ANαliZA

časopis za kritično misel

Letnik 17, številka 3, november 2013 · ISSN 1408-2969

FUNKCIONALIZEM IN VPRAŠANJE UMETNE INTELIGENCE

Janez Bregant

Funkcionalizem, računalniki in »kitajska soba«

Filozofska fakulteta Univerze v Mariboru

Citiranje: Janez Bregant, »Funkcionalizem, računalniki in »kitajska soba««, *Analiza: časopis za kritično misel* 17, št. 3 (2013): 67–95.

To delo je objavljeno pod licenco Creative Commons Priznanje avtorstva 4.0 Mednarodna (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/deed.sl.

Izdajatelj: Društvo za analitično filozofijo in filozofijo znanosti, Ljubljana.

Elektronska različica je bila leta 2026 rekonstruirana iz izvirnih stavnih datotek in preverjena s tiskano izdajo. Številke strani ustrezajo tiskani izdaji; prelomi strani znotraj prispevka se lahko nekoliko razlikujejo.

Janez Bregant
Filozofska fakulteta Univerze v Mariboru

FUNKCIONALIZEM, RAČUNALNIKI IN »KITAJSKA SOBA«

V članku je predstavljena zveza med funkcionalizmom, ki mentalna stanja obravnava kot funkcionalna stanja, računalniki, ki so po njegovi računski različici sposobni misliti, in argumentom iz »kitajske sobe«, ki naj bi enkrat za vselej pokazal, da stroji nikoli ne bodo inteligentni. Prvi del prikazuje osnovno idejo in motivacijo za funkcionalizem, skupaj z glavnimi ugovori proti njemu, drugi del članka miselni eksperiment s »kitajsko sobo«, ki formuliran kot argument zavrača idejo o mislečih strojih na podlagi tega, da manipulacija s formalnimi simboli še ni dovolj za razumevanje njihove vsebine, tretji pa povzema glavne ugovore argumentu iz »kitajske sobe« in pri odgovoru nanje prinaša nekaj novih idej, ki zagovarjajo prepričanje, da so inteligentni računalniki kljub vsemu načeloma možni.

Ključne besede: »kitajska soba«, umetna inteligenca, simbolna hipoteza, jezik misli, računska teorija duha.

1 Funkcionalizem

> V začetku 60. let 20. stol. je Putnam (1960) oblikoval teorijo duha, ki je dobila ime *funkcionalizem*, njene začetke pa lahko najdemo že pri Aristotelu (Block, 1980b/2007: 147, 158). Ta je dušo opredelil kot skupek posebnih sposobnosti živega organizma, s katerimi usmerjamo in uravnavamo življenjske funkcije z namenom, da preživimo in napredujemo kot živo, dejavno, zaznavajoče in razumno bitje:[1] »Zmožnosti, o katerih smo govorili, so: prehranjevalna, želelna ali nagonska, zaznavna, gibalna po prostoru ter razumska. Rastlinam pripada zgolj prehranjevalna zmožnost, drugim živim bitjem pa je na razpolago tako ta kot zaznavna zmožnost. Če pa jim pripada zaznavnost, jim pripada tudi želelna ali nagonska zmožnost /.../ Marsikaterim živalim pa poleg teh zmožnosti pripada še zmožnost gibanja po prostoru, drugim pa tudi razumska zmožnost ter um, tako

[1] Aristotel je npr. trdil, da imajo stvari svoj namen – telos – po svojem bistvu.

recimo ljudem /…/« (Aristotel, 1993: 123–125). Danes bi lahko funkcionalizem na kratko definirali kot filozofsko teorijo, ki mentalna stanja označuje v pojmih njihovih vzročnih vlog, kaj to pomeni, pa si bomo ogledali v nadaljevanju.

