenti da je jezik loš izbor jer eksplicitno ne obuhvaća neki aspekt inteligencije, već

[38] Ibid. str. 38-39.
[39] Allen J., Hunnicutt M. S. and Klatt D., 1987. *From Text to Speech: the MITalk System*, Cambridge University Press.

samo tvrdnje da postoji bolji izbor.

Prva je tzv. "naivna psihologija", sposobnost pripisivanja inteligencije drugim bićima. Općenito se vjeruje da je naivna psihologija svojstvo svakog bića koji posjeduje um i stoga bi svako računalo koje prođe TT trebalo biti sposobno pokazati vještine naivne psihologije. U članku *Naive Psychology and the Inverted Turing Test,* S. Watt zapravo tvrdi da nam TT kakav je danas omogućuje testiranje naivne psihologije, ali pitanja o naivnoj psihologiji moraju biti eksplicitna u svakoj instanci testa, inače to nije pravi test inteligencije. On predlaže Invertirani Turingov test kao učinkovit način testiranja naivne psihologije, a zanimljivo je da taj test zahtijeva pristup računalu koje može proći današnji TT.[40]

Druga je kombinacija jezičnih i senzomotoričkih sposobnosti. U članku *Other Bodies, Other Minds,* S. Harnad predlaže "Totalni Turingov Test" (TTT), koji od računala zahtijeva da, u stvarnom svijetu objekata i ljudi, čini sve što stvarni ljudi mogu učiniti, tako da ispitivač ne može razlikovati radnju računala od radnje stvarnih ljudi.[41] Kao što je istaknuo L. Hauser u svom radu *Reaping the Whirlwind,* Harnad zapravo navodi da jezična sposobnost ima tendenciju da bude dokaz motoričkih vještina.[42] U tom slučaju, računalo koje bi prošlo TT, moglo bi proći i TTT što znači da je TTT suvišan. Zapravo, Harnad ide toliko daleko da tvrdi da su tijelo i um vjerojatno neodvojivi, pa bi stoga i test za motoričke sposobnosti također bio suvišan.[43] Ako pretpostavimo da su tijelo i um zapravo neodvojivi, u tom slučaju nećemo moći stvoriti inteligentne strojeve koji imaju jezične vještine bez da istovremeno imaju i motoričke sposobnosti. Ipak, to je drugo pitanje i ne ukazuje na grešku u testu.

Konačno, P. Schweizer ide korak dalje od Harnada tvrdeći da bi strojevi trebali imati jezične sposobnosti, motoričke vještine i sposobnost da, kao vrsta, stvaraju. On predlaže TTTT ("Truly Total Turing Test" – stvarni totalni Turingov test), koji zahtijeva cjelokupne specifikacije robota koji se razvijaju i stvaraju te, u određenom smislu, pripisuje inteligenciju cijeloj vrsti, a ne pojedincima.[44] Iako to može biti zanimljivo proširenje TT-a, čini se da je teško generirati cijelu vrstu robota koja bi mogla proći TTTT bez da započnemo s jednim robotom koji će proći TT.

[40] Watt S., 1996. *Naive Psychology and the Inverted Turing Test,* Psycoloquy, volume 7(14).
[41] Harnad S., 1991. *Other Bodies, Other Minds*, Minds and Machines, volume 1, str. 44.
[42] Hauser L., 1993. *Reaping the Whirlwind*, Minds and Machines, Vol. 3, No. 2, str. 219-238.
[43] (Harnad 1991. str. 43–54)
[44] Schweizer P., 1998. *The Truly Total Turing Test*, Minds and Machines, volume 8, str. 263–272.

5.2.8. Sažetak kritika

Većina kritika upućena jeziku kao kriteriju TT-a tvrdi da jezik ne obuhvaća sve vrste inteligencije. Međutim, jezik je način na koji testiramo inteligenciju kod ljudi i najčešće je to testiranje manje rigorozno nego kada razgovaramo s računalom u TT-u. Osim toga, nije točno da jezik ne može obuhvatiti sve vrste inteligencije, što se očituje u mom argumentu protiv Michieve subartikulirane inteligencije. Kada bi jednoga dana otkrili određenu vrstu inteligencije koju jezik ne može obuhvatiti, onda bi nam trebao novi test za inteligenciju. Smatram da su nam inteligentna računala potrebna kako bi otkrili postoji li ta specifična vrsta inteligencije (dijelom i zato što je teško izolirati bilo koji dio inteligencije kod ljudi koji već imaju jezične sposobnosti), kao i da bi bilo koji novi test inteligencije vjerojatno bio kamen spoticaja prolasku TT-a. Uostalom, iako bi prolazak nekog novog testa inteligencije bilo veliko postignuće, računalu koji bi ga prošao bi možda nedostajala neka ključna vrsta inteligencije koja je obuhvaćena TT-om.

3.3. Završna razmatranja

Moj glavni argument u ovom poglavlju jest taj da je u ovom trenutku, većina kritika TT-a pogrešna ili irelevantna. Ovo je zasigurno smiona tvrdnja, ali vjerujem da je podržana s nekoliko činjenica. Prvo, postoji mnogo dokaza da je, unatoč kritikama, TT vrlo dobar test za umjetnu inteligenciju. To proizlazi iz činjenice da je jezik možda najbolji pokazatelj koji o inteligenciji imamo; moglo bi se čak tvrditi da koristimo manje stroge testove kada prepoznajemo ljudsku inteligenciju. Drugo, iako se TT s pravom kritizira kao test za testiranje isključivo ljudske inteligencije, nije jasno da li može postojati neljudska inteligencija niti da li bi je mi kao ljudi ikada prepoznali, a još manje vjerojatno da bi je mogli testirati. I treće, ne možemo znati jesu li druge kritike ispravne bez pristupa računalima koja su mnogo inteligentnija od onih koja imamo danas ili bez mnogo većih saznanja o ljudskoj inteligenciji. Vjerujem da je jedan od najboljih načina da bolje razumijemo ljudsku inteligenciju taj da pokušamo izgraditi inteligentna računala s ciljem da jednog dana prođu TT kakav je danas. Znanje koje ćemo dobiti od tih računala će nam pokazati kako promijeniti TT. Ne možemo izvršiti učinkovite izmjene na testu jednostavno nagađajući o tome kako bi se moglo doći do umjetne inteligencije.

Postoji još mnogo različitih kritika TT-a koje nisam spomenuo, a koje su se pojavile u literaturi tijekom proteklih gotovo sedamdeset godina, a koje, nažalost u ovom radu nisam u mogućnosti ispitati. Međutim, postoji jedna koju sam dosad namjerno preskakao, a koja se toliko često veže uz TT i filozofske probleme umjetne inteligencije da se osjećam dužnim dati joj posebno mjesto u ovom radu. Stoga, sljedeće poglavlje posvećujem Johnu Searleu i argumentu kineske sobe.

4. John Searle i argument kineske sobe

Je li moguće da stroj bude inteligentan? Da razumije jezik? Ako može razumjeti jezik znači da je u stanju razumjeti značenja riječi i rečenica kao na primjer: "U petak će kišiti". Ako to može razumjeti, onda može imati i vjerovanja: "Vjerujem da će u petak biti kiša". Ako može imati vjerovanja, onda može imati i druga mentalna stanja poput nada: "Nadam se da će kiša u petak" i strahova; "Bojim se da će u petak kišiti." Ali što je potrebno za razumijevanje značenja riječi? Vjerovanja, nade, strahovi, pa čak i bolovi su mentalna stanja. Za ono što može imati takva mentalna stanja kaže se da posjeduje "um". Dakle, ono što stvarno tražimo jest odgovor na pitanje "Je li moguće da stroj ima um?" Iako postoji mnogo filmova koji nam pokazuju robote koji se ponašaju kao da imaju umove, to je iluzija koju je stvorio Hollywood. Mi želimo znati može li stroj jednoga dana zaista posjedovati um.

Može li imati um ovisi, naravno, o tome što je um. Kroz stoljeća, pojavile su se različite teorije za koje se smatralo da objašnjavaju bitnu prirodu umova. Teorija koja je otišla najdalje kako bi potaknula ljude da vjeruju da stroj može imati um, jest teorija poznata kao funkcionalizam. Prema toj teoriji, mentalno se stanje sastoji isključivo od funkcije koju daje životu pojedinca, to jest, ima kauzalni odnos prema drugim mentalnim stanjima. Postoji nekoliko različitih vrsta funkcionalizma, ali najutjecajniji je komputacijski funkcionalizam koji se ponekad naziva i komputacijska teorija uma. Prema toj teoriji, mentalno je stanje analogno stanjima softvera računala. Mozak je hardver, a um softver. Ako je komputacijski funkcionalizam istinit, onda je moguće da stroj ima mentalna stanja. Sve što je potrebno jest da stroj izvodi pravu vrstu softvera. Funkcionalizam je ohrabrio ljude da vjeruju u mogućnost postojanja inteligentnih strojeva. Drugi takav utjecaj bilo je uvjerenje da je jedini razumni kriterij inteligencije test performansi, kao što je TT. Ako se nešto može ponašati inteligentno kao ljudsko biće, onda bi tome trebalo pripisati posjedovanje inteligencije; poput one poslovice „Ako nešto izgleda kao patka, hoda kao patka i glasa se kao patka onda je to patka.“ Ne bi bilo pošteno zanijekati inteligenciju samo zato što je tijelo sastavljeno od, primjerice, metala umjesto od organskih materijala. U prošlosti su neke skupine ljudi tvrdile da su druge manje inteligentne i manje vrijedne poštovanja zbog svog podrijetla, boje kože ili spola. Ne bismo željeli napraviti istu grešku sa strojevima, ako je doista jedina bitna razlika u materijalu od kojih su napravljeni. Mnogi smatraju da je TT razuman test za inteligenciju i razumijevanje jezika. Međutim, John Searle nije među njima. On odbacuje funkcionalizam i ne vjeruje da je TT pouzdan test za inteligenciju. Zapravo, on vjeruje da ima argument koji

pokazuje da niti jedan klasični program UI, koji se izvodi na digitalnom računalu, ne može stroju omogućiti razumijevanje jezika. Svoj argument naziva „kineska soba“.

Kako bi bolje shvatili argument kineske sobe, moramo se vratiti malo unatrag. Naime, Searleovoj kineskoj sobi prethode tri važna argumenta: Leibnizov mlin, Turingov stroj i kineski mozak.[45]

Gottfried Leibniz je 1714. godine, u svom djelu *Monadologija*, iznio Leibnizov mlin, oblik misaonog eksperimenta usmjerenog protiv fizikalizma. Leibniz nas poziva da zamislimo fizički sustav; stroj, tako konstruiran da navodno razmišlja, osjeća i percipira, odnosno pokazuje znakove inteligencije i svijesti poput ljudskog mozga. Kada bi ga mogli proširili do te mjere da bismo u njega mogli ući kao u mlin, pronašli bi samo mehaničke dijelove (zupčanike, poluge i remene), a ne mentalna stanja (misli, vjerovanja, percepcije, emocije). Leibniz ističe da su unutarnje mehaničke operacije stroja samo dijelovi koji pokreću jedni druge, ništa što je svjesno ili što može objasniti razmišljanje, osjećaje ili percepciju. On je koristio misaoni eksperiment proširenja mozga da se suprotstavi stroju kao entitetu koji proizvodi razmišljanje. Za Leibniza fizička stanja nisu dovoljna, niti konstitutivna stanja mentalnih stanja.[46]

Drugi argument koji prethodi kineskoj sobi jest ideja Turingovog stroja; računala implementiranog od strane čovjeka. Alan Turing, u svom radu *Intelligent Machinery* iz 1948. godine, opisuje stroj za igranje šaha – program koji uključuje niz jednostavnih koraka poput računalnog programa, ali napisanih prirodnim jezikom (na primjer na engleskom) koje čovjek može slijediti. Ljudski operater tog stroja za igranje šaha ne treba uopće znati igrati šah. Sve što mora raditi jest slijediti upute za generiranje poteza na šahovskoj ploči.[47]

Treći argument koji je neposredno prethodio kineskoj sobi pojavio se u ranoj raspravi o funkcionalnim teorijama uma i spoznaje. Funkcionalisti smatraju da su mentalna stanja definirana kauzalnom ulogom koju igraju u sustavu (baš kao što su pekač, guljač i držač određeni onim što čine, a ne od čega su napravljeni). Kritičari funkcionalizma iznijeli su zanimljive misaone eksperimente u kojima se pitaju je li stvarno uvjerljivo da anorganski sustavi mogu imati mentalna stanja. 1974. godine, Lawrence Davis je zamislio dupliciranje mozga pomoću telefonskih linija i ureda s ljudima, a 1978. godine, u *Troubles with Functionalism*, Ned Block je zamislio cijelo stanovništvo Kine uključeno u takvu simulaciju

[45] (Cole 2014.)
[46] Ibid.
[47] Ibid.

mozga. Ovaj je misaoni eksperiment kasnije nazvan kineski mozak, kineska nacija ili kineska dvorana. U filozofiji uma, misaoni eksperiment kineskog mozga razmatra što bi se dogodilo kad bi svaki član kineskog naroda simulirao djelovanje jednog neurona u mozgu, koristeći telefone za simulaciju aksona i dendrita koji povezuju neurone. Pretpostavimo da svaki kineski državljanin dobiva popis telefonskih brojeva te na određeni dan unaprijed određeni građani pokreću postupak pozivom brojeva s njihovog popisa. Kad bi bilo kojem kineskom građaninu zazvonio telefon, on ili ona bi zatim telefonirali onima na njihovom popisu koji bi opet kontaktirali druge i tako dalje. Nije potrebno razmijeniti nikakvu telefonsku poruku, sve što je potrebno je obrazac pozivanja. Popisi poziva bili bi konstruirani na takav način da obrasci poziva implementiraju iste obrasce aktivacije koji se pojavljuju u nečijem mozgu kada je ta osoba, na primjer, u stanju mentalnog stanja boli. Telefonski pozivi igraju istu funkcionalnu ulogu kao i neuroni koji uzrokuju međusobno paljenje. Block je bio prvenstveno zainteresiran za to je li moguće tvrditi da bi stanovništvo Kine moglo biti u boli, dok niti jedan pojedini član stanovništva nije doživio nikakvu bol. Može li postupak kineskog mozga proizvesti um ili svijest? Block tvrdi kako ne može dok Daniel Dennett tvrdi kako može.[48]

Ova su tri misaona eksperimenta zasigurno jako utjecala na Searleovo razmišljanje i formiranje argumenta kineske sobe.

4.1. Argument kineske sobe

Jezgra argumenta kineske sobe je vrlo jednostavni misaoni eksperiment. Searle se zamišlja u zatvorenoj sobi u kojoj se nalazi veliki broj priručnika na engleskom jeziku. Ljudi izvan sobe su izvorni kineski govornici koji kroz otvor ubacuju komadiće papira s kineskim znakovima. Searle ne prepoznaje niti jedan od tih simbola, oni su mu jednostavno besmisleni oblici. Međutim, ti simboli imaju značenje. Oni su kineski znakovi, štoviše, to su pitanja na kineskom. Searleov zadatak je pregledavati priručnike dok ne pronađe niz simbola koji izgledaju baš poput onih napisanih na papiru. Kada pronađe taj niz, priručnik (koji je na engleskom jeziku) će mu otkriti novi niz simbola (u obliku kineskih znakova) koji da napiše kao odgovor i gurne natrag kroz otvor. Ulazni listovi papira predstavljaju pitanja, a izlazni odgovarajuće odgovore na ta pitanja. Kada kineski govornici dobiju odgovore na njihova

[48] Ibid.

pitanja, razumno zaključuju da u sobi postoji inteligentna osoba koja razumije kineski. No Searle, iako je točno odgovorio na pitanja, ne razumije ni riječi.[49]

Priručnici djeluju kao računalni program. Svaka stranica daje konkretne upute za rukovanje tim simbolom, ali ne i njegovo značenje. U priručnicima ne postoji engleski prijevod kineskih znakova (nisu ni slični kinesko-engleskom rječniku) već, poput samog računalnog programa, upućuju čitatelja kako manipulirati simbolima na temelju njihovih formalnih svojstava, njihovog oblika i položaja, a ne njihovog značenja. Na primjer, ako ovdje vidite znak "X", tamo napišite znak "Y" i tako dalje.[50]

Pretpostavimo da su priručnici u kineskoj sobi dokazali da su učinkoviti UI programi koji mogu proći Turingov test na kineskom jeziku. Kineski znakovi koji su ušli u sobu predstavljaju pitanje: "Koji je glavni grad Kine?", a kineski znakovi koje je Searle napisao kao odgovor na to pitanje znače: "Peking". Ako kineska soba može proći Turingov test, treba li reći da soba "razumije" kineski? Searle kaže da ne. Kada bi postavljena pitanja bila napisana na engleskom jeziku, Searle bi ih mogao normalno pročitati i odgovoriti na njih te što je najvažnije, razumjeti ih. Međutim, njegov odnos prema kineskim znakovima je sasvim drukčiji. Mogao bi provesti godine, uzimajući tisuće komada papira s kineskim znakovima i mogao bi napisati tisuće inteligentnih odgovora na kineskom, a da ni u kojem trenutku ne razumije kineski. Dakle, osoba unutar kineske sobe predstavlja ljudsku simulaciju rada računala, koja je slična Turingovom stroju. Ona slijedi engleske upute za manipuliranje kineskim znakovima kao što računalo izvodi program koji je napisan na određenom računalnom jeziku. Budući da računalo radi isto što i osoba u kineskoj sobi; manipulira simbolima na temelju same sintakse, Searle zaključuje da niti jedno računalo samim izvođenjem programa ne može zaista razumjeti kineski niti bilo koji drugi jezik, čak i ako prođe TT.[51]

Ovaj argument usmjeren je prema poziciji koju Searle naziva "jaka UI". Jaka UI je stajalište koje smatra da odgovarajuće programirana računala (ili čak sami programi) mogu razumjeti prirodni jezik te imaju druge mentalne sposobnosti slične ljudskima. Prema jakoj UI, računalo može inteligentno igrati šah, riješiti kompleksni matematički algoritam ili razumjeti jezik. Nasuprot tome, "slaba UI" je stajalište da su računala korisna samo u psihologiji, lingvistici i

[49] Searle, J., 1984, *Minds, Brains and Science*, Cambridge: Harvard University Press, str. 32.
[50] Ibid.
[51] Ibid. str. 33-34.

drugim područjima kognitivne znanosti zato što mogu *simulirati* mentalne sposobnosti. Ali slaba UI ne tvrdi da računala zapravo razumiju ili da su inteligentna. Argument kineske sobe nije usmjeren na slabu UI, već na stajalište da formalna komputacija nad simbolima može proizvesti razumijevanje.[52]

Argument kineske sobe možemo sažeti kao *reductio ad absurdum* protiv Jake UI na sljedeći način:

(1) Ako je jaka UI istinita, onda postoji program za kineski, takav da ako bilo koji računalni sustav pokreće taj program, taj sustav time razumije kineski.

(2) Ja bih mogao pokrenuti program za kineski, a da time ne bi razumio kineski.

(3) Stoga je jaka UI lažna.

Argument kineske sobe podržava drugu premisu. Zaključak ovog argumenta je da pokretanjem programa ne možemo stvoriti razumijevanje. U razumijevanju jezika, ljudi ne samo da manipuliraju simbolima na temelju njihovih formalnih svojstava, već čine još nešto (što točno još nismo otkrili) na temelju čega razumijemo značenje simbola, što u kineskoj sobi nije slučaj.

Važno je napomenuti kako Searle ne smatra da strojevi ne mogu razumjeti jezik, odnosno misliti, u smislu da su ljudska bića biološki strojevi koji razumiju. Također, ne tvrdi niti da računalo ne može misliti. Opet, ljudi su računala u smislu da se određene operacije ljudskog mozga mogu ispravno opisati kao "računanje". Searle smatra da je pravo pitanje: 'Može li digitalno računalo, kako je definirano, misliti?' I na ovo pitanje odgovor je "ne" i to stoga što je računalni program isključivo sintaktički definiran. Međutim, razmišljanje zahtijeva nešto više od pukog manipuliranja beznačajnim simbolima, ono uključuje smislene semantičke sadržaje. Ovi semantički sadržaji su ono što Searle naziva 'značenjem'.[53]

Važno je naglasiti kako Searle ne govori o određenoj fazi računalne tehnologije. Argument nema nikakve veze s napretkom u računalnoj znanosti, programima, memoriji, brzini računalnih operacija ili čak s mogućim izumom robota. Možemo razumno očekivati da će u budućnosti doći do još većeg napretka u tehnologiji i nema sumnje da ćemo tada moći puno bolje simulirati ljudsko ponašanje na računalima nego što to možemo u ovom trenutku. No,

[52] Ibid. str. 28-29.
[53] Ibid. str. 35-36.

ono što Searle tvrdi jest da su, ako govorimo o mentalnim stanjima, odnosno o umu, sve te simulacije jednostavno nebitne. Nije važno koliko je tehnologija dobra ili koliko su brza računala. U digitalnom se računalu operacije moraju definirati sintaktički, dok svijest, misli, osjećaji, emocije uključuju više od same sintakse; uključuju semantiku. Kineska soba pokazuje da se ne može dobiti semantika (značenje) iz sintakse (manipulacija formalnim simbolima). Te značajke, po definiciji, računalo ne može duplicirati koliko god jaka bila njegova sposobnost da simulira. Nikakva simulacija sama po sebi ne predstavlja dupliciranje. Ovdje je ključna razlika između dupliciranja i simulacije. O simulaciji, duplikaciji, semantici i sintaksi raspravljat ću kasnije, a sada ću se okrenuti najvažnijim prigovorima argumentu kineske sobe.[54]

4.2. Prigovori argumentu kineske sobe

Ne postoji konsenzus o zaslugama argumenta kineske sobe. Neki vjeruju kako on dokazuje da se digitalna računala ne mogu programirati da bi razumjela jezik pa je stoga Turingov test pretjerano vezan za ponašanje. Drugi vjeruju da argument kineske sobe ne dokazuje ništa. Prigovori argumentu kineske sobe prate određene linije misli prema kojima se mogu svrstati u tri kategorije:[55]

(1) Neki kritičari priznaju da osoba u prostoriji ne razumije kineski, ali u isto vrijeme smatraju kako postoje neke druge stvari koje razumije. Ovi kritičari smatraju da tvrdnja da čovjek u prostoriji ne razumije kineski ne vodi do zaključka da nikakvo razumijevanje nije stvoreno. Tu bi moglo biti razumijevanja od strane većeg ili drugačijeg entiteta. To je strategija prigovora sustava i prigovora virtualnog uma. Ti prigovori smatraju da bi moglo biti razumijevanja u izvornom scenariju kineske sobe.[56]

(2) Drugi kritičari priznaju Searleovu tvrdnju da se samim pokretanjem programa za obradu prirodnog jezika, kako je opisano u scenariju kineske sobe, ne stvara nikakvo razumijevanje, bilo kod čovjeka ili računalnog sustava. No, ovi kritičari smatraju da je moguća određena varijacija na scenarij kineske sobe koja bi mogla dovesti do

[54] Ibid. str. 36-37.
[55] (Cole 2014.)
[56] Ibid.

razumijevanja. Varijacija bi mogla biti da se računalo ugradi u tijelo robota koji ima interakciju s fizičkim svijetom putem senzora i motora (prigovor robota) ili bi to moglo biti računalo koje simulira detaljne operacije cijelog mozga, neuron po neuron (prigovor simulatora mozga).[57]

(3) Na kraju, neki kritičari smatraju da osoba u izvornom scenariju kineske sobe može razumjeti kineski unatoč Searleovom poricanju ili smatraju da je takav scenarij nemoguć. Na primjer, kritičari su tvrdili da je naša intuicija u takvim slučajevima nepouzdana. Drugi prigovaraju tvrdnji da bilo koji sustav (na primjer Searle u sobi) može pokrenuti bilo koji računalni program. I konačno, neki su tvrdili da ako nije razumno pripisati razumijevanje na temelju ponašanja izloženog u kineskoj sobi, onda ne bi bilo razumno niti pripisati razumijevanje ljudima na temelju dokaza sličnih ponašanja (Searle to naziva prigovorom drugih umova).[58]

Osim ovih prigovora argumentu kineske sobe, postoje prigovori i protiv Searlove tvrdnje da nije moguće dobiti semantiku (to jest značenje) iz sintaktičke manipulacije simbolima, uključujući one vrste manipulacije koja se odvija unutar digitalnog računala. Ovaj ću prigovor detaljnije elaborirati u odjeljku *Sintaksa nije semantika*, a sada ću predstaviti glavne prigovore argumentu kineske sobe.

4.2.1. Prigovor sustava

U knjizi *Minds, Brains and Science*, Searle identificira i raspravlja o nekoliko prigovora argumentu kineske sobe na koje je naišao predstavljajući ga na različitim sveučilištima. Searle kaže kako je vjerojatno najčešći prigovor argumentu kineske sobe prigovor sustava.[59]

Prigovor sustava, za kojeg Searle kaže da je bio izvorno povezan s Yaleom, priznaje da osoba u sobi ne razumije kineski, ali smatra da je ona samo dio, središnja procesorska jedinica (CPU), većeg sustava.[60] Veći sustav obuhvaća memoriju (koja sadrži prijelazna stanja) i upute potrebne za odgovaranje na kineska pitanja. Dok osoba koja izvodi program ne razumije kineski, sustav kao cjelina ga razumije. Ključna tvrdnja je da entitet koji razumije kineski nije osoba u sobi, već cijeli sustav. Sustav nije samo Searle koji hoda po sobi, uzima papire i vraća

[57] Ibid.
[58] Ibid.
[59] Ibid.
[60] (Searle 1980. str. 419.)

odgovore, već Searle plus program za obradu kineskih znakova i svim informacijama koje on utjelovljuje. Takav sustav razumije kineski.[61]

Ned Block bio je jedan od prvih predstavnika prigovora sustava, uz mnoge druge, uključujući Jacka Copelanda, Daniela Dennetta, Jerrya Fodora, Johna Haugelanda, Raya Kurzweila i Georgesa Reya. Rey (1986.) kaže kako je osoba u sobi samo CPU sustava. Kurzweil (2002.) kaže kako je ljudsko biće samo izvršitelj i nema značaja, vjerojatno smatrajući da svojstva izvršitelja nisu nužno i svojstva sustava. Kurzweil, u duhu Turingovog testa, kaže da ako sustav pokaže očiglednu sposobnost razumijevanja kineskog, on bi zaista trebao razumjeti kineski. On smatra da se Searle suprotstavlja samome sebi kad tvrdi da stroj govori kineski, ali istovremeno ga ne razumije.[62]

Searleov odgovor na prigovor sustava prilično je jednostavan; Searle u principu kaže kako osoba može internalizirati cijeli sustav, zapamtiti sve upute i izvršiti sve izračune u glavi. Osoba bi mogla napustiti sobu i lutati vanjskim svijetom, možda čak i razgovarati na kineskom, ali još uvijek ne bi imala načina da formalnim simbolima pridoda neko značenje.[63] Osoba bi u ovom slučaju bila cijeli sustav, ali još uvijek ne bi razumjela kineski. Na primjer, ne bi znala značenje kineske riječi za hamburger. Još uvijek ne bi mogla dobiti semantiku iz sintakse.[64]

4.2.2. Prigovor robota

Prigovor robota priznaje da je Searle u pravu kod scenarija kineske sobe koji nam pokazuje da osoba zarobljena u sobi ne može razumjeti jezik ili znati značenje pojedinih riječi, ali da je pogriješio kada misli da je pokazao da su računala intrinzično nesposobna za proizvodnju razumijevanja. Što ako računalu dodamo senzore (kamere, mikrofone, termistore, detektore mirisa i tako dalje) i pustimo ga da luta svijetom u robotskom tijelu? Može li onda, baš kao i ljudi, biti u stanju naučiti značenja riječi povezujući ih sa svijetom? Prigovor robota predlaže da stavimo digitalno računalo u tijelo robota sa senzorima, kao što su videokamere i mikrofoni, i dodamo pokretače poput kotača za pomicanje i ruke kojima će manipulirati stvarima u svijetu. Takvo digitalno računalo u tijelu robota, oslobođeno iz sobe moglo bi pridati značenja simbolima i zapravo razumjeti prirodni jezik; moglo bi, poput djeteta, učiti

[61] (Searle 1984. str. 34.)
[62] (Cole 2014.)
[63] (Searle 1980. str. str. 420.)
[64] Ibid.

promatranjem i oponašanjem. Među onima koji su podržali verzije ovog prigovora na jedan ili drugi način su Margaret Boden, Tim Crane, Daniel Dennett, Jerry Fodor, Stevan Harnad, Hans Moravec i Georges Rey.[65]

Sada bi, na primjer, prema prigovoru robota došlo bi do razumijevanja kineskog znaka za hamburger.[66] Nama se čini razumno smatrati da znamo što je hamburger jer smo ga vidjeli, a vjerojatno i kušali ili barem čuli druge ljude kako govore o njemu. Razumijemo što je hamburger povezujući ga sa stvarima koje gledamo, koje kušamo ili o kojima slušamo. Tako bi i robot, pomoću interakcije sa svijetom, formalnim simbolima mogao pridati značenje. Prigovor robota apelira na „široki sadržaj" ili „eksternalističku semantiku". Time ovaj prigovor može zadržati konzistentnost sa Searleovom tvrdnjom da su sintaksa i unutarnje veze nedovoljne za semantiku, istovremeno implicirajući da odgovarajuća uzročna povezanost sa svijetom može pružiti sadržaj unutarnjim simbolima.[67]

Searle misli da ovaj prigovor argumentu kineske sobe nije ništa bolji, a ni bitno drugačiji od prigovora sustava. Sve što senzori daju jest dodatan input računalu, ali to je i dalje samo sintaktički input. To možemo vidjeti ako napravimo promjenu u scenariju kineske sobe tako da cijelu kinesku sobu zajedno s osobom preselimo u glavu robota. Osoba sad, pored kineskih znakova, dobiva i niz binarnih brojeva. Priručnik, uz kineske znakove, sadrži i upute za te brojeve. Osoba u sobi ne zna da su brojevi koje dobiva zapravo digitalizirani izlazi video kamere i moguće drugih senzora robota. Odgovore koje daje služe kako bi motori unutar robota pomakli noge, ruke ili "oči" robota. Searle tvrdi da osoba unutar sobe ni u ovom scenariju neće razumjeti kineski jer sve što čini jest slijedi upute za manipulaciju formalnim simbolima. Ovi dodatni sintaktički inputi neće promijeniti ništa kako bi osobi omogućili povezivanje značenja s kineskim znakovima. Jedina razlika u ovom scenariju je što će osoba imati više posla.[68]

4.2.3. Prigovor simulatora mozga

Prigovor simulatora mozga traži od nas da zamislimo računalo koje radi na potpuno drugačiji način od uobičajenog; takvo računalo simulira paljenje neurona koji se javljaju u mozgu izvornog kineskog govornika kada razumije kineski, simulirajući svaki pojedini neuron i

[65] (Cole 2014.)
[66] (Searle 1984. str. 34.)
[67] (Cole 2014.)
[68] (Searle 1980. str. 420.)

svako paljenje. Budući da računalo tada funkcionira isto kao i mozak izvornog kineskog govornika, obrađujući informacije na isti način, razumjet će kineski. Kada bi tvrdili da računalo ne razumije kineski morali bi poreći da izvorni kineski govornici razumiju kineski jer na razini sinapsa nema nikakve razlike između programa računala i mozga kineskih govornika. Glavni zastupnici ovog prigovora su Paul i Patricia Churchland te Andy Clark.[69]

Kao odgovor na ovaj prigovor, Searle tvrdi da simuliranjem operacija mozga i dalje ne postižemo razumijevanje; što se može vidjeti iz sljedeće varijacije na scenarij kineske sobe. Pretpostavimo da soba sadrži veliki set ventila i vodovodnih cijevi postavljenih u istom rasporedu kao i neuroni u mozgu kineskog govornika. Umjesto manipulacije simbolima, osoba sada mora upravljati složenim skupom vodovodnih cijevi s ventilima koji ih povezuju. Služeći se prilagođenim priručnikom, osoba zna koje ventile mora otvoriti, a koje zatvoriti kao odgovor na dobiveni input. Svaka cijev odgovara sinapsi u kineskom mozgu, a cijeli je sustav opremljen tako da nakon što osoba otvori prave ventile, kineski odgovor izlazi van. Searle tvrdi da je očito da osoba ni ovom slučaju neće razumjeti kineski. Problem simulatora mozga, kako ga Searle dijagnosticira, jest da i dalje simulira samo formalnu strukturu neuronskih paljenja koja nije dovoljna za proizvodnju značenja i mentalnih stanja.[70] Simulacija aktivnosti mozga nije što i duplikacija mozga. Searleov odgovor je blizak scenarijima u Leibnizovom mlinu i kineskoj dvorani.[71]

U siječnju 1990. godine argument kineske sobe predstavljen je u poznatom znanstvenom časopisu *Scientific American*. Vodeća opozicija Searleovom argumentu u tom broju pripala je filozofima Paulu i Patriciji Churchland. Churchlandi se slažu sa Searleom da kineska soba ne razumije kineski, ali smatraju da sam argument iskorištava naše neznanje kognitivnih i semantičkih pojava. Oni zagovaraju pogled na mozak kao konekcionistički sustav, a ne sustav manipuliranja simbola prema formalnim pravilima. Smatraju da su brzina i povezanost mozga ono što je bitno za svijest. Predstavljaju paralelni slučaj kojeg nazivaju *Svjetleća soba*, gdje osoba maše magnetom i tvrdi da odsutnost vidljivog svjetla pokazuje da je Maxwellova elektromagnetska teorija lažna. Prema Maxwellovoj teoriji, magnet bi trebao početi svijetliti, međutim, da bi došlo do toga trebalo bi mahati magnetom mnogo brže nego što to čovjek može. Churchlandi stoga smatraju da kineska soba ne može u potpunosti duplicirati cijeli sustav i brzinu koja je potrebna da bi se obradila informacija i proizvelo razumijevanje. Oni se