Strojni funkcionalizem

Putnam je kot osnovo za funkcionalizem predlagal analogijo med človeško duševnostjo in t.i. Turingovim strojem. To je enostavni abstraktni sistem za dokazovanje teoretičnih predpostavk, ki si ga je leta 1936 izmislil Alan Turing,[2] sestavljen pa je iz gibljive *glave* in neskončnega *traku*. Glava se pomika preko polj, na katere je razdeljen trak in katera so zasedena s simboli (vsako polje z enim), ki so predvideni na končnem seznamu znakov. Delovanje stroja določa pet elementov/parametrov: (i) simbol, ki ga zazna (*z*), (ii) notranje stanje, v katerem se v danem trenutku nahaja (*s*), (iii) naslednji simbol v istem polju (*z + 1*), (iv) naslednje stanje (*s + 1*) in (v) naslednje polje/smer gibanja (*p*). Prva dva elementa se nanašata na sedanje obnašanje, ostali trije pa na obnašanje v naslednjem trenutku. V splošnem lahko glede na to stroj naredi naslednje: (a) ostane v stanju, v katerem je ali preide v neko drugo stanje, (b) pusti na traku zapisan isti simbol, ali ga izbriše, ali ga izbriše in vpiše novega ter (c) premakne se po traku levo ali desno ali ostane na istem mestu. Kje se ustavi, kaj s poljem stori in kam se giblje, je vnaprej določeno s pravili, ki imajo obliko pogojnika naslednje vrste: če je stroj v stanju *s* in dobi vnos *z*, proizvede naslednji iznos *z + 1* in gre v naslednje stanje *s + 1*.

Takšen niz pogojnikov je pogosto formuliran kot *strojna tabela*, sistem, ki je tako opisan, pa imenujemo avtomat: v bistvu je vsak avtomat zgolj ena vrsta Turingova stroja, kar lahko ponazorimo z enostavnim primerom stroja za coca-colo (Block, 1980b/2007, Podor, 1981). Predpostavimo, da je njegova naloga, da izvrže pločevinko coca-cole, ko prejme 1 evro. Vanj lahko vržemo kovanec za 50 centov in za 1 evro. Avtomat se lahko nahaja v dveh stanjih, *s*, in *s 2*, ki sta določeni glede na vnos, tj. kateri kovanec prejme. Možnosti z vidika obeh stanj so potem naslednje:

> (s_1): (i) če je avtomat v s_1 in prejme 50 centov, ne izvrže ničesar, zgolj preide v s_2, (ii) če pa je avtomat v s_1 in prejme 1 evro, izvrže coca-colo in ostane v s_1;
>
> (s_2): (i) če je avtomat v s_2 in prejme 50 centov, izvrže coca-colo in preide v s_1, (ii) če pa je avtomat v s_2 in prejme 1 evro, izvrže coca-colo in 50 centov ter preide v s_1.

Shematsko lahko to ponazorimo z naslednjo tabelo:

[2] Za podrobnosti *glej* Turing (1936).

vnos/stanje	s_1	s_2
50 centov	ni iznosa pojdi v s_2	izvrži coca-colo pojdi v s_1
1 evro	izvrži coca-colo ostani v s_1	izvrži coca-colo in 50 centov pojdi v s_1

Putnam bistvo takšnih avtomatov izrazi takole: »Ko je Turingov stroj opisan s pomočjo 'strojne tabele', je opisan kot nekaj, kar ima trak, tiskalnik, 'skener' (to mogoče ne rabi biti nič drugega kot točka stroja, ki je v kateremkoli danem trenutku poravnana zgolj z enim kvadratom na traku) in končni niz (A, B, C itd.) 'stanj'. /.../ Razen tega je opisan samo še z vnaprej določenimi pravili, ki opredeljujejo vrstni red, po katerem si stanja sledijo in kaj je kje natisnjeno« (Putnam, 1960: 158–159). Pomembno je, da ostane v takšnem opisu Turingovega stroja vprašanje narave njegovih stanj, tj. njegove notranje zgradbe, povsem odprto oziroma nedoločeno: »Zlasti pa 'logični opis' Turingovega stroja ne vključuje nobene specifikacije fizične narave teh 'stanj' – ali celo fizične narave celotnega stroja« (Putnam, 1960: 159). Njegova stanja so definirana v odnosu do simbolov na traku (vnosov) in v odnosu do drugih stanj, o tem, kako se med seboj razlikujejo glede fizične strukture, pa ni govora, realizira jih lahko karkoli, tako nekaj fizičnega kot tudi nefizičnega.[3] Poudariti je treba, da opisujejo strojne tabele takšne vrste operacije *deterministčnega avtomata*, kjer program za vsako stanje in niz vnosov določi naslednje stanje sistema in iznos, ki ga proizvede. Danes se večini avtorjev to ne zdi več pravi model za opredelitev človeškega duha, kar je hitro ugotovil tudi Putnam, ki je nekaj let kasneje razvil idejo, da smo ljudje *probabilistični avtomati*, kjer program za vsako stanje in niz vnosov določa zgolj verjetnost naslednjega stanja sistema in iznosa, ki ga proizvede (Putnam, 1967/1 980a).