[69] (Cole 2014.)
[70] (Searle 1980. str. 420-421.)
[71] (Cole 2014.)

slažu sa Searleovom tvrdnjom protiv jake UI, ali podržavaju prigovor simulatora mozga, tvrdeći da nas, kao kod svjetleće sobe, naša intuicija može iznevjeriti kada uzimamo u obzir ovako složen sustav. Iako niti jedan neuron u našem mozgu ne razumije engleski, naš cijeli mozak razumije, zaključuju Churchlandi.[72]

U svojoj knjizi *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing* iz 1991.godine, Andy Clark drži da je Searle u pravu kada tvrdi da računalo koje pokreće Schankov program ne zna ništa o restoranima i hamburgerima, barem ako pod "znati nešto" mislimo na nešto što razumijemo. Međutim, Searle misli da se to odnosi na bilo koji računalni model, dok Clark smatra da, kada se radi o konekcionističkom modelu, Searle nije u pravu. Clark stoga brani prigovor simulatora mozga. Mozak misli na temelju svojih fizičkih svojstava, no koja su fizička svojstva mozga važna? Clark odgovara da ono što je važno u mozgu su „varijabilne i fleksibilne podstrukture“ koje nedostaju konvencionalnim UI sustavima, ali to ne znači da je komputacionalizam ili funkcionalizam lažan. Sve ovisi o tome na kojoj razini primjenjujemo funkcionalne jedinice. Clark brani „mikrofunkcionalizam“ i smatra da bi trebali gledati mikro razinu funkcionalnog opisa, na primjer razinu neuronskih mreža. Stoga su Clarkovi pogledi slični onima Paula i Patricije Churchland, priznajući da Searle ima pravo u vezi s Schankovim sustavom i sustavima za obradu na simboličkoj razini, ali smatrajući da je u krivu u vezi s konekcionističkim modelom.[73]

Searle koristi prilagodbu argumenata kineskog mozga kako bi pokazao da instancijacija konekcionističke mreže nije dovoljan uvjet za stvaranje razumijevanja kineskog. Pretpostavimo da je osnovna kineska soba proširena na veliku dvoranu punu engleskih monolingvista koji su rasprostranjeni poput čvorova u mreži te slijede engleske priručnike koji im govore koje simbole trebaju dodavati jedni drugima. Ovim postupkom oni izvode iste procese koje izvodi mozak prilikom kineskog govora, samo što nitko u dvorani ne razumije kineski.[74]

Ray Kurzweil (2002.) tvrdi da se Searleov argument može izokrenuti kako bi pokazao da ni ljudski mozak ne može razumjeti; mozak uspijeva razumjeti manipulacijom neurotransmitera i drugim mehanizmima koji su sami po sebi besmisleni. U kritici Searleovog odgovora na prigovor simulatora mozga, Kurzweil kaže da ako drastično povećamo Searleovu kinesku sobu, tko kaže da cijeli sustav od sto trilijuna ljudi koji simuliraju kineski mozak neuron po

[72] Ibid.
[73] Ibid.
[74] Searle, J., 1990. *Is the Brain's Mind a Computer Program?*, Scientific American, 262, no.1, str. 26-31.

neuron ne bi bio svjestan? Svakako, bilo bi ispravno reći da takav sustav zna kineski i ne bi mogli tvrditi da nije svjestan više nego što bi to mogli tvrditi za bilo koji drugi entitet. Ne možemo znati subjektivno iskustvo drugog entiteta, zaključuje Kurzweil.[75]

Searle tvrdi da ova kritika i dalje ne uspijeva doprijeti do jezgre argumenta kineske sobe. Možemo pokrenuti ovu simulaciju i još uvijek ne znati što bilo koja kineska riječ znači. Simulacija koja se odvija na računalu, da li to bila simulacija probave, oluje ili mozga, je i dalje samo simulacija i različita je od duplikacije (više u odjeljku *Simulacija nije duplikacija*).[76] Nitko ne pokisne na računalnoj simulaciji oluje, računalo ne probavlja pizzu u računalnoj simulaciji probave kao što ne može niti razumjeti pomoću simuliranja razumijevanja, zaključuje Searle.[77]

4.2.4. Kombinirani prigovor

Kombinirani prigovor pretpostavlja sva tri prethodna prigovora: robot s digitalnim računalom koji simulira mozak tako da se sustav u cjelini ponaša nerazlučivo od čovjeka. U ovom slučaju, razumno je pripisati razumijevanje i intencionalnost takvom sustavu kao cjelini.[78]

Searle razmatra i kombinirani prigovor i smatra da, budući da ni jedan od prethodnih prigovora nije uspio odbaciti argument kineske sobe, ne mogu ga niti sva tri uzeta zajedno, zato što je nula puta tri i dalje nula. Iako priznaje da bi bilo racionalno prihvatiti hipotezu da takav robot ima intencionalnost, ali samo dok ne bi saznali kako on funkcionira. Čim bi saznali da je to računalo koje bez razumijevanja manipulira simbolima na temelju sintakse, a ne značenja, prestali bi mu pripisivati intencionalnost[79] čak i kad bi takav robot bio naš najbolji prijatelj ili partner s kojim je izgrađen cjeloživotni odnos (priče iz znanstvene fantastike uključujući epizode televizijske serije *Westworld* bazirane su na takvim mogućnostima).

4.2.5. Prigovor drugih umova

O prigovoru drugih umova ili, kako ga je Turing nazvao, argumentu iz svijesti, već sam raspravljao u prethodnom poglavlju. Prigovor drugih umova podsjeća nas da promatranjem ponašanja znamo jesu li drugi ljudi inteligentni i razumiju li. Prema tome, ako računalo može

[75] (Cole 2014.)
[76] (Searle, 1990.)
[77] Searle, J., 2015. *Consciousness in Artificial Intelligence*, Talks at Google, intervju.
[78] (Searle 1980. str. 421.).
[79] Ibid.

proći testove ponašanja (poput TT-a) te ako pripisujemo razumijevanje drugim ljudima, onda ga načelno moramo pripisati i računalima. S obzirom na to da se kineska soba bihevioralno ne razlikuje od izvornog kineskog govornika, moramo zaključiti ili da kineska soba razumije kineski ili da ni kineski govornik ne razumije kineski.[80]

Searle odbacuje biheviorizam i tvrdi da zna da je njegov pas Tarski svjestan, a da njegov *smartphone* nije, i to ne iz biheviorističkih razloga, već zato što vidi da Tarski ima mehanizam relativno sličan ljudskom; ima oči, uši, usta, kožu i tako dalje, odnosno posjeduje mehanizme kojima posreduje između ulaznih podražaja i izlaznog ponašanja. Zato je Searle potpuno siguran da je njegov pas svjestan, kao što je siguran i da su drugi ljudi svjesni, a njegov *smatrphone* nije. Searleov odgovor na prigovor drugih umova vrlo je kratak; pretpostavljamo da drugi ljudi imaju umove i razumiju baš kao što u fizici pretpostavljamo postojanje objekata.[81]

Hans Moravec, direktor laboratorija robotike na Sveučilištu Carnegie Mellon i autor knjige *Robot: Mere Machine to Transcendent Mind,* tvrdi da Searleov argument odražava intuicije iz tradicionalne filozofije uma koje nisu u koraku s novom kognitivnom znanošću. Moravec podržava verziju prigovora drugih umova i smatra da računalima ima smisla pripisati intencionalnost iz istih razloga iz kojih ima smisla pripisati je ljudima. Njegov „interpretativni položaj“ sličan je stajalištima Daniela Dennetta. Moravec nadalje napominje da je sposobnost pripisivanja atribucija intencionalnosti jedna od stvari koju pripisujemo drugim ljudima, a zatim takve atribucije stavljamo na sebe. Takva samoreprezentacija je u srcu svijesti, smatra Moravec.[82] Slično tome, Kurzweil smatra da ima smisla pripisati svijest računalima ako je pripisujemo ljudima, jer kako znamo da su ljudi svjesni? Kurzweil se slaže sa Searleom da postojeća računala ne razumiju jezik, što se očituje i činjenicom da još uvijek nisu prošla TT, ali on predviđa da ćemo u budućnosti razviti inteligentna računala i pita Searla kako ćemo onda znati jesu li ta računala svjesna ili se samo čine takvima.[83]

Searle smatra da nema nikakve sumnje u to da su ljudi svjesni. To čak nije ni teorija koju brani već je pozadinska pretpostavka; isto kao što pretpostavlja da je pod na kojem stoji čvrst, tako pretpostavlja da su ljudi svjesni.[84] Searle međutim smatra da ključna točka nije pitanje

[80] Ibid.
[81] (Searle, 2015.)
[82] (Cole 2014.)
[83] (Searle, 2015.)
[84] Ibid.

kako znamo jesu li drugi ljudi svjesni, već je ključno ono što im pripisujemo kada im pripisujemo svijest i kognitivna stanja. On smatra da to nisu samo komputacijski procesi i njihovi izlazi jer oni mogu postojati i bez kognitivnih stanja.[85] Searle kaže da je mozak mehanizam koji stvara svijest prilično složenim i još uvijek ne potpuno shvaćenim neurobiološkim procesima. Problem računala je da nema veze sa specifičnostima implementacije; bilo koja implementacija je dobra za provođenje koraka u programu. Programi su čisto formalni ili sintaktički, dok mozak nije. Mozak je specifičan biološki organ koji djeluje nad specifičnim biokemijskim principima i stvara svijest.[86] Vjerojatni razlog zbog kojeg Searle misli da možemo odbaciti tvrdnje da su digitalna računala svjesna je taj što znamo da obrađuju informacije isključivo sintaktički. Također, poznavajući unutarnje djelovanja kineske sobe znamo da osoba u sobi ne razumije kineski. Doista, Searle vjeruje da je to glavna točka koju kineska soba pokazuje, a o kojoj ću još detaljnije raspraviti u odjeljku *Sintaksa nije semantika.*

4.2.6. Prigovor napretka tehnologije

Mnogi znanstvenici u području UI sugeriraju da čak i ako Searle ima pravo da programiranje ne može dovesti do toga da računala steknu intencionalnost i kognitivna stanja, u budućnosti bi se moglo otkriti da, osim programiranja, postoji drugi način implementacije istih. Smatraju da će UI jednog dana razviti tehnološke resurse potrebne za rješavanje zagonetke kineske sobe.

Ova su predviđanja odlična, ali ne pomaže projektu jake UI. Doista, u svjetlu Church-Turingove teze, očekivana tehnologija morala bi ići daleko iznad onoga što trenutno možemo zamisliti. Searle smatra da ovo trivijalizira projekt jake UI redefinirajući i napuštajući izvorni zahtjev jake UI da su mentalni procesi računalni procesi nad formalno definiranim elementima. Možda ćemo u budućnosti razviti tehnologiju superračunala, ali jaka UI tvrdi da nam je za razmišljanje dovoljan Turingov stroj. Ako jaka UI napusti tu preciznu i dobro definiranu tezu, njegove primjedbe se više neće moći primjenjivati, kaže Searle.[87]

[85] (Searle, 1980. str. 420-421.)
[86] (Searle, 2015.)
[87] (Searle, 1980. str. 422.)

4.2.7. Prigovor virtualnog uma

Prigovor virtualnog uma priznaje, poput prigovora sustava, da osoba u kineskoj sobi ne razumije kineski samim pokretanjem programa. Međutim, za razliku od prigovora sustava, ovaj prigovor smatra da pokretanje programa može stvoriti nove entitete različite od sustava i njegovih dijelova (osoba u sobi ili CPU računala). Konkretno, može stvoriti virtualni um koji razumije kineski. Ovaj virtualni um bit će različit od osobe u sobi i cijelog sustava. Psihološke osobine virtualnog uma (uključujući jezične sposobnosti) ovisit će o načinu programiranja i neće biti identične s osobinama i sposobnostima CPU-a ili osobe u sobi. Prema prigovoru virtualnog uma, pogreška u argumentu kineske sobe je ta što uzima kao tvrdnju jake UI da računalo razumije kineski dok tvrdnja koju treba razmatrati je ta da računalo stvara um koji razumije kineski. Ovdje možemo povući analogiju s računalnim video igrama. Likovi u tim igricama imaju različite osobine i sposobnosti te nisu identični s hardverom ili programom koji ih stvara.[88]

Prigovor virtualnog uma prvi su predložili Marvin Minsky (1980.) te Aaron Sloman i Monika Croucher (1980.) kada se argument kineske sobe prvi puta pojavio.[89] U svom članku *Computation and Consciousness*, iz 1989. godine, Tim Maudlin razmatra i argument kineske sobe. Navodeći Minskog te Slomana i Crouchera, ističe da bi entitet koji razumije mogao biti različit od fizičkog sustava te smatra da Searle nije učinio ništa kako bi isključio mogućnost istovremenog postojanja disjunktnih mentaliteta.[90]

David Cole najopsežnije razvija prigovor virtualnog uma i u svom članku *Articial Intelligence and Personal Identity* iz 1990. godine, tvrdi da je Searle argumentom kineske sobe uspio dokazati da nijedno računalo nikad neće razumjeti prirodni jezik. Coleov naglasak, međutim, nije na tome da računalo ne razumije, već na tome da računalo uzrokuje novi entitet (a) koji nije identičan računalu, ali (b) koji postoji isključivo zbog računalne aktivnosti stroja i (c) razumije engleski jezik (ili u slučaju kineske sobe (c) razumije kineski). Novi entitet je virtualna osoba u skladu s onim što računalni znanstvenici nazivaju virtualnim strojem. Cole smatra da to što računalo ne razumije ne dokazuje da ne postoji ništa što razumije. Iz činjenice da netko (na primjer, Searle u sobi) ne razumije kineski ne slijedi da nitko ne razumije kineski. Searleov neuspjeh u razumijevanju kineskog unutar sobe ne dokazuje da nikakvo razumijevanje nije stvoreno. Cole tvrdi da mentalne osobine koje čine Johna Searlea, njegova

[88] (Cole 2014.)
[89] Ibid.
[90] Maudlin, T., 1989. *Computation and Consciousness*, Journal of Philosophy, LXXXVI: str. 414-415.

vjerovanja, želje, sjećanja i osobine ličnosti su sve nevažne i uzročno inertne u proizvodnji odgovora na kineska pitanja. Ako se pokretanjem programa stvara razumijevanje, um koji razumije kineski neće biti računalo, kao što ni u kineskoj sobi um koji razumije kineski nije osoba u sobi. Cole smatra da se pokretanjem programa stvara različit entitet od uma osobe u sobi, takozvani virtualni um koji razumije. U prilog prigovoru virtualnog uma, Cole nudi dodatni argument koji se često naziva korejska soba, varijaciju Searleovog originalnog argumenta kineske sobe.[91]

Argument korejske sobe

Cole nas poziva da zamislimo Searleovu kinesku sobu, ali ovaj put, umjesto da imamo samo pisane kineske znakove, imamo i korejske znakove. Pretpostavimo da postoje i dva priručnika, jedan za manipuliranje kineskim znakovima, a drugi za manipuliranje korejskim znakovima. Mogli bismo zamisliti da sustav javlja Searlu u sobi koji da priručnik koristi za odgovor ovisno je li input na kineskom ili na korejskom (primjerice, ako je input na kineskom upali se lampica kod kineskog priručnika i obrnuto). Izvan sobe nalaze se kineski i korejski govornici koji postavljaju pitanja. Pretpostavimo, na primjer, da je korejski priručnik nešto lošije napisan od kineskog tako da kod čitanja korejskih odgovora, ljudi izvan sobe zaključuju da ih piše neobrazovani dječak, dok kod čitanja kineskih odgovora, zaključuju da ih piše vrlo dobro obrazovana starija kineska žena. Ljudima izvan sobe se opravdano čini da su u sobi dvije osobe, korejski dječak i kineska žena. Cole zaključuje da Searle, pokretanjem takvog programa, stvara dva virtualna uma; jedan koji razumije kineski i jedan koji razumije korejski. Nijedan od njih ne razumije engleski tako da nema nikakve veze s onim što se događa u Searleovom umu.[92]

Cole, pomoću primjera korejske sobe, u kojoj se pokretanjem prikladno strukturiranog računalnog programa mogu proizvesti odgovori na kineskom, ali i na korejskom, pokazuje da um koji razumije nije niti um osobe u sobi niti sustav koji se sastoji od osobe i programa. S obzirom na to da kineski odgovori mogu naizgled prikazati različita znanja, sjećanja, vjerovanja i želje od korejskih odgovora, imamo bihevioralni dokaz da postoje dva nejednaka uma (jedan koji razumije samo kineski, a drugi koji razumije samo korejski). Budući da umovi mogu imati međusobno isključive psihološke osobine, ne mogu biti jednaki i *ipso facto*, ne mogu biti identični s umom operatera u sobi. Analogno s ovim primjerom, video igra

[91] Cole, D., 1991. *Artificial Intelligence and Personal Identity*, Synthese, 88. str. 399–417.
[92] Ibid.

može uključivati lika s jednim skupom kognitivnih sposobnosti (pametan, razumije kineski), kao i drugog lika s na nespojivim setom kognitivnih sposobnosti (glup, engleski monolingvist). Ove nekonzistentne kognitivne osobine ne mogu biti osobine računala, Playstationa ili Xbox sustava koji ih realizira. Čini se da je implikacija ovog primjera ta da su umovi generalno apstraktniji od sustava koji ih realiziraju.