Kakorkoli, Putnam je bil prvi, ki je značilnosti delovanja Turingovega stroja prenesel na področje našega mentalnega življenja. Tudi psihologijo zanima zgolj funkcioniranje organizma v različnih okoliščinah, ne pa to, iz česa je narejen. Pri svojem početju povsem abstrahira od fizičnih lastnosti stanj, skozi katere bitje prehaja, ko opravlja različne operacije. Glede na to, da obstajajo številni materialni in nematerialni sistemi, ki so funkcionalno izomorfni, ne moremo pričakovati redukcije funkcionalne vloge na zgolj enega izmed njih. Isti psihološki opis lahko zadovoljijo z vidika zgradbe povsem različni organizmi, tudi takšni brez nevronov, zaradi česar psihologije ne moremo skrčiti na nevrofiziologijo, saj je po svojem namenu splošnejša: nanaša se na vse funkcionalno izomorfne sisteme ne glede na razlike v njihovi zgradbi (Sesardić, 1984: 86).

[3] S tem se funkcionalizem nasploh izogne t.i. *argumentu iz večvrstne realizacije*, ki preprečuje identifikacijo mentalnih stanj s fizičnimi stanji (Putnam, 1967/1980a) in *onemogoča identitetno teorijo duha*, ki mentalna stanja enači z možganskimi (Smart, 1959/2007).

Tako med Turingovim strojem in človeško duševnostjo obstaja analogija, saj so strojeve strukturalne lastnosti (hardver) podobne človeškim nevrofiziološkim, v tem primeru gre za opis obnašanja v fizičnih pojmih (v slovarju kemije in biologije), strojeve funkcionalne lastnosti (softver) pa našim mentalnim stanjem in njihovi z zakoni določeni organizaciji v smislu vrstnega reda pojavljanja, v tem primeru gre za opis obnašanja v psihičnih pojmih (v slovarju psihologije). »Zanimiva je ugotovitev, da tako kot obstajata dva možna opisa obnašanja Turingovega stroja – inženirjev strukturalni načrt in logikova 'strojna tabela' – obstajata tudi dva možna opisa človeške psihologije. Behavioristični pristop /.../, katerega cilj je priskrbeti popoln fizikalistični opis človeškega obnašanja /.../ To je skladno z inženirjevim ali fizikovim opisom fizično realiziranega Turingovega stroja. Toda ravno tako bi bilo možno najti bolj abstrakten opis človeških mentalnih procesov v pojmih 'mentalnih stanj' (z nedoločeno fizično realizacijo, če sploh) in 'vtisov' (ki imajo na strojevem traku vlogo simbolov) – opis, ki bi specificiral zakone, ki bi nadzorovali vrstni red, po katerem si stanja sledijo /.../ Ta opis bi bil analogen 'strojni tabeli'/.../« (Putnam, 1960: 160–161). Takšen funkcionalizem je dobil ime *strojni funkcionalizem*, v svoji najbolj preprosti različici pa trdi, da je vsaka vrsta mentalnega stanja identična enemu izmed stanj v strojni tabeli. V našem primeru s coca-colo (če pustimo ob strani dejstvo, da si stroj ne more ničesar želeti) tako velja, da je s_1 želja po 50 centih, s_2 pa želja po 1 evru, res pa je, da opis človeških mentalnih stanj v pojmih Turingovega stroja zahteva kompleksnejše strojne tabele.

Metafizični funkcionalizem

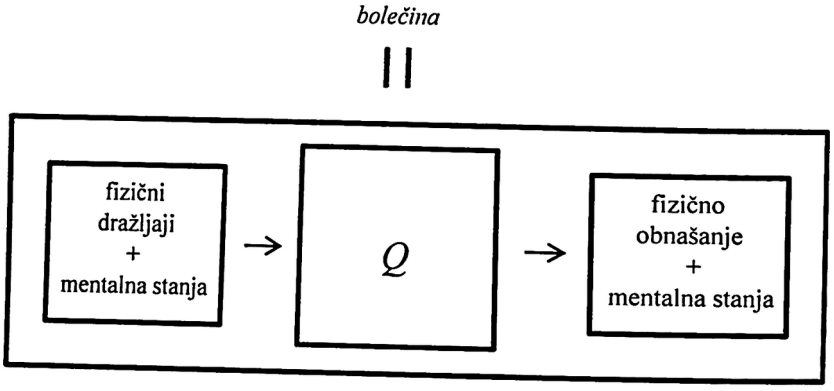
Kljub temu, da ima strojni funkcionalizem prednost pred *behaviorizmom*,[4] saj so stanja strojnih tabel opisana v pojmih relacij do vnosov in iznosov ter v pojmih relacij do trenutnega stanja strojne tabele, npr. dežnik vzamem (*iznos*), če sem prepričan, da bo deževalo (*vnos*) in če želim ostati suh (*trenutno stanje*), se je večina funkcionalistov od njega kmalu poslovila, saj je prevladalo prepričanje, da ne zagotavlja splošnega opisa tega, kaj funkcionalizem pomeni (Block, 1980b/2007). Ker smo ljudje lahko v istem trenutku v več kot zgolj enem mentalnem stanju oziroma lahko v istem trenutku opravljamo več kot zgolj eno mentalno operacijo, se zdi trditev, da smo pri proizvodnji naših duševnih stanj podobni avtomatu za coca-colo, neverjetna (Block in Fodor, 1972). Tako se je kmalu zatem uveljavil pojem funkcionalizma, ki mentalnega življenja ni definiral zgolj v pojmih determiniranih stanj iz strojne tabele, ampak tudi v pojmih vzročnih zvez med čutnimi vnosi in behaviorističnimi iznosi, ki ga lahko imenujemo *metafizični funkcionalizem*.[5] Armstrong je o strojnem funkcionalizmu denimo zapisal: »Splošna ideja je primerjava duha z računalnikom in mentalnih stanj z računalniškimi stanji v smislu njihovega označevanja v pojmih programa softvera. Ena stvar glede te ideje pa je povsem jasna: 'program, ki teče v naših

[4] Za behaviorizem in njegovo kritiko *glej* npr. Putnam (1963/1980a)
[5] Za ostale različice funkcionalizma *glej* Block (1980b/2007).

možganih, bi moral biti drugačen od kateregakoli programa, ki ga lahko napišemo zdaj in v bližnji prihodnosti« (Armstrong, 1999: 105). Kljub temu pa je ideja, da lahko notranja stanja organizma opišemo z njihovimi s pravili določenimi relacijami do vnosa, iznosa in trenutnih notranjih stanj, ostala bistvo vseh funkcionalističnih različic.

Podobno kot strojni funkcionalizem tudi metafizični funkcionalizem išče odgovor na vprašanje, kaj je narava mentalnih stanj, pri čemer pa jih ne enači zgolj z ozkim strojnim stanjem, ampak s širšo *vzročno-funkcionalno vlogo*, zaradi česar bolečina ni več strojno, ampak vzročno stanje: »Bolečina je identificirana z abstraktno vzročno lastnostjo, ki je povezana z dejanskim svetom samo preko svojih posrednih in neposrednih zvez z vnosi in iznosi« (Block, 1980b/2007: 154). To, kar naredi bolečino za bolečino, je določena vzročno-funkcionalna vloga, ki jo ima, ta pa je definirana v pojmih tipičnih vzrokov in tipičnih učinkov ter drugih stanj, ki so v vzročni interakciji in se nahajajo med njima. Ponazorimo to z naslednjim primerom. Predpostavimo, da ureznina na roki in prepričanje, da smo jo staknili, povzročita določeno stanje (fizično ali psihično) in hkrati s tem celo vrsto vzročnih interakcij med njim in drugimi stanji (imenujmo to »črna škatla« – *Q*), to pa povzroči glasen krik in željo po prenehanju. Mentalni pojem, ki ga v primeru, ko se to zgodi, uporabimo za opis tega, kar občutimo, je seveda bolečina. Fizično informacijo in mentalno stanje, ki povzročita verižno vzročno reakcijo v »črni škatli«, sta vnos, druga mentalna stanja in fizično obnašanje, ki so njena posledica, pa iznos. Tako ni bolečina (po tej preprosti teoriji) nič drugega kot lastnost, ki jo nekdo ima, ko je v stanju, ki je povzročeno z urezom z nožem in ki povzroča glasen krik in željo po prenehanju. »Tako bi lahko teorija bolečine, ki bi jo ponudil metafizični funkcionalist, bolečino označevala delno s tem, da je povzročena s poškodbo tkiva, delno s tem, da povzroči željo, da bi se je rešili, in delno s tem, da povzroči dejanje, namenjeno temu, da se poškodovani del telesa loči od tistega, za kar se misli, da je bolečino povzročalo« (Block, 1980b/2007: 149).