Ukratko, srž prigovora virtualnog uma je ta da, s obzirom na to da dokaz koji Searle pruža da nema razumijevanja u scenariju kineske sobe je taj da on ne bi razumio kineski u sobi, argument kineske sobe ne može opovrgnuti drugačije formuliranu, a jednako jaku tvrdnju UI da računalo stvara virtualni um koji razumije kineski. Maudlin[93] kaže kako Searle nije adekvatno odgovorio na kritike, međutim, drugi autori odgovorili su na prigovor virtualnog uma, uključujući matematičkog fizičara Rogera Penrosea. Penrose generalno simpatizira pitanja koja Searle povlači s argumentom kineske sobe te i sam argumentira protivno prigovoru virtualnog uma. Penrose ne vjeruje da računalni procesi mogu objasniti svijest, što zbog argumenta kineske sobe, što zbog ograničenja formalnih sustava koje je otkrio Kurt Gödel teoremom nepotpunosti. U svom članku *Consciousness, Computation, and the Chinese Room* iz 2002. godine, koji se posebice odnosi na argument kineske sobe, Penrose tvrdi da varijacija kineske dvorane s dvoranom veličine Indije i s Indijcima koji izvode procesiranje, pokazuje da je vrlo neplauzibilno tvrditi da postoji neka vrsta bestjelesnog razumijevanja povezana s osobom koja pokreće program, a čija se prisutnost ne sudara na bilo koji način s njenom sviješću„ (230-1). Penrose zaključuje da argument kineske sobe pobija jaku UI.[94]

4.3. Sintaksa nije semantika

Stekao sam dojam da se većina prigovora Searleovoj kineskoj sobi fokusira na strukturu misaonog eksperimenta umjesto na temeljno načelo koje Searle pokušava ilustrirati, a to je da „sintaksa nije semantika“. Računalni program, čak i najnapredniji zamisliv, manipulira simbolima prema skupu sintaktičkih pravila, bez obzira na njihovo značenje. Searle vjeruje da ovo načelo objašnjava neuspjeh kineske sobe da stvori razumijevanje. Računalne operacije su „formalne“ u smislu da reagiraju samo na fizički oblik nizova simbola, a ne na njihovo značenje. S druge strane, umovi imaju stanja sa značenjem; s mentalnim sadržajem. Riječima

[93] (Maudlin, 1989.)
[94] (Cole 2014.)

i znakovima u jeziku mi pripisujemo značenje, a odgovore dajemo ovisno o njihovu značenju, a ne fizičkom izgledu. Ukratko, mi razumijemo. Prema Searleu, ključna točka je činjenica da sintaksa sama po sebi nije dovoljna za semantiku.[95] Formalni simboli sami po sebi ne mogu biti dovoljni za mentalni sadržaj, jer simboli, po definiciji, nemaju značenje (ili semantiku) osim u slučaju kad im ga netko izvan sustava daje. Dakle, iako računala mogu manipulirati sintaksom kako bi proizvela odgovarajuće odgovore na pitanja prirodnog jezika, računala ne razumiju rečenice koje dobivaju ili daju jer ne mogu riječima pridati značenja. Ideja da sintaktička manipulacija nije dovoljna za proizvodnju značenja predstavlja problem koji ima šire implikacije od UI i definicije razumijevanja. Neke teorije uma smatraju da je ljudska spoznaja uglavnom računalna. U jednom obliku, drže da razmišljanje uključuje operacije nad simbolima na temelju njihovih fizičkih svojstava. Ako je Searle u pravu, ne samo da je jaka UI pogrešna, već su pogrešni i takvi pristupi razumijevanju.[96]

Searle smatra kako scenarij kineske sobe pokazuje da se semantika ne može dobiti samo iz sintakse. Na primjer, u formalnim logičkim sustavima, pravila za sintaksu unaprijed su poznata i čini se kako je postupak računanja prilično neovisan o semantici. Svatko tko je proučavao formalnu logiku zna da su zakoni poput De Morganovih teorema ili zakona idempotencije (na primjer $x \wedge x = x$) neovisni o značenju simbola koje obrađuju. Logičari određuju osnovni skup simbola i određena pravila za manipulaciju istih. Ta su pravila čisto formalna ili sintaktička te se primjenjuju na simbole isključivo prema njihovoj sintaksi ili obliku. Semantika se mora zasebno dodati. Semantika je, dakle, sasvim neovisna o sintaksi i ne može se dobiti iz same sintakse.[97]

Kada se radi o formalnom logičkom sustavu, Searleova je tvrdnja istinita. Međutim, kada se iz formalnog logičkog sustava preselimo na računalni, situacija postaje složenija. Kao što je Cole napomenuo, računalo koje pokreće program nije istovjetno samoj sintaksi. Računalo je veliki kauzalni sustav koji mijenja stanja u skladu s programom. Stanja su sintaktički određena od strane programera, ali su u osnovi stanja složenog kauzalnog sustava koji je ugrađen u stvarni svijet. To se sasvim razlikuje od apstraktnih formalnih sustava koje logičari proučavaju. Daljnje komplikacije nastaju zbog toga što nije sasvim jasno obavljaju li računala sintaktičke operacije na isti način kako ih obavlja čovjek, točnije, nije jasno razumije li računalo sintaksu ili sintaktičke operacije. Računalo ne zna da manipulira nizovima nula i

[95] (Searle 1984. str. 39)
[96] (Cole 2014.)
[97] Ibid.

jedinica. Za razliku od osobe unutar kineske sobe, računalo ne prepoznaje da binarni podatkovni nizovi imaju određeni oblik pa se na njih mogu primijeniti određena sintaktička pravila. Unutar računala ne postoji ništa što doslovno čita ulazne podatke ili "zna" što su simboli. Umjesto toga, postoje milijuni tranzistora koji mijenjaju stanja, a sekvence napona uzrokuju izvođenje operacija. Mi (ljudi) možemo tumačiti napone kao binarne brojeve i promjene napona kao sintaktičke operacije, ali računalo svoje radnje ne tumači tako. Stoga se računalo možda ne treba prebaciti iz sintakse na semantiku, kao što to misli Searle, već se mora pomaknuti iz složenih kauzalnih veza na semantiku.[98] Nisam u potpunosti siguran je li Searle pokazao da se pokretanjem programa ne stvara razumijevanje. Iako se slažem sa Searleovom tvrdnjom da se semantika ne može dobiti samo iz sintakse, smatram da je pravo pitanje može li se iz masivnog elektroničkog kauzalnog sustava dobiti razumijevanje i značenje? To je teško pitanje, možda tako teško kao što je i pitanje kako se iz neuronskih paljenja u ljudskom mozgu proizvodi razumijevanje, značenje, um i svijest. Siguran sam da će nam znanost jednog dana dati odgovor na ova pitanja.

4.4. Simulacija nije duplikacija

Drugo važno načelo u argumentu kineske sobe je razlika između duplikacije i simulacije ili, drugim riječima, između stvarnih i simuliranih stvari. Searle tvrdi da postoji ključna razlika između simulacije i duplikacije. Kao što je pogrešno tvrditi da je računalna simulacija oluje prava oluja ili računalna simulacija probave prava probava, jednako je pogrešno zamijeniti računalnu simulaciju razumijevanja s razumijevanjem.[99]

Na prvi pogled to se čini istinitim, međutim, pojavljuju se dva problema. Nije jasno možemo li uvijek činiti razliku između simulacije i duplikacije. Na primjer, je li umjetno srce simulacija srca ili je ono funkcionalni duplikat srca izrađen od drugačijih materijala nego pravo srce? Je li nožna proteza simulacija ili duplikacija prave noge? Hodaju li osobe s umjetnim udovima ili simuliraju hodanje? Hodaju li roboti? Ako smo skloni reći da su ovi primjeri funkcionalni duplikati, a ne simulacije srca i nogu, zašto ne bismo rekli da je umjetna inteligencija duplikacija ljudskog uma, a ne simulacija? Postoji veliki i brzo rastući broj dokaza koji upućuju na to da je ljudski mozak neka vrsta "komputacijskog stroja" koji bi se,

[98] Ibid.
[99] (Searle 1984. str. 37-38)

barem u načelu, mogao od temelja rekonstruirati. Pretpostavimo da smo doslovno stvorili kopiju nečijeg mozga, atom po atom, koristeći vrlo naprednu tehnologiju. Ima li ovaj novi mozak "um"? Razumije li? Je li svjestan? Smatram da je očigledan odgovor na ova pitanja "da", barem ako se slažemo da ljudski mozak ima ove osobine. Ako su fizički identični, zašto bi se onda razlikovali od prirodno razvijenih ljudskih mozgova? Ovo povlači nekoliko pitanja, ponajprije u kojoj točki nastaju mentalne stvari i što su one točno, te u kojem trenutku simulacija postaje duplikacija?

Postoji još jedan problem sa simulacijom i duplikacijom, koji proizlazi iz procesa evolucije. Searle smatra da su intencionalnost i razumijevanje svojstva samo određenih bioloških sustava i vjerojatno proizvodi evolucije te da računala ova svojstva mogu samo simulirati. Istodobno, u scenariju kineske sobe, Searle tvrdi da sustav može prikazati ponašanje koje je jednako složeno kao i ljudsko, simulirajući svaki stupanj inteligencije i jezičnog razumijevanja koji se može zamisliti i simulirajući svaku sposobnost suočavanja sa svijetom, ali i dalje bez razumijevanja. On također kaže da bi takvi kompleksni bihevioralni sustavi mogli biti implementirani s vrlo jednostavnim materijalima, na primjer, s vodenim cijevima i ventilima (odgovor prigovoru simulatora mozga). Osim problema drugih umova, ovdje postoji i biološki problem. Iako mi možemo pretpostaviti da drugi ljudi imaju umove, evolucija ne čini takve pretpostavke. Selekcijske snage koje pokreću biološku evoluciju odabiru na temelju ponašanja. Evolucija može odabrati sposobnost kreativnog i inteligentnog korištenja informacija o okolišu dok god se ona očituje u ponašanju organizma. Pod uvjetom da se prirodna selekcija vrši protivno ponašanju (odbacujemo biheviorizam), tada ona ne brine o tome hoće li entitet razumjeti ili ne. Ako nema razlike u ponašanju, u bilo kojem nizu okolnosti, između sustava koji razumije i onoga koji ne razumije, evolucija ne može odabrati izvorno razumijevanje. Tada se čini da umovi, koji doista razumiju i posjeduju sintaksu, nemaju prednost nad entitetima koji samo obrađuju informacije koristeći čisto računalne procese. Stoga, ako se simulacija razumijevanja može jednako dobro biološki prilagoditi svijetu kao što to može i istinsko razumijevanje, pitanje je kako i zašto bi se sustavi s "istinskim" razumijevanjem mogli razvijati. Cole zaključuje da izvorna intencionalnost i istinsko razumijevanje postaju epifenomenalni.[100]

[100] (Cole 2014.)

4.5. Završna razmatranja

Mislim da Searleov argument kineske sobe otvara mnoga zanimljiva pitanja te daje ideju o tome što stvarno znači biti čovjek. Za sada ne vidim protuargument koji ga može pobiti, iako smatram da njegova istinitost uopće ne blokira potragu za umjetnom inteligencijom, već ju zapravo jača. Osobno se slažem sa Searleom i njegovim misaonim eksperimentom kojim dokazuje kako niti jedno računalo ne može razumjeti kineski samim pokretanjem programa. Mislim da je logika argumenta besprijekorna. Međutim, u isto vrijeme, ne mislim da argument kineske sobe odbacuje u potpunosti TT. To se može činiti paradoksalnim, ali to je zapravo jednostavno kao da uzmemo da je prolazak TT-a prvi korak ka umjetnoj inteligenciji, a ne njezin konačni cilj.

Smatram da niti jedan od prigovora argumentu kineske sobe nije adekvatno pokazao da je Searleov misaoni eksperiment pogrešan. Čini mi se da je Searle upio pokazati nekoliko važnih načela pomoću kojih odbacuje sve prigovore upućene argumentu. Dva temeljna načela argumenta kineske sobe su; da sintaksa nije semantika, odnosno da simulacija nije duplikacija. Čak i ako procese u mozgu možemo simulirati (kao što je to slučaj s vodovodnim cijevima koje oponašaju paljenje sinapsi u mozgu), to nije dovoljno za proizvodnju sintakse i mentalnih stanja. Odgovori koje osoba u sobi daje ne predstavljaju njeno znanje, već su samo reakcije na input. Primjerice, osoba u kineskoj sobi može dati identičan odgovor kao i izvorni kineski govornik, iako ona zapravo ne razumije značenje dobivenog pitanja, a ni odgovora kojeg je dala.

Kada bih morao izabrati jedan prigovor za koji smatram da predstavlja najveću prijetnju argumentu kineske sobe, izabrao bih onaj kojeg Searle najbrže odbacuje; prigovor drugih umova. Odgovor koji Searle daje na prigovor drugih umova je pomalo razočaravajuć jer on samo odbacuje uvjete ovog argumenta (iako prigovor drugih umova čini isto, ali doći ću i do toga). Searle kaže da prigovor drugih umova nije primjenjiv na njegov argument jer nema smisla pitati se jesu li ljudi svjesni, već je to pretpostavka koju uzimamo zdravo za gotovo. Kako bi se mogli baviti kognitivnom znanošću, moramo priznati postojanje mentalnih stanja kod ljudi kao i postojanje fizičkih objekata u svijetu. Ako prihvatimo ove pretpostavke, jasno je da prigovor drugih umova nema osnove za nastavak rasprave, međutim smatram da je to još uvijek najjači prigovor koji se trenutno može uputiti argumentu kineske sobe. Searle na isti način napada prigovor drugih umova kao što prigovor drugih umova napada njegov argument. Naime, dok prigovor drugih umova pretpostavlja da ne možemo znati subjektivno

iskustvo drugog entiteta, Searle pretpostavlja da možemo. Mislim da je ovdje riječ o Searleovom lukavom načinu da zaobiđe ovaj prigovor bez da adekvatno brani svoje pretpostavke. Ipak, kao što je najlakši način za odbacivanje prigovora drugih umova samo pretpostaviti da su ljudi svjesni, da imaju mentalna stanja i da razumiju, najlakši način za odbacivanje argumenta kineske sobe je pretpostaviti da ne možemo znati jesu li drugi ljudi svjesni osim ako ne promatramo njihovo ponašanje. No to bi nas u konačnici moglo odvesti u solipsizam; stav da svijet i drugi umovi ne postoje, što nikako ne želimo.

Iako vjerujem da su svi ostali Searleovi odgovori prigovorima argumentu kineske sobe dobro potkrijepljeni i potpuno valjani, postoji jedan prigovor koji bi u budućnosti možda mogao ugroziti argument kineske sobe. Prigovor napretka tehnologije tvrdi kako postoji mogućnost da jednog dana čovjek napravi računala koja posjeduju ljudsku razinu svijesti, a kao problem navodi činjenicu da Searle u obzir uzima samo računala bazirana na sadašnjoj tehnologiji. Searle tvrdi da bez obzira na to kako je računalo realizirano (kao stroj u obliku mozga, vodovodni sustav koji simulira neurone i sinaptičke veze ili osoba u sobi), ono nema sposobnost imati ili steći mentalna stanja i intencionalnost te stoga nema mogućnost razumijevanja kao čovjek. Iako je moja prva pomisao da se u potpunosti složim sa Searleom, moram se sjetiti da ideja računala nije uopće postojala prije stotinu ili nešto manje godina. Takvo što nitko nije mogao shvatiti, sve dok nije nastalo. Dakle, dopuštam si da vjerujem da tehnologija može imati sposobnost nadmašiti bilo koje od naših sadašnjih očekivanja. Budući da argument kineske sobe pokušavamo objasniti pomoću ideje tehnologije koja je sada dostupna, Searleov argument i dalje stoji. U principu, prigovor napretka tehnologije i Searleov argument kineske sobe, međusobno se ne isključuju te stoga mogu biti uspješni i neovisni jedan o drugom. Ovaj prigovor ne poništava Searleov argument da niti jedno računalo, kako je sada definirano, ne može samo po sebi biti dovoljno za stvaranje ljudskog razumijevanja. Ljudsko razumijevanje uključuje i semantiku te postoji jasna distinkcija između simulacije i duplikacije. Međutim, prigovor napretka tehnologije jasno pokazuje da stvaranje računala s istinskim razumijevanjem nije izvan mogućnosti. Ovo je važno zbog toga što bi, bez vjerovanja da jednog dana možemo stvoriti računalo koje razumije, imali malo razloga da nastavimo proučavati umjetnu inteligenciju.

Sa sigurnošću mogu zaključiti da niti jedan argument koji se bavi prirodom misli, spoznaje i razumijevanja umjetne inteligencije nije stvorio toliku razinu negodovanja i prigovora kao što je to učinio argument kineske sobe. Mislim da je razlog tome taj što ovaj argument pokazuje koliko je problematično točno odrediti što je razumijevanje te pokazuje koliko je čudan odnos

između stroja i naših mentalnih iskustava. Dok Searle tvrdi da smo i mi strojevi, ne pokušava objasniti kako je moguće da razumijemo kao strojevi. Bez objašnjenja prirode ljudskog razumijevanja možda nismo spremni donositi odluke o razumijevanju drugih entiteta.

Searleov misaoni eksperiment kineske sobe otvara mnoga pitanja te upućuje na ozbiljne probleme s kojima se suočavamo u razumijevanju značenja i umova. O tim se problemima kontinuirano raspravlja i vjerojatno će još dugo vremena proći prije nego što se oni riješe. Najveći je problem što se ovi problemi ne mogu riješiti sve dok ne nastane konsenzus o prirodi značenja, njegovom odnosu prema sintaksi i prirodi svijesti. I dalje postoje značajna neslaganja o tome koji procesi stvaraju značenje, razumijevanje i svijest. Iako postoji mala vjerojatnost da se ti teški problemi mogu riješiti jednostavnim misaonim eksperimentima, ipak, u najmanju ruku, argument kineske sobe zajedno s prigovorima predstavlja važan doprinos kognitivnoj znanosti u shvaćanju ljudskog mozga, uma, svijesti i općenito umjetne inteligencije.

5. Hubert Dreyfus i utjelovljena inteligencija

Posljednje poglavlje ovog diplomskog rada koncipirano je kao prikaz Hubert L. Dreyfusove kritike umjetne inteligencije kroz osvrt na knjigu *What Computers Can't Do: The Limits of Artificial Intelligence*. Prije samog izlaganja Dreyfusove kritike klasične teorije umjetne inteligencije, s ciljem da se ista lakše shvati, ubacio sam odjeljak o racionalizmu u kojem su, kroz kratku povijest misli ove popularne filozofske pozicije, prikazane tri karakteristične ideje racionalizma koje su preuzete u klasičnoj teoriji umjetne inteligencije. Zatim predstavljam Dreyfusovo briljantno filozofsko izlaganje o četiri skrivene pretpostavke pomoću kojih se pogrešno tumači inteligencija, nakon kojeg slijedi drugačiji pokušaj pristupa istraživanju o inteligenciji bez tih pretpostavki. Ovdje se rad najviše fokusira na treći dio knjige *What Computers Can't Do* i na prikaz Dreyfusove alternativne teorije inteligencije kroz radikalnu trostruku tezu da je ljudska inteligencija utjelovljena, da su inteligentna tijela situirana te da je svijet u suštini ljudski. Smatram kako potonje predstavlja najveći doprinos Dreyfusovog rada u shvaćanju ljudske inteligencije i mogućem napretku u razvoju umjetne inteligencije te sam iz tog razloga nastojao da čini i najvažniji dio ovog poglavlja. Konačno, u zadnjem sam odjeljku procijenio pouzdanost i utjecaj Dreyfusove kritike.

5.1. Racionalizam i simbolička umjetna inteligencija

Ideje o prirodi inteligencije koje promiče simbolička UI ponekad su opisane kao inovativne, no Dreyfus naglašava da je to samo najnovija reinkarnacija drevnog pogleda koji se općenito naziva racionalizam i povremeno se pojavljuje u povijesti misli. Platon, koji je postavio temelje za tu teoriju, predložio je da se mudrost sastoji od sposobnosti formuliranja znanja u eksplicitne definicije,[101] a smatrao je da ljudska bića, čije se ponašanje temelji na naučenim sposobnostima ili intuiciji, nemaju vrijednosti. On je vjerovao u mogućnost otkrivanja sustava teoretskih i objektivnih načela koji se, baš kao i temeljni aksiomi geometrije, mogu koristiti za opravdavanje ponašanja i objašnjenja stvarnosti na racionalnoj osnovi.

Najvažniji predstavnik ove ideje u moderno doba bio je René Descartes, koji je u sedamnaestom stoljeću tvrdio da se svaki problem može podijeliti na jednostavne i nezavisne

[101] Dreyfus, H. L. 1972. *What Computers Can't Do*, New York, Harper & Row, Intorduction xv-xvii

elemente te da svaka složena situacija ili misao može biti razjašnjena otkrivanjem sustava pravila koja reguliraju na koji je način ta situacija ili misao izgrađena iz jednostavnih elemenata. Tvrdio je da čak i ljudski um djeluje u skladu s takvim pravilima i jednostavnim elementima.[102] Nakon Descartesa, ova koncepcija prirode inteligencije bila je prisutna i u drugim racionalističkim istomišljenicima; mislioci kao što su Leibniz, Kant i Husserl, a u nešto manjoj mjeri i empiričari poput Lockea i Humea, te u novije vrijeme znanstvenici poput lingvista Noama Chomskog, psihologa Jerrya Fodora, kao i različiti predstavnici simboličke UI dijelili su i dijele ovu koncepciju prirode ljudske inteligencije.

Prema Dreyfusu, u simboličkoj UI mogu se naći tri karakteristična pojma racionalizma. Prvi je pojam nazvao psihološka pretpostavka, koji je već spomenuta pretpostavka da je ljudska inteligencija pitanje manipulacije simbolima prema formalnim pravilima. Ova pretpostavka daje teoretsku osnovu tvrdnji da računala mogu biti programirana da razmišljaju poput ljudskih bića. Primjerice, ona je eksplicitna točka polaska pristupa kognitivne simulacije. Iako joj je većina sklona, ovu pretpostavku ne prihvaćaju baš svi istraživači UI. Sve varijacije simbolične UI, međutim, prihvaćaju drugu, epistemološku pretpostavku, da je svo znanje moguće formalizirati, da se sve što mogu razumjeti ljudska bića može izraziti pomoću kontekstno-neovisnih, formalnih pravila ili definicija. Ako je to istina, ta bi pretpostavka jamčila uspjeh ideje simboličke UI čak i u slučaju da je psihološka pretpostavka lažna, jer bi formalizirana verzija neformalnog ljudskog znanja i ponašanja imala istu kognitivnu vrijednost kao i neformalizirana originalna verzija. Time računalo možda ne bi bilo u poziciji da simulira ljudske misaone procese, ali bi bilo u stanju reproducirati inteligentna ljudska ponašanja.

I epistemološka i psihološka pretpostavka često se temelje na ontološkoj pretpostavci da stvarnost, utoliko što može biti poznata ljudima, ima formaliziranu strukturu koja je izgrađena od niza objektivnih, determiniranih elemenata, od kojih je svaki nezavisan od drugih. Kada stvarnost ne bi imala takvu strukturu, bilo bi vrlo malo vjerojatno da se može spoznati uz pomoć skupa kontekstno-neovisnih, formalnih pravila čije je postojanje uzeto zdravo za gotovo, kako u epistemološkoj tako i u psihološkoj pretpostavci.

Činjenica da se klasična UI temelji na racionalističkom poimanju inteligencije, koje pripada povijesti filozofije, ne bi nam bila od interesa da povijest ne uključuje i važne filozofske

[102] (Dreyfus 1972. str. 89.)

kritike te pozicije. Vidljivo je da je Dreyfus pod jakim utjecajem ovih antiracionalističih kritika, pogotovo onih Heideggera, Merleau-Pontyja i Wittgensteina, čije argumente upotrebljava u svojoj kritici simboličke UI i njenih triju središnjih pretpostavki, kao i u svom razvoju alternativne teorije inteligencije.

5.2. Četiri pretpostavke simboličke UI i Dreyfusova kritika

Dreyfusova kritika simboličke UI se odnosi na ono što on smatra da su četiri primarne pretpostavke simboličkog istraživanja UI.[103] U prethodnom sam poglavlju spomenuo tri pretpostavke koje je simbolička UI na neki način preuzela od racionalizma. To su psihološka, epistemološka i ontološka pretpostavka. Psihološka pretpostavka je da um radi tako što izvodi diskretna računanja (u obliku algoritamskih pravila) na zasebnim reprezentacijama ili simbolima. Dreyfus tvrdi da plauzibilnost psihološke pretpostavke polazi od druge dvije pretpostavke: epistemološke i ontološke. Epistemološka pretpostavka je da se sva aktivnost, bilo živog ili neživog objekta, može matematički formalizirati u obliku predviđenih pravila ili zakona. Ontološka pretpostavka je da se stvarnost u potpunosti sastoji od skupa međusobno nezavisnih, atomskih nedjeljivih činjenica. Zbog epistemološke pretpostavke znanstvenici i teoretičari na polju UI argumentiraju da je inteligencija isto što i formalno slijeđenje pravila, a zbog ontološke tvrde da se ljudsko znanje u potpunosti sastoji od unutarnjih reprezentacija stvarnosti. Dva znanstvena polja koja bi mogla pružiti dokaz u prilog psihološkoj pretpostavci simboličke UI su psihologija i neuroznanost. Neuroznanost je važna jer smatra da je razmišljanje uz pomoć pravila i simbola moguće samo ako se ta pravila i simboli u ljudskom mozgu provode na isti način na koji se računalni program provodi od strane hardvera računala. Pretpostavku da je mozak analogan računalnom hardveru, a um analogan računalnom softveru Dreyfus naziva biološkom pretpostavkom i ona predstavlja četvrtu pretpostavku koja se često pojavljuje u simboličkoj UI.

Imamo, dakle, četiri pretpostavke koje su temelj klasične paradigme umjetne inteligencije i zbog kojih se smatra da ljudska inteligencija ovisi o manipulaciji simbolima:[104]

[103] (Dreyfus 1972. str. 68.)
[104] Ibid.

1. **Biološka pretpostavka:** mozak obrađuje informacije diskretnim operacijama putem nekog biološkog ekvivalenta on/off prekidača.

2. **Psihološka pretpostavka:** um se može promatrati kao uređaj koji radi na bitovima informacija u skladu s formalnim pravilima.

3. **Epistemološka pretpostavka:** svo znanje može biti formalizirano.

4. **Ontološka pretpostavka:** svijet se sastoji od neovisnih činjenica koje se mogu prikazati neovisnim simbolima.

Dreyfus kaže da su te pretpostavke uzete od strane znanstvenika UI kao zdravo za gotovo, kao aksiomi koji jamče rezultate, a u stvari su samo hipoteze koje moraju biti ispitane i dokazane.[105] U nastavku slijedi Dreyfusova kritika pojedinih pretpostavki.

5.2.1. Kritika biološke pretpostavke

U ranim danima istraživanja neurologije, znanstvenici su shvatili da neuroni u mozgu ili šalju impulse ili ih ne šalju. Takvi „ili-ili“ „sve ili ništa“ impulsi slični su binarnom jeziku. Dakle, mozak računa koristeći nešto poput binarnih signala. Umjesto jedinica i nula, ili „on“ i „off“ naredbi, mozak koristi „šalji“ ili „ne šalji“ naredbe pri otpuštanju impulsa u neuronima. Nekoliko istraživača, kao što su Walter Pitts i Warren McCulloch, tvrdili su da neuroni funkcioniraju slično kao Booleov logički sklop pa se mogu oponašati elektroničkim sklopom na razini neurona. Kada su se digitalna računala počela koristiti u ranim 50-im godinama, ovaj je argument proširen na to da je mozak velik fizički sustav simbola koji manipulira binarnim simbolima; nulama i jedinicama. Dreyfus je vrlo lako pobio biološku pretpostavku navodeći istraživanja u neurologiji koja su ukazala na to da je slanje impulsa neurona sličnije analognom signalu nego onom digitalnom.

Međutim, treba reći da biološka pretpostavka, iako uobičajena u četrdesetim i ranim pedesetim godinama, više nije bila prihvaćena među većinom istraživača UI u vrijeme kada je Dreyfus objavio knjigu *What Computers Can't Do*. 1970-ih godina jako je malo istraživača UI još vjerovalo u biološku pretpostavku te nitko nije argumentirao protiv Dreyfusove kritike.[106] Iako mnogi i danas smatraju da se djelovanje analognih neurona može simulirati digitalnim

[105] (Dreyfus 1972. str. 137.)

[106] (Crevier 1993. str.126.)

računalom do razumne razine točnosti (kao što su Ray Kurzweil ili Jeff Hawkins), oni ne pretpostavljaju da su neuroni esencijalno digitalni (Alan Turing je napravio isto opažanje već 1950. godine).[107]

5.2.2. Kritika psihološke pretpostavke

Dreyfus je opovrgnuo psihološku pretpostavku pokazujući da se većina onoga što "znamo" o svijetu sastoji od kompleksnih stavova ili tendencija zbog kojih naginjemo jednom tumačenju nasuprot drugom. Tvrdio je da, čak i ako koristimo eksplicitne simbole, koristimo ih u skladu s nesvjesnom pozadinom zdravorazumskog znanja, bez koje naši simboli nemaju značenje. Ova pozadina, prema Dreyfusovom mišljenju, nije implementirana u naše mozgove poput eksplicitnih individualnih simbola s eksplicitnim individualnim značenjem.

Mnogi bi se istraživači UI danas složili da se ljudsko razmišljanje ne sastoji prvenstveno od manipulacije simbola na visokoj razini. Zapravo, od kad je Dreyfus prvi put objavio svoje kritike, istraživanja UI su se općenito udaljila od ideje manipulacije simbola na visokoj razini prema novim modelima koji više uključuju naše nesvjesno rasuđivanje.[108]

5.2.3. Kritika epistemološke pretpostavke

Dreyfusova najvažnija kritika usmjerena je protiv epistemološke pretpostavke koja je u podlozi svih oblika simboličke UI. Čak i ako se složimo da je psihološka pretpostavka lažna, istraživači UI još uvijek mogu tvrditi (kao što je tvrdio John McCarthy, osnivač UI) da postoji mogućnost da računalo reproducira svo ljudsko znanje, bez obzira reproducira li to znanje na isti način kao i ljudska bića ili ne. Dreyfus je tvrdio da ne postoji opravdanje za tu pretpostavku zbog toga što postoji toliko ljudskog znanja koje nije simboličko.

Epistemološkom pretpostavkom se, dakle, tvrdi da se inteligentno ponašanje može reproducirati formaliziranjem ljudskog znanja (primjerice, kodiranjem u pravilima) na način da ga računalo može slijediti. Iako takva formalna pravila mogu biti jedan od načina da opišemo ljudsko znanje, Dreyfus smatra da ona ne mogu pružiti osnovu za reprodukciju takvog znanja. Posjedovanje znanja, Dreyfus ističe, podrazumijeva sposobnost primjene tog znanja u relevantnim situacijama – u razmišljanju, komunikaciji i ponašanju. Ideja da je vatra vruća, primjerice, podrazumijeva da možemo primijeniti to znanje u odgovarajućim trenucima

[107] Turing 1950. u "(7) Argument from Continuity in the Nervous System." str. 451.
[108] (Crevier 1993. str.126.)

kada razmišljamo i komuniciramo s nekime o vatri ili kada smo u interakciji s njom; ako se to nije dogodilo ne možemo istinski reći da je to znanje bilo prisutno.

5.2.4. Kritika ontološke pretpostavke

Prisjetimo se, ontološkom se pretpostavkom tvrdi da se svijet, odnosno naša stvarnost, u potpunosti sastoji od skupa međusobno nezavisnih i nedjeljivih činjenica koje se mogu prikazati nezavisnim simbolima. Ovdje je Dreyfus također identificirao suptilniju pretpostavku o svijetu. istraživači UI (kao i futuristi i pisci znanstvene fantastike) često pretpostavljaju da ne postoji ograničenje formalne znanstvene spoznaje jer pretpostavljaju da se bilo koja pojava u svemiru može opisati simbolima ili znanstvenim teorijama. Time se pretpostavlja da sve što postoji možemo shvatiti kao objekte, svojstva objekata, klase objekata, odnosa objekata i tako dalje; to su upravo one stvari koje mogu biti opisane logikom, jezikom i matematikom. Pitanje o tome što postoji zove se ontologija i zato Dreyfus to naziva ontološkom pretpostavkom. Ako je ona netočna, onda to izaziva sumnju što u konačnici uopće možemo znati, a u tome će nam inteligentni strojevi moći pomoći.