To lahko ponazorimo z naslednjo splošnejšo shemo:



Kljub temu, da metafizični funkcionalizem mentalna stanja identificira z

njihovimi vzročno-funkcionalnimi vlogami, se zdi, da ohranja *fizikalizem*, ki v svoji najsplošnejši obliki pravi, da je vse, kar obstaja, sestavljeno iz fizičnih delcev, saj zavrača zgolj trditev, da so mentalna stanja *identična* s fizičnimi.[6] »Funkcionalisti so lahko fizikalisti s tem, ko dopuščajo, da so vse entitete (stvari, stanja, dogodki itn.), ki obstajajo, fizične entitete, pri čemer zanikajo zgolj to, da je tisto, kar povezuje posamezne vrste stvari, fizična lastnost« (Block, 1980b/2007: 149).

Računsko-reprezentacijski funkcionalizem ali računska teorija duha[7]

Posebno vrsto funkcionalizma, »ki je postavil temelje klasični kognitivni znanosti in simbolnemu modeliranju kot empirični teoriji /…/«. (Markič, 2010: 34), pa predstavlja različica funkcionalizma, ki primerja psihološko razlago tega, kako je lahko misel proizvod fizičnih predmetov, npr. naših možganov, z računalniškim programom.[8] Njeno bistvo je priskrbeti/odkriti funkcionalno analizo mehanskih procesov, na katere lahko skrčimo naše mentalne operacije, ker pa ta običajno vsebuje algoritemski zapis, tj. zaporedje korakov računanja, ki je podoben temu, kako deluje Turingov stroj, je tudi ta vrsta funkcionalizma z njim povezana. Funkcionalizem v tem smislu je v grobem kombinacija reprezentacijske teorije, ki predstavlja svet preko *jezika misli* in računalniške analogije, po kateri so psihični procesi obravnavani kot računanje s takšnimi reprezentacijami, zaradi česar ga imenujemo *računsko-reprezentacijski funkcionalizem* ali *računska teorija duha*. Block ga rahlo ironično opiše takole: »Kakršno koli skrivnostnost naj bi naše mentalno življenje sprva imelo, jo funkcionalna analiza mentalnih procesov razčleni do te mere, da se zdi, kot da je sestavljena iz računskih operacij, ki so tako primitivne kot operacije digitalnega računalnika – procesi, ki so tako neumni, da psihološke razlage, ki se na njih sklicujejo, niti slučajno niso krožne« (Block, 1980b: 147).

V resnici je računska teorija duha sestavljena iz treh sicer povezanih delov: (i) *jezika misli*, (ii) *strukturne občutljivosti sistemskih procesov* in (iii) *reprezentacijske teorije duha. (i)* Jezik misli ali »mentalese« je leta 1975 predstavil Fodor v svoji knjigi *The Language of Thought*, gre pa za trditev, da so mentalne predstave zgrajene na isti način kot jezik, kar se včasih opisuje s frazo »misli so stavki v glavi«. Jezikovna struktura, ki je za opis misli primerna, pa vsebuje dva elementa, *kombinatorno sintakso* in *kompozicionalno semantiko* (Fodor in Pylyshyn, 1988). Za njuno ponazoritev vzemimo jezik stavčne logike, ki s simboli in vezniki tipično predstavlja stavke naravnega jezika in predpostavimo, da *p* označuje stavek »Peter je suh«, *s* stavek »Sara je zabavna«, *m* pa stavek »Marko je debel«. Vsak izmed simbolov je *atomarna predstava*

[6] Za argumente v prid trditvi, da iz resničnosti funkcionalizma sledi neresničnost fizikalizma *glej* Block (1980c), Block in Fodor (1972/1980), Putnam (1966/1975).

[7] Uporablja se tudi izraz *računski funkcionalizem*.

[8] Trditev, da so človeški možgani zelo podobni računalnikom in da se lahko z njihovim preučevanjem dokopljemo do pomembnih ugotovitev glede našega duha, sicer imenujemo *računalniška metafora*. (Miščević in Smokrović, 2001: 9)

določenega naravnega stavka, če pa jih združimo v stavek »Peter je suh« in »Sara je pametna« ali »Marko je debel« in to označimo kot (p a s) v *m*, dobimo njegovo *sestavljeno predstavo*. V tem primeru lahko rečemo, da jezik stavčne logike (in katerikoli drug jezik) uporablja kombinatorno semantiko, saj za označevanje uporablja atomarne in sestavljene predstave, kar je nujni in zadostni pogoj, ki ga mora jezik izpolniti, da ima to lastnost.

Razen tega imajo predstave v jeziku stavčne logike tudi pomen (vsebino), tj. lahko so resnične ali neresnične, kakšne so, pa je odvisno od resničnostne vrednosti sintaktičnih delov, njihove razvrstitve in veznikov med njimi. Če kateregakoli izmed naštetih elementov spremenimo, se spremeni tudi resničnostna vrednost celotne predstave/izraza, zaradi česar pravimo, da vsebuje jezik stavčne logike (in katerikoli drug jezik) tudi kompozicionalno semantiko, »ki za vsako možno kombinacijo določa, kako pomeni sintaktičnih delov sestavljenega izraza prispevajo k pomenu celote« (Markič, 2010: 36).

(ii) Strukturno občutljivost sistemskih procesov, ki je drugi del računske teorije duha, Fodor in Pylyshyn (1988) opredelita kot zmožnost sistema, da spremembe mentalnih stanja definira z načeli, ki so nad strukturo mentalnih predstav. Tega pa je sposoben zgolj v primeru, ko uporablja računske procese, ki niso občutljivi na vsebino, ampak samo na sintakso: modus ponens ($p \rightarrow q$, p, $\therefore q$) je npr. takšna računska operacija, pri kateri se možnost izpeljave iz predstave p -> q skupaj s predstavo p do predstave q, ne nanaša niti na pomen obeh propozicij niti na pomen veznika. »Z vidika formalnega procesa je pri uporabi *modusa ponensa* pomembno le to, da ima p v drugi premisi isto obliko kot p na levi strani prve premise in da ima q v sklepu isto obliko kot q na desni strani prve premise« (Markič, 2010: 37). To sistemu daje prostor, da spremembe mentalnih stanj npr. definira v pojmih vzročno-funkcionalnih vlog, ki so implementirane z njegovo sintakso ter tako realizira mentalnost/inteligentnost ne glede na svojo naravo.

(iii) Reprezentacijska teorija duha pa je sestavljena iz dveh predpostavk. Prva trdi, da lahko *intencionalna stanja*, tj. stanja, ki se na nekaj nanašajo (predmet ali situacijo) ali so na nekaj usmerjena, kar imenujemo njihova *vsebina*, zaradi česar so stanja *»o* nečem«, opredelimo s t.i. *propozicionalnimi naravnanostmi*, izrazom, ki se nanaša na povedi, v katerih v glavnem stavku tipično nastopajo glagoli verjeti, želeti ali upati, njim pa sledi predmetni odvisnik z veznikom ‘da’, npr. »Želim, da bi nehalo deževati«. Tako so propozicionalne naravnanosti sestavljene iz relacije med subjektom, tj. tistim, ki si »želi« (ima to naravnanost) in propozicijo (»da bi nehalo deževati«), ki je identična s vsebino našega mentalnega stanja. Ali: »Za katerikoli organizem *O* in katerokoli naravnanost do propozicije *P* obstajata (‘računska’/‘funkcionalna’) relacija *R* in mentalna reprezentacija *MP*, ki sta takšni, da *MP* pomeni *P* in da *O* ima *A*, če in samo če, je *R* med *0* in *MP«* (Podor, 1987:17).

Poleg te trditve, ki se nanaša na *naravo propozicionalnih naravnanosti*, pa reprezentacijska teorija duha vsebuje še eno, ki se dotika *narave mentalnih*

procesov, ki so v njej opisani kot vzročna zaporedja primerkov mentalnih predstav. Zaradi tega je sosledje misli v bistvu vzročno zaporedje primerkov mentalnih predstav, primerki mentalnih predstav pa izražajo propozicije, ki so objekti misli. Npr. »Misliti 'Deževalo bo. Torej grem noter.' je uprimeriti mentalno predstavo, ki pomeni *Grem noter*, povzročeno na določen način z uprimerjanjem mentalne predstave, ki pomeni *Deževalo bo«* (Podor, 1987: 17).

Turing je v svojem članku *Stroji, ki računajo in inteligenca* iz leta 1950 pompozno trdil, da v primeru, ko se »verbalno« obnašanje stroja, ki poteka preko ekrana in tipkovnice, ne razlikuje od verbalnega obnašanja človeka, ko odgovarjata na ista vprašanja, stroj, ki je tega sposoben, *misli*.[9] No, računska teorija duha je v resnici obrnjena Turingova trditev, da je, če mu to uspe, računalnik duh, pravi namreč, da je mišljenje računski proces oziroma da je duh računalnik. Kakorkoli, njena razlaga tega, kako lahko možgani proizvedejo misel, ki črpa svoj navdih iz stavčne logike (teorije dokaza) in Turingove ideje, da so formalni postopki lahko mehanizirani, je še danes osrednja tema kognitivne znanosti in temelj umetne inteligence, ki pa ima tako svoje zagovornike kot tudi nasprotnike.