5.2.5. Beskontekstualizam simboličke UI

Na temelju epistemološke i ontološke pretpostavke, znanstvenici i teoretičari na polju UI tvrde da je spoznaja manipulacija unutarnjih simbola pomoću internih pravila te da je zato ljudsko ponašanje u velikoj mjeri bez konteksta (engl. *context free*). Stoga je znanstvena psihologija koja detaljno određuje „unutarnja“ pravila ljudskog uma doista moguća, na isti način na koji zakoni fizike detaljno određuju “vanjske” zakone fizičkog svijeta. Ali, to je i ključna pretpostavka koju Dreyfus negira. Drugim riječima, on tvrdi da ne možemo sada, niti ćemo ikada moći, razumjeti svoje ponašanje na isti način kao što razumijemo objekte u, primjerice, fizici ili kemiji; odnosno, ne možemo smatrati ljude poput stvari čije se ponašanje može predvidjeti pomoću „objektivnih“ i „beskontekstnih“ znanstvenih zakona. Prema Dreyfusu, beskontekstna psihologija je kontradikcija u terminu.

Dreyfusovi argumenti protiv ove pozicije uzeti su iz fenomenološke i hermeneutičke tradicije, posebno iz djela Martina Heideggera. Heidegger je tvrdio, protivno kognitivističkoj poziciji na kojoj se temelji simbolička UI, da je naše biće u stvari visoko kontekstno vezano (engl. *highly context bound*), zbog čega su ove dvije beskontekstne pretpostavke lažne. Dreyfus ne poriče da, ako to želimo, možemo gledati na ljudsko ili bilo koje drugo djelovanje kao da je „regulirano zakonom“, na isti način na koji možemo gledati na stvarnost kao da se sastoji od

nedjeljivih atomskih činjenica. No velika je razlika između tvrditi nešto zato što želimo ili možemo vidjeti stvari na taj način i zaključiti da je to objektivna činjenica i da su stvari zaista takve kakve nam se čine. Zapravo, Dreyfus tvrdi da one nisu nužno takve te da će, stoga, bilo koji istraživački program umjetne inteligencije koji to pretpostavlja brzo upasti u duboke teorijske i praktične probleme. Dosadašnji su napori znanstvenika na polju klasične paradigme umjetne inteligencije osuđeni na neuspjeh, smatra Dreyfus.

5.3. Mozak, tijelo i svijet

U dosadašnjem radu izložio sam Dreyfusov oštar kritički osvrt na povijest i stanje područja umjetne inteligencije 1970-ih godina te njegovu filozofsku kritiku četiriju pretpostavki pomoću kojih simbolička UI pogrešno tumači inteligenciju. Kao što sam u uvodu i najavio, ono što slijedi u ovom poglavlju predstavlja Dreyfusov pokušaj drugačijeg pristupa istraživanju i razmišljanju o inteligenciji. Na neki način ono, u grubo, obuhvaća treći dio knjige *What Computers Can't Do.*

Prvi dio originalne knjige dobio je najviše pažnje s obzirom na to da je bio najkritičniji prema simboličkoj UI, a ujedno i najlakši za razumjeti. Isto tako, s obzirom na to da je ta kritika bila pravodobna, time je i najbrže zastarjela, iako je bila temelj raznih rasprava unutar zajednice UI do dana današnjeg. Nezgodna posljedica toga je u tome što su zanimljiviji i značajniji dijelovi knjige, drugi i treći, u većoj ili manjoj mjeri potisnuti pa čak i zanemareni. Treći je dio knjige najdublji i najnapredniji, a istovremeno najteži i najnerazvijeniji te sukladno tome, mislim da je najisplativiji za preispitati. Njegove glavne teze su: da je ljudska inteligencija u osnovi utjelovljena, da su inteligentna tijela u osnovi situirana (ugrađena u svijetu) te da je svijet u suštini ljudski. Sve se ove teze svode na istu stvar; naime, razumjeti mogućnosti inteligencije ne znači razumjeti svojstvo nekog eventualno izoliranog sustava kao što je „intelekt“ ili „um“ (intelekt + afekt), ili čak „agent“ (intelekt + afekt + tijelo). Umjesto toga, razumjeti mogućnosti inteligencije znači razumjeti veću cjelinu koja obuhvaća niz kulturnih utjelovljenih pojedinaca koji zajedno žive u već smislenom svijetu. U onome što slijedi, htio bih dati dovoljno dokaza za ovu radikalnu trostruku tezu jer smatram da njena prihvatljivost i položaj u trenutnim istraživanjima umjetne inteligencije moraju biti razmotreni ako želimo ostvariti napredak.

Za početak, jedno preliminarno pojašnjenje; tvrdnja o utjelovljenoj situiranosti nije epistemološka već ontološka tvrdnja. Ovdje se ne radi o preduvjetima teoretiziranja ili o empirijskim otkrićima, čak se ne radi ni o praktičnoj primjeni teorije (kao npr. rubnih uvjeta). „Razumjeti mogućnosti inteligencije“ znači smatrati inteligenciju autentičnim fenomenom – nešto čemu se može pristupiti istragom i objašnjenjem. Dreyfus tvrdi da ideja inteligencije bez utjelovljenog postojanja u svijetu nema smisla, pa ne samo da takva stvar ne postoji, već ona ne može postojati (to je besmislica). Ova je Dreyfusova teza prilično nezavisna i tendenciozna; ne tiče se konceptualne analize ili *a priori* rasuđivanja prvih principa i možda nije pogodna za izravnu potvrdu ili odbijanje, ali ipak je u kontinuitetu sa znanstvenim istraživanjima na isti način na koji su temeljna pitanja u kvantnoj mehanici, teoriji evolucije, ekonomiji i tako dalje, u kontinuitetu s empirijskim znanstvenim pitanjima u svojim domenama.

Većina je istraživača UI odlučilo ignorirati Dreyfusovu kritiku, ili su ga s druge strane napali za filozofiranje. Primjerice, Marvin Minsky je o Dreyfusovoj kritici i tadašnjim drugim filozofskim kritikama rekao da oni pogrešno razumiju rad istraživača UI te ih treba ignorirati.[109] Ironično, mislim da su upravo istraživača UI ti koji su češće krivi za "filozofiranje" u ovom slučaju, nego što su filozofi poput Dreyfusa. Obično je opozicija teorije situiranosti inteligencije ta koja je zasnovana na pristupu *a priori* pretpostavke (to je na neki način i Dreyfusova optužba iz drugog dijela knjige *What Computers Can't Do*). Na primjer, protivnici situiranosti mogu postaviti pitanje: Kako situiranost može biti esencijalna za inteligenciju? Uostalom, sve što znamo (a samo bi luđak tako nešto porekao) je to da se inteligencija ostvaruje u živčanom sustavu, uglavnom u neokorteksu, i da komunicira s vanjskim svijetom samo preko određenih senzora i pretvornika. Kako bi funkcionirao, prigovor se nastavlja, mozak treba metaboličku podršku ostatka organizma, a time i okoliša, ali to su „implementacijski detalji“ u smislu da su na krivoj razini opisa za analizu inteligencije same po sebi. Kada izostavimo te detalje, jasno je da inteligencija, esencijalna funkcionalna struktura koju živčani sustav implementira, može biti samo nešto što je odvojeno i od tijela i od svijeta.

Zanimljivo je koliko metafizike neprovjereno vreba u ovim, naizgled ležernim frazama. Počnimo s „esencijalnom i funkcionalnom strukturom koju živčani sustav implementira“. Koja je to struktura? Mozak, naravno, koji kao i bilo koji razumno složen fizički sustav

[109] (Crevier 1993. str. 143.)

prikazuje određeni broj različitih apstraktnih struktura koje bi mogle u principu biti identificirane i definirane zasebno izvan njega. No, kako je određeno da je inteligencija jedna od ovih struktura? Zapravo, takva stvar nikada nije određena jer se pitanje nikada nije postavilo na ovaj način. Ono što znamo je da je, u slučaju ljudskih bića, nešto u strukturi našeg mozga presudno za našu inteligenciju. Znamo to jer, ako se nečiji mozak fizički ošteti, ošteti se i njegova inteligencija. No ta činjenica ne dokazuje ništa o tome da je sam mozak, ili primjerice neka struktura u području mozga, samostalno dovoljna za inteligenciju kao razumljiv fenomen. Ono što zadovoljava kriterij za samostalan razumljiv sustav ili podsustav, ovisi o aspektima ili „razini“ na kojoj ga treba shvatiti. Na primjer, procesor ili CPU (engl. *central processing unit* – centralna procesorska jedinica) stolnog računala je iz mnogih točki gledišta savršeno razumljiv samostalni podsustav, no ako netko želi razumjeti sustav kao program za obradu teksta, onda se ne može usredotočiti samo na procesor jer je obrada teksta kao takva razumljiva samo kao karakteristika većeg, sveobuhvatnog sustava. Napomena: poanta nije samo da se mora opisati na višoj razini, nego je procesor jednostavno krivo mjesto za tražiti tako nešto – procesor je prejednostavan da bi se razumjela obrada teksta na bilo kojoj razini (iako se ovdje ne tvrdi da procesor nije presudna komponenta, u smislu da relevantni sustav ne može funkcionirati bez njega). Isto tako, mozak može biti, iz neke točke gledišta, savršeno razumljiv samostalni podsustav, ali ako je Dreyfus u pravu, ljudska je inteligencija razumljiva samo kao obilježje većeg sveobuhvatnog sustava; sam je mozak krivo mjesto za tražiti inteligenciju (na bilo kojoj razini opisa). Doduše, ni cijeli organizam pojedinca ne mora biti, sam po sebi, dovoljno sveobuhvatan.

Kako netko može razviti i braniti takvu točku? Koje su vrste razmatranja relevantne za utvrđivanje opsega i granica inteligentnih sustava? Dreyfus piše:

> “Generally, in acquiring a skill-in learning to drive, dance, or pronounce a foreign language, for example-at first we must slowly, awkwardly, and consciously follow the rules. But then there comes a moment when we finally can perform automatically. At this point we do not seem to be simply dropping these same rigid rules into unconsciousness; rather we seem to have picked up the muscular gestalt which gives our behavior a new flexibility and smoothness. The same holds for acquiring the skill of perception.”[110]

[110] (Dreyfus 1972. str. 160-161.)

„Muscular gestalt“?[111] Kakve veze imaju mišići? Možemo reagirati pitanjima poput ovih, ali to je, smatra Dreyfus, zbog vrlo zavodljive tradicionalne priče. Kad mi djelujemo inteligentno, naš racionalni intelekt (svjesno i/ili nesvjesno) uzima u obzir različite činjenice koje ima na raspolaganju i pokušava otkriti što treba učiniti, a zatim izdaje odgovarajuće upute. Te se upute onda preračunavaju pomoću izlaznih pretvornika u fizičke konfiguracije (mehaničke snage, električne struje, kemijske koncentracije...) koje za posljedicu imaju određeno tjelesno ponašanje. Pretvarači funkcioniraju kao neka vrsta sučelja između racionalnog i fizičkog i, kao takvi, pružaju prirodnu točku subdivizije u smislu da bilo koji drugi izlazni podsustav, koji je odgovorio istim uputama s istim ponašanjem, može biti zamijenjen bez da se dogodi bilo kakva bitna razlika u intelektualnom dijelu. Prema toj teoriji, mišići spadaju u fizičku razinu, a ne u inteligentan (pod)sustav.

No ima li zaista pretvarača između našeg uma i našeg tijela? Iz određene perspektive, pitanje se može učiniti banalnim – naravno da ima. Gotovo po definiciji, mora postojati određena pretvorba između simboličkog ili konceptualnog sadržaja naših umova i fizičkih procesa u našem tijelu i ta je pretvorba transdukcija. Dreyfus ovo opovrgava, ali ne negirajući da postoje umovi i da postoje tijela, već negirajući da postoji potreba za sučeljem ili pretvorbom između njih. Transdukcija je, prisjetimo se, funkcija koju je Descartes dodijelio epifizi;[112] ona je potrebna ako i samo ako su um i tijelo fundamentalno različiti, to jest, razumljivi u vlastitim uvjetima, sasvim odvojeni jedan od drugoga.[113]

Presudan trenutak je već određen u slici intelekta koji otkriva činjenice, a zatim izdaje upute. No što je uputa? Po prilično konvencionalnoj definiciji, to je sintaktički izraz koji, na temelju pripadnosti prikladno interpretirajućeg formalnog sustava, nosi određenu vrstu semantičkog sadržaja. Konkretno, njegov sadržaj ne ovisi o tome na koji način i kada će se izvršiti u nekom određenom fizičkom sustavu. Na primjer, ako odlučim upisati slovo "S" na tipkovnici, sadržaj te upute neće ovisiti o tome je li on namijenjen za moje prste ili za neku robotsku protezu koju imam ugrađenu umjesto ruke. Bilo koji izlazni sustav koji može preuzeti upute i upisati slovo

[111] Gestalt je nešto što se sastoji od mnogo dijelova, a ipak je nešto više ili drugačije od same kombinacije dijelova od kojih se sastoji. (Vidi: Sabar, S. 2013. *What's a Gestalt?*, Gestalt Review, 17(1):6-34, Gestalt International Study Center)

[112] Pinealna žlijezda ili epifiza (lat. *Glandula pinealis*) je mala endokrina žlijezda češerastog oblika smještena u mozgu, upravo na mjestu gdje se razdvajaju dvije polutke.

[113] Francuski filozof René Descartes među prvima je načeo ovu temu u svojoj knjizi *Traté de l' Homme* iz 1664. godine. Vjerovao je u tezu da upravo epifiza predstavlja poveznicu između uma i tijela.

"S " ili neku drugu uputu, uz određene prilagodbe, bio bi dobar. Ideja da postoje takve upute moralno je ekvivalentna ideji da postoje pretvornici.

Suprotni, odnosno inkompatibilni pogled s prethodnim je sljedeći. Postoje deseci milijardi neuronskih puteva koji vode iz mog mozga ili neokorteksa do različitih mišićnih vlakana u mojim prstima, rukama, zglobovima, ramenima, itd., te iz različitih *proprioceptivnih* i taktilnih stanica natrag. Svaki put kad upišem slovo na tipkovnici, znatan dio tih impulsa putuje na različitim frekvencijama i u različitim vremenskim odnosima jedan s drugim. Taj određeni obrazac, u ovom slučaju obrazac upisivanja slova "S" na tipkovnici, ovisi o mnogim drugim faktorima povrh same upute. U prvom redu, on ovisi o jačini i brzini mojih mišića, o duljini mojih prstiju, o obliku mojih zglobova i slično. Drugim riječima, u načelu ne postoji način da, primjerice, Bog mikrokirurgijom spoji moje neurone na nečije tuđe prste i ruke tako da bi ja mogao uspješno pisati njima. Set veza koje bi mogle uspješno raditi za jedno slovo u jednom položaju će se sasvim promijeniti kad ću morati upisati sljedeće slovo jer će, zbog fizičke promijene ruke koja je sada uparena s mojim obrascima nastalim na staroj ruci, nova ruka zauzeti drugačiji položaj. U tom slučaju, ono što je moj obrazac „značio“ ovisilo je o tome da je to bio posebni obrazac stvoren posebno za moje prste, tj. za prste mog „mišićnog gestalta“.

Jedna analogija u Turingovom stilu možda može pomoći da se ovo bolje shvati. Zamislimo sustav šifriranja koji se temelji na vrlo velikim ključevima za šifriranje, u kojem ispada da se sve kratke šifrirane poruke mogu usporediti po veličini sa samim ključevima (na primjer, nekoliko desetaka milijuna bitova). Sada, razmotrimo jednu takvu poruku i pitajmo se bi li ona mogla značiti išta bez svog određenog ključa za šifriranje. Teško je vidjeti kako bi mogla. Tada analogija funkcionira ovako: određeno tijelo određenog pojedinca ima svoj vlastiti „mišićni gestalt“ koji, da tako kažemo, funkcionira poput velikog ključa za šifriranje i bez kojeg su „poruke“ čista besmislica.

Ipak, čak i to može biti pretjerano optimistično. To što će određeni obrazac rezultirati upisom slova "S" ovisi i o tome koliko su moji prsti već zagrijani za tipkanje, koliko sam umoran, jesam li gladan, pišem li na prijenosnom ili stolnom računalu, odnosno koju tipkovnicu koristim i tako dalje. U različitim okolnostima, isti će obrazac dati različita slova, a različiti obrasci isto slovo. Drugim riječima, ne postoji slična struktura između obrazaca, na bilo kojoj razini opisa, koja pouzdano korelira s radnjom koju obrazac proizvodi. Razlog zbog kojeg mogu tipkati, bez obzira na sve to, je taj da postoje podjednako bogati aferentni obrasci koji

tvore neku vrstu povratne petlje koja stalno podešava sustav (u smislu gore navedene analogije, to je kao da se ključ za šifriranje stalno mijenja u realnom vremenu). Ali to bi značilo da „sadržaj“ u bilo kojem neuronskom izlaznom obrascu ovisi ne samo o određenom tijelu na koje je spojen, već i o konkretnim detaljima trenutne situacije. Ako prihvatimo ovo, nužno moramo odbaciti ideju definiranih uputa i zamjenjivih pretvarača. No bez koherentnog pojma mentalne/fizičke transdukcije, granica, a time i sama razlika između uma i tijela, počinje lagano nestajati.