Ugovori funkcionalizmu

Vrnimo se k najbolj splošni trditvi/obliki funkcionalizma, ki pravi, da so mentalna stanja sestavljena iz dispozicije delovati na določen način in drugih mentalnih stanj, kar je posledica določenih čutnih in mentalnih vnosov (Block, 1980c: 268), ter si oglejmo nekaj težav, ki ga spremljajo. Ugovore, ki izhajajo iz njegove predpostavke, da ima mentalno življenje vse, kar izpolni ali realizira ustrezno vzročno-funkcionalno vlogo, lahko razvrstimo v dve skupini: (i) v prvo spadajo tisti, ki pravijo, da iz tega sledi podobnost sistemov, za katere niti v sanjah ne bi rekli, da imajo skupnega karkoli bistvenega za njihovo naravo, (ii) v drugo pa spadajo tisti, ki trdijo, da iz tega sledi priznavanje zavesti sistemom, katere niti slučajno ne dojemamo kot mentalne, kar kaže na to, da se pri funkcionalizaciji človeške duševnosti izgubi ena od njenih bistvenih značilnosti.

(i)

Ugovor iz funkcionalne izomorfnosti

Funkcionalizem predpostavlja, da sta dva stroja izomorfna takrat, kadar izpolnita isti funkcionalni opis ne glede na to, v kolikšni meri se razlikujeta glede svojih strukturalnih lastnosti. To pa vodi do govorjenja o funkcionalni izomorfnosti sistemov, ki si z vidika funkcioniranja še zdaleč niso podobni, saj je treba pri njihovem definiranju upoštevati tudi opis na nižjih ravneh abstrakcije (Kalke, 1969). Vzemimo mačko in mišolovko. Če ju opišemo dovolj abstraktno, obe realizirata isto funkcionalno shemo, in sicer *P* (pojav miši – vnos), *x* (notranje stanje – »črna škatla«), *Q* (lovljenje miši – iznos). Če pa ju opišemo bolj

[9] Za razpravo o tem *glej* Bregant (2011, 2012).

podrobno, tj. osredotočimo se na to, kaj se dogaja znotraj mačke in kaj znotraj mišolovke, takoj ugotovimo, da sta povsem različni. Sklep je, da sta dva sistema lahko funkcionalno izomorfna zgolj na visoki ravni abstrakcije, pri podrobnejšem opisu njune funkcionalne organizacije, pa se izkažeta za bistveno drugačna: očitno je, da je funkcionalna shema pri živih organizmih bolj zapletena, kot pa pri neživih artefaktih.

Poanta Kalkejevega argumenta zoper funkcionalizem je tako naslednja: bolj, ko je nek stroj abstraktno opisan, tj. več kot je stvari, ki se ne upoštevajo, večja je verjetnost, da bomo našli stroj, ki mu je funkcionalno podoben; in manj, ko je nek stroj abstraktno opisan, tj. manj kot je stvari, ki jih šteje za nepomembne, manjša je verjetnost, da obstaja stroj, ki mu je funkcionalno izomorfen. In govoriti, da je psihologija danes neodvisna znanost, ker pojasnjuje naše obnašanje s pomočjo prikazanih funkcionalnih shem, že dolgo ni več resnično, zadnje čase se vse bolj opira na strukturalne lastnosti sistemov, tj. na nevrofiziološko organizacijo naših možganov (Churchland, 2011), ki pa jo funkcionalizem povsem zanemarja.