Dreyfus, međutim, želi otići i korak dalje kada tvrdi da je razlika između nas i našeg svijeta isto tako pod sumnjom. Baš kao što se mozak može identificirati kao odvojen od ostatka ljudskog organizma za određene namjene, tako se i organizam može identificirati kao odvojen od njegove okolice. Pitanje nije je li površina kože vidljiva, već iznosi li ta površina važnu podjelu ili sučelje kada želimo shvatiti ljudsku inteligenciju. Vraćamo se na ono što čini velik dio knjige *What Computers Can't Do,* a to je napad na „klasičnu paradigmu umjetne inteligencije“ ili „simboličku UI“ prema kojoj interne simboličke reprezentacije predstavljaju područje smislenosti u kojem inteligencija obitava.

Sada Dreyfus, kao što već znamo, naglašeno odbacuje primat unutarnjih simboličkih reprezentacija, međutim, on sa svojim protivnicima dijeli uvjerenje da inteligencija obitava u smislenom. Postavlja se pitanje: Je li to područje smislenosti koje je mjesto inteligencije uopće reprezentacionalno i je li ograničeno kožom? Dreyfusov je odgovor jasan:

> “When we are at home in the world, the meaningful objects embedded in their context of references among which we live are not a model of the world stored in our mind or brain: they *are the world itself.”*[114]

Ovdje zapravo imamo dvije usko povezane teze; negativnu tezu protiv simboličke UI i pozitivnu tezu o smislenom kao takvom.

Negativna teza je jednostavno odbacivanje stajališta, gotovo sveprisutnog u simboličkoj UI, da su smisleni predmeti usred kojih inteligencija obitava, u prvom stupnju, unutarnji. „Klasični“ kognitivni znanstvenici ograničavaju ove unutarnje objekte na simboličke izraze i modele, dok su drugi nešto liberalniji kada govore o mentalnim slikama, kognitivnim mapama i raspoređenim reprezentacijama. Ali Dreyfus želi proširiti smisleno i izvan unutarnjeg;

[114] (Dreyfus 1972. 177-178.)

smisleni objekti su „sam svijet“. Ne samo da odbacuje simboličke modele, već i općenito reprezentacijsku teoriju kao što ćemo vidjeti detaljnije kada dođemo do pozitivne teze.

Ali prvo, treba pripaziti da ne dođe do nesporazuma. Svatko bi se složio da svjetovni objekti mogu biti smisleni u derivativnom smislu, kao kada im mi dodijelimo značenja koja su već prisutna „u našim glavama“. Primjerice, dvoje se vojnika mogu složiti da koriste određeni signal koji znači da neprijatelj dolazi, a taj bi signal onda doista značio to, ali samo derivativno iz njihove odluke (mnogi bi filozofi i znanstvenici i dalje smatrali da je to jedini način da vanjski objekti postanu smisleni). S druge strane, kada Dreyfus kaže da su smisleni predmeti sam svijet, on misli na njihovo originalno značenje, ne samo derivativno. To znači da inteligencija obitava „izvan“, u svijetu, a ne samo „unutra“ – suprotno kognitivnoj znanosti i klasičnoj teoriji umjetne inteligencije.

Pozitivna teza, nažalost manje razrađena u Dreyfusovoj knjizi, je o smilenosti kao takvoj. Ako mogu pokušati objasniti tu tezu svojim riječima, rekao bih, vrlo grubo, da je smisleno ono što je značajno u smislu nečega izvan sebe i podložno je normativnoj evaluaciji prema tom značenju. U tom su smislu *reprezentacije* slične paradigme smislenosti, a kad kognitivni znanstvenici govore o smislenim unutarnjim objektima, oni uvijek misle na reprezentacije (jedino između njih postoji spor o mogućim oblicima tih reprezentacija, npr. jesu li su one simboličke ili ne). Reprezentacija je značajna kada podrazumijeva određeni sadržaj izvan sebe, a smislena ako taj sadržaj reprezentira pravilno i točno. Descartes je, u stvari, izumio „unutarnje carstvo“ (engl. *inner realm*) kao spremište za kognitivne reprezentacije – reprezentacije onoga što je izvan njih. Simbolička UI nije u suštini promijenila ovo; ništa drugo osim reprezentacija nije predloženo kao unutarnje i smisleno.

Kada Dreyfus tvrdi da su smisleni predmeti u samome svijetu, on ne misli samo ili većinom na reprezentacije. Svijet ne može biti reprezentacija u svim fazama, ali to ne znači da ne može biti smislen, jer postoji više vrsta značenja nego reprezentacijskog sadržaja. Veliki broj filozofa u dvadesetom stoljeću: Dewey, Heidegger, Wittgenstein i Merleau-Ponty, samo da imenujem neke od najistaknutijih – su mnogo raspravljali o značenju opreme, javnih mjesta i prakse u zajednici, a Dreyfus je to znao i iskoristio za svoju teoriju. Čekić je, na primjer, značajan izvan sebe, u smislu onoga za što je namijenjen; zabijanje raznih vrsta čavli, razbijanje nekog materijala i tako dalje. Čavli, drvo, projekt i sam stolar su isto tako uhvaćeni u ovoj „mreži značenja“, svatko na svoj način. To su smisleni objekti koji čine sam svijet i niti jedan od njih nije reprezentacija.

Tu je očiti problem; cijela stvar ovisi o igri riječima. Naravno, čekić i slične stvari su „značajne“ pa čak i „smislene“ u smislu da su one važne za nas i ovisne o drugim stvarima za njihovu pravilnu uporabu. Ali to nije isto što i smislenost u smislu da imaju vlastiti sadržaj ili semantiku. To je i razlog zašto one nisu reprezentacije. Možemo reći da su one smislene u širem smislu, iako ne i u užem. Pravo pitanje je: Koji je smisao bitan u kontekstu razumijevanja ljudske inteligencije?

Da bi riješili ovo pitanje, prvo se moramo zapitati kakve veze smislenost ima s inteligencijom. Tvrditi da inteligencija obitava u smislenom nije isto kao i tvrditi da je okružena ili usmjerena prema smislenom, kao da su to dvije odvojene stvari. Umjesto toga, Dreyfus tvrdi da inteligencija ima svoje postojanje u smislenom kao takvom – na način na koji, primjerice, državno bogatstvo leži u njegovom produktivnom kapacitetu ili se snaga korporacije sastoji u njenoj tržišnoj poziciji. Dakle, inteligencija nije ništa drugo nego cjelokupna interaktivna i međuovisna struktura smislenih ponašanja i objekata.

Osnovnu ideju možda možemo predstaviti na sljedeći način. Inteligencija je sposobnost da se pouzdano bavimo nečim višim od onog što nam je dano i što nam je jasno. Ovo sigurno nije odgovarajuća definicija inteligencije, ali njome dolazimo do nečeg esencijalnog i posebno do nečega što ima veze sa smislenošću. Reprezentacije su očito važne za ono što nam je odsutno i skriveno, jer su one same prisutne i „zamjenjuju“ to nešto drugo što „reprezentiraju“. Kako to mogu raditi? Pojedine reprezentacije mogu funkcionirati kao takve samo sudjelovanjem u skladu s mnogim drugima u većoj i pravilno reguliranoj shemi reprezentacija. Zatim, uz pretpostavku da je sama shema „u dobroj formi“ i pravilno se koristi, sustav može posredno pratiti i istraživati odsutne i skrivene pojave prateći i istražujući dane zamjenske reprezentacije (da je shema „u dobroj formi“ znači da taj sistem u pravilu dobro radi). Struktura postojećih reprezentacija zajedno s onom same sheme, „kodira“ nešto u strukturi onog što reprezentira, na takav način da se potonje može uzeti u obzir, čak i kad nam nije eksplicitno dano.

Obitavati u smislenom znači obitavati u tim strukturama, inertno i dinamično, kako bi ova proširena učinkovitost bila moguća. Dovoljno je jasno kako su reprezentacije prihvatljive, ali jednako bi tako, trebalo biti jasno kako alati, strukturirana mjesta i institucionalizirana praksa proširuju postojeće kapacitete „kodiranjem“ nejasnog. Da uzmemo najokrutniji primjer: ljudski problem osiguravanja zaklona od divljine je sve samo ne jednostavan i očit. Trebalo je proći mnogo generacija naših predaka kako bi došli do osnovnog rješenja koje mi danas

uzimamo, više ili manje, zdravo za gotovo. I kako nam je točno to rješenje dodijeljeno? Na različite načine, ali jedan od najvažnijih leži u oblicima i kvalitetama čekića i čavala, daski i pila, zajedno s našim standardnim postupcima za njihovo korištenje. Ovo ne „reprezentira“ akumuliranu mudrost naših predaka, barem ne u bilo kojem semantičkom smislu, ali nekako ju uključuje i prenosi do nas na jedinstven i učinkovit način.

Evo još jednog kuta gledanja na istu ideju. C.P.U. se ne može shvatiti kao „obrada teksta“ sam po sebi već samo u kombinaciji s barem nekim prikladnim softverom i tekstom za obradu (oba dostupna u RAM memoriji) te prigodnim ulazno/izlaznim dodaci (tipkovnica, zaslon, pisač, na primjer); tek tada možemo reći da imamo obradu teksta. No, niti jedan od ovih drugih objekata nije u izolaciji niti sam po sebi dostatan; softver reagira na upute tipkovnice i radi izmjenu teksta na ekranu samo kada operativni sustav prihvati njegov zahtjev koji se provodi kroz C.P.U. koji opet radi pomoću operativnog sustava koji se nalazi u nekom hardveru gdje su spojeni tipkovnica i ekran i tako dalje.

Klasična kognitivna znanost i simbolička UI su htjele uvesti esencijalnu strukturu u mozgu koja je dostatna za inteligenciju, a upravo paralelno s primjerom obrade teksta, imamo inteligenciju samo kada u obzir uzmemo sve dijelove zajedno – primitivne operacije (kognitivnu arhitekturu), naučene ekspertize (skripte, produkcije, zdrav razum), aktualne modele i planove, itd. Svi se ovi dijelovi mogu karakterizirati na relevantnoj razini opisa – razini koja je, međutim, koherentna i ima smisla samo kada su svi dijelovi karakterizirani zajedno. Ono što imamo ovdje je vrsta holizma (osim što je ovaj širi od poznatog holizma u semantičkoj interpretaciji jer uključuje procesore i operacije u stvarnom vremenu). Temeljne smislene aktivnosti i objekti u kojima inteligencija nekog sustava obitava, imaju smisla samo na temelju načina na koji cjelokupni sustav funkcionira kao cjelina.

Sada možemo shvatiti da Dreyfus predlaže još i treću varijaciju na istoj osnovnoj slici. U toj analogiji, svaka je osoba zapravo samo „procesor“ u konfiguraciji više procesora, a relevantne operacije odvijaju se u javnom svijetu. Smislene „podatkovne strukture“ na kojima „procesori“ rade su javni objekti: čekići, gradovi, filmovi, izborne kampanje, korporacije, tehnologije, revolucije, itd., a ti smisleni objekti predstavljaju sam svijet. Kao što je i prije rečeno, smislenost i inteligencija obitavaju samo u integriranoj cjelini, a ne u samim

„procesorima“. Ipak, ovi su procesori presudni, u smislu da bez njih ništa od ostatka integrirane cjeline ne bi funkcioniralo i cijela bi se struktura srušila.[115]

Ovdje dolazimo do treće i završne točke. Dreyfus kaže:

> “The human world, then, is prestructured in terms of human purposes and concerns in such a way that what counts as an object or is significant about an object already is a function of, or embodies, that concern. This cannot be matched by a computer, for a computer can only deal with already determinate objects, and in trying to simulate this field of concern, the programmer can only assign to the already determinate facts further determinate facts called values, which only complicates the retrieval problem for the machine.“[116]

Na što Dreyfus misli kada kaže ljudski svijet? Očito, misli na sav svijet oko nas, onaj u kojem živimo svaki dan. No to još uvijek može dovesti do zablude; je li naš (ljudski) svijet različit i u kontrastu sa nekim drugim svjetovima – životinjskim svijetom, Božjim svijetom, izvanzemaljskim svijetom ili nekim sličnim svijetom. Međutim, Dreyfus tvrdi da postoji samo jedan svijet, ovaj ovdje i on je naš.[117]

Dobro, no u kojem je smislu to „naš“ svijet? To je naš svijet u smislu da ga razumijemo i živimo; on je suština i oblik naših života. Međutim svijet je prilično raznolik u mjeri u kojoj ponekad možemo govoriti o „različitim“ svjetovima: svijet kazališta, veliki sportski svijet, novi svijet kognitivne znanosti, a da ne spominjem nomadski beduinski svijet, agrarni Hopi svijet i tako dalje. To su sve još uvijek ljudski svjetovi, razumljivi, u kojima obitava ljudska inteligencija. Po mom mišljenju, a pretpostavljam i Dreyfusovom, ne postoji takva stvar kao životinjska ili božanska inteligencija. Čak i ako postoji, mi ju ne bi mogli razumjeti te je stoga besmisleno govoriti o inteligenciji. A isto se može reći i o mogućnosti inteligentnih vanzemaljaca. A čak i onda kada bi ju mogli razumjeti postala bi dio našeg proširenog svijeta. Svijet je sam po sebi područje smislenosti, drugim riječima, on je mjesto gdje inteligencija obitava.

Ali što je s fizičkom svemirom: bezbroj zvijezda i galaksija, velikih sila, materije, energije, čestica, petnaest milijardi godina atoma te početno ništavilo svemira? Nije li to pravi svijet

[115] (Dreyfus 1972. str. 178.)
[116] (Dreyfus 1972. str. 173.)
[117] (Dreyfus 1972. str. 180-183.)

koji je izvan našeg i čije dijelove pokušavamo tumačiti i gledati kao smisleno? Ne, fizički je svemir dio našeg svijeta. To je neobičan i svojstven dio jer neobična i svojstvena inteligencija obitava u njemu, inteligencija koja je izmislila značenje „besmisleno“ i zajedno s „našom“ tvori novu vrstu cjeline. Međutim, to nije mjesto da ispitujemo smislenost, jer ona nije primarna za nas. Prema tome, Dreyfus zaključuje da, s ovakvim razumijevanjem fizike istraživači UI i kognitivna znanost trebali bi graditi svoje teorije o inteligenciji od vrha prema dnu, a ne obrnuto. Temelj inteligencije ne leži u nejasnom, već ovdje, u zemaljskom.[118]

5.4. Pozadina i utjecaj Dreyfusove kritike

Godine 1972., prije četrdeset i pet godina, filozof Hubert L. Drefus objavio je knjigu *What Computers Can't Do*, kritiku tadašnje prve generacije istraživanja umjetne inteligencije čime je uzdrmao temelje svijeta UI. U to vrijeme, mnogi su ga znanstvenici oštro napali, ali danas možemo zaključiti da su mnoge od njegovih kritika prošle test vremena. U današnjim terminima, ono o čemu je Dreyfus diskutirao bila je potreba da roboti imaju autonomiju kako bi mogli učiti iz iskustva i kategorizirati svoje svjetove u interakciji među sobom, upravo ono čime se današnji vodeći znanstvenici robotike bave. Mnogo ispred svog vremena.

Značajan aspekt Dreyfusove kritike je taj da je motiviran filozofskom tradicijom fenomenologijom koja u to vrijeme nije bila često povezivana sa znanošću i tehnologijom te naizgled daleko u svojim problemima. Fenomenologija, kao što se može vidjeti u radovima Martina Heideggera i Mauricea Merleau-Pontya, se odnosi na opisivanje međuodnosa između čovjeka i svijeta i kao polazišnu točku uzima iskustva ljudskih bića u prvom licu. I dok Heidegger, Merleau-Ponty i drugi filozofi fenomenologije tvrde prilično određene stvari o prirodi ljudske percepcije, prirodi razmišljanja i ponašanja, njihove tvrdnje o znanosti i tehnologiji imaju tendenciju da budu prilično opće i apstraktne. Dreyfus je bio u stanju vješto primijeniti njihove ideje u svojoj kritici umjetne inteligencije da bi došao do vrlo specifičnih i konkretnih zaključka.

Od njegovog najranijeg rada na tu temu pod nazivom *Alkemija i umjetna inteligencija*, Dreyfus je progresivno proširio svoju filozofsku kritiku umjetne inteligencije upotrebljavajući ideje fenomenologa kao što su Heidegger, Merleau-Ponty i Husserl, ali i koristeći spoznaje

[118] (Dreyfus 1972. str. 177-183.)

drugih filozofa uključujući Ludwiga Wittgensteina, Michela Foucaulta i Sörena Kierkegaarda. Jedan od glavnih Dreyfusovih problema, koji se regularno pojavljuje tijekom njegovih spisa, je kako artikulirati različite načine na koje ljudska bića doživljavaju svijet i razvijaju načine snalaženja u njemu.

Druga važna točka je njegova kritika kartezijanskog racionalizma. U Dreyfusovom tumačenju, ključne pretpostavke racionalizma su da je stvarnost racionalna struktura izgrađena od nezavisnih elemenata na pravilno uređen način, da ljudsko razmišljanje djeluje na isti racionalan način te da je sve što nije racionalizirajuće – koje se ne može izraziti i braniti racionalnim načelima, ima malu, ako ikakvu vrijednost. Dreyfus je uvjeren da je zapadna kultura još uvijek, u velikoj mjeri, oblikovana ovim racionalističkim pretpostavkama, ali je također uvjeren, na temelju čitanja Heideggera, Merleau-Pontya i Wittgensteina, da su ove pretpostavke fundamentalno slabe. Racionalne formalne strukture su, prema njegovim riječima, ljudska konstrukcija koja je samo naknadno nametnuta stvarnosti. Spoznajna stvarnost sama po sebi nema racionalnu strukturu, njezin sadržaj određuju ljudske potrebe i aktivnosti. Najosnovniji način dolaska do znanja je intuitivan, a ne racionalan. Racionalizam, koji se javlja u UI i drugdje, ne zna ništa o tim izvornim strukturama stvarnosti i ne može obavljati pravednu ulogu u odnosu na ulogu intuitivnog znanja i vještina. Dreyfus je nepopustljiv zagovornik intuitivnog znanja i vještina, a oštar kritičar racionalizma u svim svojim suvremenim oblicima.

Važno je naglasiti da Dreyfus nigdje ne kaže da je postizanje umjetne inteligencije u osnovi nemoguće; ono što on tvrdi je to da su trenutni istraživački programi i smjer u kojem oni idu fatalno pogrešni. Tvrdi da ako želimo stvoriti uređaj ili uređaje koje posjeduju ljudsku inteligenciju, morat ćemo implementirati u njih i svojstvo čovjeka kao bića u svijetu, svojstvo koje zahtjeva posjedovanje tijela i organa, više ili manje, poput naših i društvenu akulturaciju, odnosno društvo koje je više ili manje isto kao naše. Ovaj stav danas dijele mnogi psiholozi i filozofi utjelovljene i distribuirane spoznaje, ali i neki istraživači u području robotike kao i u području umjetnog života (ALife, engl. *Artificial Life*).

Umjesto zaključka

U ovom sam radu prikazao Turingov test te njegove kritike. Osim toga, detaljnije sam se posvetio onome što smatram da su dva najvažnija filozofska argumenta protiv UI: Searleovom argumentu kineske sobe i Dreyfusovoj utjelovljenoj inteligenciji.

Osjećam veliko divljenje prema liku i djelu Alana Turinga, jednom od najvećih mislilaca 20. stoljeća. Iako u njegovim predviđanjima postoje neki nedostaci, smatram da bi nam još puno toga uspio pojasniti te bi danas sigurno imali bolju situaciju u polju UI da nije tako tragično rano okončao svoj život. Mnogo se toga promijenilo u svijetu umjetne inteligencije otkad je Turing, 1950. godine, napisao svoj rad, ali čini mi se da je igra oponašanja koju je opisao još uvijek dobar test za umjetnu inteligenciju. Dovoljno dobar da se u potpunosti ne odbaci dok ga ne poboljšamo ili zamijenimo nekim novim načinom za testiranje inteligencije.

Mjesto koje je odigralo važnu ulogu u računalnoj industriji, a time i u UI bez sumnje je Silicijska dolina, smještena u južnome dijelu Zaljevskog područja San Francisca u sjevernoj Kaliforniji. Zanimljivo je da se, niti stotinjak kilometara sjevernije, nalazi UC Berkeley, jedno od najboljih svjetskih sveučilišta s odjelom za filozofiju koji je desetljećima bio dom dvojici najpoznatijih kritičara UI; Hubertu Dreyfusu i Johnu Searleu koji su svojim kritikama uzdrmali temelje umjetne inteligencije. U onoj mjeri u kojoj svijet sada shvaća da mozak nije samo veliko računalo kao što ga je prva generacija istraživača UI opisala, velik dio zasluga ide Dreyfusu i Searleu. Iako se smjer istraživanja UI s vremenom promijenio, njihove kritike se i dalje mogu primijeniti.

John Searle i Hubert Dreyfus za mene su dva velikana suvremene filozofije. U ovom diplomskom radu nisam htio ulaziti u raspravu o prioritetu njihovih filozofskih metodologija, redom logičke analize i fenomenologije. Moj cilj nije bio napasti ova dva filozofska pristupa i pokazati koji je bolji; u fenomenologiju sam upućen tek posljednjih godinu dana kojih sam proveo na studiju u Ljubljani, te o njoj znam premalo da bi ju mogao adekvatno braniti ili napadati. S druge strane, logička analiza je ono prema čemu oduvijek prirodno težim i čemu su me podučavali na Filozofskom fakultetu u Rijeci pa bi takva usporedba onda bila presubjektivna. Moj je cilj bio, dakle, pokušati što objektivnije prikazati kritiku svakog od ovih filozofa i izvući iz njih ono što mi se čini najvažnijim za filozofski prikaz problema umjetne inteligencije. Iako se njihove metodologije razlikuju, vjerujem da važnost filozofskih kritika koje upućuju UI prevladava te razlike.

Jedna od prvih stvari koje su me privukle razmišljanju o filozofskim problemima umjetne inteligencije, bila je Searleova kineska soba. Smatram da se Searleov argument jasno usredotočuje na pravu vrstu problema; naglašava važnost i zagonetnu prirodu promatranja svijeta iz perspektive prvog lica. Searleov naglasak na važnost perspektive prvog lica trebao bi ga činiti prirodnim saveznikom Dreyfusa. Na neki način oni i jesu saveznici, budući da se oboje slažu da je perspektiva prvog lica ono što je bitno za filozofsko mišljenje o umu i svijetu. Posvećujući pozornost važnosti perspektive prvog lica u svojim kritikama, čini mi se da su Searle i Dreyfus učinili duboke i važne korake u shvaćanju našeg uma i umjetne inteligencije. Ako je prava filozofska metodologija ona koja radi, onda smatram da su Dreyfusova fenomenologija i Searleaova logička analiza uzete zajedno, bitne za puni prikaz inteligencije, svijesti i društvene stvarnosti.

Prije nego što okončam ovu raspravu htio bih napomenuti da se u ovom radu nisam bavio moralnim i etičkim pitanjima koje implicira umjetna inteligencija, a s kojima bi se jednog dana mogli suočiti, ako računalo ikada uspije proći Turingov test. Moguća pitanja bila bi: Ima li računalo prava? Može li računalo glasati na izborima? Može li računalo počiniti zločin? Tko plaća račun za napajanje? Što ako se računalo reproducira? Ova su pitanja predmet mnogih filmova znanstvene fantastike, ali vjerujem da mogu postati i ozbiljna pitanja na koja ćemo jednoga dana možda biti prisiljeni odgovoriti.

Veliko pitanje kojim sam se u ovom radu bavio, i koje je držalo računalne znanstvenike i filozofe budne noću, jest; „Mogu li računala razmišljati?“ Moj je zaključak da ne mogu. Ipak, to nije i konačan odgovor. Vrijeme će pokazati hoće li tehnologija moći izgraditi inteligentno računalo koje može proći TT. Vjerujem da će se to ubrzo i dogoditi, a tada ćemo morati razmišljati o tome kako izgraditi računalo koje posjeduje svijest. Hoćemo li ikada moći izgraditi svjesno računalo? Na ovo pitanje znat ćemo odgovor, ali tek kada u potpunosti shvatimo kako funkcionira naš mozak.

Filozofija je često meta kritika koje tvrde da se ona ne bavi implikacijama u stvarnome svijetu, ali je, u ovom slučaju, pokazala najbolji smjer istraživačima UI. Istraživači u području umjetne inteligencije i kognitivne znanosti, koji će prepoznati da je osjetljivost na perspektivu prve osobe ključna za istraživanje, ostvarit će napredak unutar svojih domena. Kakav god da se napredak unutar UI postigne, morat će se temeljiti na napretku koji su Searle i Dreyfus prvi predstavili.

Niti je mozak računalo, niti je um potpun bez tijela. Prihvaćanje ove važne ideje promijenit će način na koji istražujemo naš um, a time i pristup u pokušaju reprodukcije i stvaranju umjetne inteligencije. Tijelo je ono što čini svijet važnim za nas. Mi ne bi imali interes za stvari koje se događaju oko nas kada taj interes ne bi bio tjelesan. Bilo bi nam svejedno imamo li tijelo ili smo mozgovi u posudi.

Umjetna inteligencija ne ostvaruje značajan napredak jer, unatoč suprotnim tvrdnjama, i dalje koristi reprezentacijski pristup ili simboličku UI. Moraju se poduzeti konstruktivne akcije. Umjetna inteligencija i kognitivna znanost moraju širom otvoriti svoja vrata drugim znanostima poput psihologije, etnologije, neurobiologije, antropologije, neuroznanosti i fizike, kao i prema humanističkim znanostima poput filozofije, lingvistike, sociologije i povijesti umjetnosti. Napredak u shvaćanju ljudskog uma ovisit će o raznolikim interdisciplinarnim prijedlozima koji će uslijediti iz tih znanosti. Potrebna su nam nova konceptualna istraživanja za stvaranje hipoteze o tome kako naš um i tijelo interagiraju jedno s drugim te se uklapaju u smisleni, društveni svijet. Nova se istraživanja moraju fokusirati na nove ideje i koncepte koji toliko ne ovise o tehnološkim resursima. Tek kada kognitivna znanost u potpunosti shvati interakciju ljudskog uma i tijela moći će to znanje prenijeti istraživačima UI koji će, uz već postojeće tehnološke izume, moći stvoriti pravu umjetnu inteligenciju. Vrijeme je ključ koji otključava tajne inteligencije.

Literatura

Boden, M. 1996. *Artificial Intelligence*, Handbook of Perception and Cognition, Second Edition, Academic Press.

Cole, D., 1991. *Artificial Intelligence and Personal Identity*, Synthese, 88. str. 399-417. URL=https://pdfs.semanticscholar.org/59fd/331d1db3ef2650f3245d6d97e880d19e2274.pdf preuzeto 2.9.2017.

Cole, D. 2014. The Chinese Room Argument. *The Stanford Encyclopedia of Philosophy*. Edward N. Zalta (ur.), URL=https://plato.stanford.edu/entries/chinese-room/ pristupljeno 6.8.2017.

Crevier, D. 1993. *AI: The Tumultuous Search for Artificial Intelligence*, New York, NY: BasicBooks.

Douglas, J. R. 1978. *Chess 4.7 versus David Levy*, BYTE. str. 84. pristupljeno 11.1.2017. https://archive.org/stream/byte-magazine-1978-12/1978_12_BYTE_03-12_Life#page/n85/mode/2up

Dreyfus, H. L. 1965. *Alchemy and Artificial Intelligence*, RAND Corporation, preuzeto sa http://www.rand.org/content/dam/rand/pubs/papers/2006/P3244.pdf (18.01.2017.)

Dreyfus, H. L. 1972. *What Computers Can't Do*, New York, Harper & Row.

Dreyfus, H. L. 2007. *Why Heideggerian AI Failed and How Fixing it Would Require Making it More Heideggerian*, Oxford University Press, preuzeto sa http://cid.nada.kth.se/en/HeideggerianAI.pdf

Gillies, D. A. 1996. *Artificial Intelligence and Scientific Method*, Oxford University Press.

Gunderson K., 1964. *The Imitation Game*. Mind, vol. 73. str. 234–245.

Harnad S., 1991. *Other Bodies, Other Minds*, Minds and Machines, volume 1. str. 43–54.

Hauser L., 1993. *Reaping the Whirlwind*, Minds and Machines, vol. 3, No. 2. str. 219-238.

Maudlin, T., 1989. *Computation and Consciousness*, Journal of Philosophy, LXXXVI. str, 407-432. preuzeto sa http://web.csulb.edu/~cwallis/labs/stanford/Computation&consc.pdf (30.08.2017.)

Millican P. and Clark A. (ur.) 1999. *Machines and Thought: The Legacy of Alan Turing*, Volume I, Oxford: Clarendon Press.

Minsky, M. 1961. *Steps Toward Artificial Intelligence,* Proceedings of the IRE 49(1):8-30, February 1961, preuzeto sa (12.1.2017.) https://www.cs.utexas.edu/~jsinapov/teaching/cs378/readings/W2/Minsky60steps.pdf

Miščević N., Smokrović N. 2001. *Računala, mozak i ljudski um: zbornik tekstova iz teorije umjetne inteligencije i kognitivne teorije*, Izdavački centar Rijeka, Rijeka.

Newell, A. and Simon, H. A. 1961. *Computer Simulation of Human Thinking*, *Science,* New Series, Vol. 134, No. 3495, preuzeto 21.01.2017. sa http://www1.rdaprendizagem.net/biblioteca/newell-simon.pdf

Norvig, P. 1992. *Paradigms of artificial intelligence programming:case studies in Common Lisp.* San Francisco, California: Morgan Kaufmann Publishers Inc. str. 109–149.

Penrose, R. 2004. Carev novi um : razmišljanja o računalima, razumu i zakonima fizike, Izvori, Zagreb.

Perron, C. 2016. *Can You Tell Which Diamonds Are Lab Grown*, preuzeto 17.01.2017. sa http://www.brilliantearth.com/news/can-you-tell-which-diamonds-are-lab-grown/

Randell, B. 1980. *The Colossus*, u A History of Computing in the Twentieth Century, preuzeto sa http://www.cs.ncl.ac.uk/research/pubs/books/papers/133.pdf, (16.01.2017.)

Rajaraman, V. 2014. *John McCarthy – Father of Artificial Intelligence*, u Asia Pacific Mathematics Newsletter, Vol. 4 No. 3, preuzeto sa http://www.asiapacific-mathnews.com/04/0403/0015_0020.pdf (16.01.2017.)

Russel, S. J. and Norvig, P. 2010 *Artificial Intelligence: A Modern Approach (3rd Edition),* Prentice Hall, Englewood Cliffs, New Jersey.

Sabar, S. 2013. *What's a Gestalt?*, Gestalt Review, 17(1):6-34, Gestalt International Study Center, preuzeto sa http://www.gisc.org/gestaltreview/documents/whatsagestalt.pdf

Schweizer P., 1998. *The Truly Total Turing Test.* Minds and Machines, vol. 8. str. 263–272.

Searle, J., 1980. *Minds, Brains and Programs*, Behavioral and Brain Sciences.

Searle, J., 1984, *Minds, Brains and Science*, Cambridge: Harvard University Press.

Searle, J., 1990. *Is the Brain's Mind a Computer Program?*, Scientific American, 262, no.1, str. 26-31., preuzeto 29.08.2017 sa http://www.cs.princeton.edu/courses/archive/spr06/cos116/Is_The_Brains_Mind_A_Computer_Program.pdf

Searle, J., 2015. *Consciousness in Artificial Intelligence*, Talks at Google, intervju, https://www.youtube.com/watch?v=rHKwIYsPXLg&t=2594s pristupljeno 3.8.2017.

Simon, H. A. 1965. *The Shape of Automation for Men and Management*, Harper & Row.

Simon, H. A. 1996. *Machine as Mind*, u The Legacy of Alan Turing, Volume 1: Machine and Thought, edited by Millican, P. and Clark, A., Oxford: Claredon Press.

Stevenson J. G., 1976. *On the Imitation Game.* Philosophia, vol. 6, str. 131-133.

Turing, A. 1950. *Computing Machinery and Intelligence, Mind*, vol. 59, No.236. str. 433-460.

van Heerden, P. J. 1968. *The Basic Principles of Artificial Intelligence*, The Foundation of Empirical Knowledge with a Theory of Artificial Intelligence, Wassenaar: N. V. Uitgeverij Wistik.

Watt S., 1996. Naive Psychology and the Inverted Turing Test, Psycoloquy, volume 7(14). URL=http://www.cogsci.ecs.soton.ac.uk/cgi/psyc/newpsy?7.14, pristupljeno 3.8.2017.

Weizenbaum, J. 1966. *ELIZA—a computer program for the study of natural language communication between man and machine*, Magazine Communications of the ACM, Volume 9 Issue 1, Jan. 1966., preuzeto sa http://web.stanford.edu/class/linguist238/p36-weizenabaum.pdf