Ugovor iz splošnih in abstraktnih sistemov

Funkcionalizem trdi, da so mentalna stanja definirana s svojim kontekstom in da spadajo v isto naravno vrsto izključno na osnovi mesta, ki ga imajo v funkcionalni shemi, ne glede na razlike v svojih fizičnih značilnostih. To pa pripelje do tega, da klasificiramo kot pojave iste vrste tudi pojave, ki nimajo med sabo skupnega ničesar razen, da so primerki iste funkcionalne sheme. Vzemimo funkcionalni model (Sesardić, 1984), ki ga lahko uprimerita tako mehanski kot tudi elektronski sistem in ki opisuje oscilacije, v enem primeru mehanske, v drugem pa elektronske: $A\ (d^2x/dt^2) + B\ (dx/dt) + Cx = D$. Skladno s kontekstom simbolom v tej enačbi pripišemo različne pomene, oba sistema pa se s funkcionalnega vidika kljub temu obnašata izomorfno, zaradi česar se lahko mehanski in elektronski parametri prikažejo z istim funkcionalnim modelom. Predpostavimo, da imajo simboli v teh dveh sistemih naslednje pomene:

	A	B	C	D	x
1. sistem (mehanski):	masa	koeficient trenja	togost	sila	razdalja
2. sistem (elektronski):	električni tok	električni upor	kapacitivnost $^{-1}$	električni potencial	naboj

Problem je, da bi glede na funkcionalistični princip, ki razvršča stanja v vrste zgolj na osnovi položaja, ki ga v abstraktnih shemah zasedajo, v prikazani funkcionalni shemi v isto vrsto spadala npr. razdalja in naboj, česar pa noben resen teoretik ne bi zastopal, saj si nista podobna v nobeni bistveni podrobnosti, na osnovi katere bi ju lahko razvrstili v isto vrsto. Da bi ju tja uvrstili, bi o obeh morale obstajate zanimive posplošitve, ampak o razdalji in naboju jih ni, edino, kar imata skupnega, je, da uprimerjata isti funkcionalni model, da pa je to že dovolj za to, da sta tudi stanji iste naravne vrste, je, milo rečeno, prehuda

poenostavitev. To se zgodi zato, pravi Gendron (1971), ker funkcionalizem svoj model predstavlja kot *splošen*, kot sta npr. Newtonova mehanika ali Maxwellova elektromagnetika, ki opredeljujeta obnašanje posameznih primerkov in vnaprej določata, katere vrste so, v resnici pa je *abstrakten*, kot je npr. McCullochov in Pittsov logični model nevrona, ki pa ne pojasnjujeta obnašanja posameznih realizatorjev in ne določata, v katero vrsto spadajo.

Z vidika psihologije kot posebne znanosti definiranje človeške duševnosti v pojmih vzročno-funkcionalnih vlog tako ni zanimivo, saj ne vodi do nobenih posplošitev o pojavih, kar bi nam zagotavljalo klasifikacije v vrste in prispevek k razumevanju sveta, ampak gre zgolj za računanje s simboli, kar pa je domena matematike, ki ji tako jemlje njeno neodvisnost.

(ii)

Ugovor iz zombijev[10]

Obstaja več definicij zavesti, najbolj zanimiva in za funkcionalizem najbolj problematična pa je tista, ki pravi, da ima bitje zavest, ko njegove čutne interakcije z okoljem vsebujejo tudi kvalitativna subjektivna izkustva tega, kar zaznava, tj. *takšnosti* ali *kvalije*. Gre za tipične fenomenalne značilnosti občutkov, ki se običajno opisujejo kot načini čutenja, pojasnjujejo pa s pomočjo stavkov ‘kako je biti’, npr. kako je videti rdeč paradižnik ali kako je čutiti bolečino v kolenu. V zvezi z njimi je problematično predvsem to, kako naj bodo opisane, da bodo ustrezno predstavljene drugim, saj se predpostavlja, da to, ‘kako je biti’, ve samo tisti, ki to izkusi.

Kaj imajo torej vsi primerki bolečine skupnega, da so vsi tudi primerki bolečine? Funkcionalistov odgovor na to vprašanje je jasen, in sicer tipične vzroke, tj. fizične dražljaje ali mentalna stanja, npr. poškodbo ali travmo, in tipične učinke, npr. kričanje in željo po prenehanju. Vendar se zdi, da mentalni pojavi, za katere so značilne takšnosti, *fenomenalni pojavi*, vsebujejo še nekaj več od golega opisa v pojmih njihovih vzročno-funkcionalnih vlog. Kar imajo vsi primerki bolečine skupnega, da so vsi tudi primerki bolečine, je, da vsi bolijo. Kaj to pomeni? To pomeni, da takšnosti ne moremo opisati v funkcionalnih pojmih, pravi Jackson (1995). Pri tem se sklicuje na primer z Marijo, ki ve o izkustvu rdečega vse, kar se da o njem vedeti na osnovi funkcionalnih informacij, pa o izkustvu rdečega kljub temu nima popolnega znanja. Manjka ji takšnost rdečega, ki pa jo dobi šele takrat, ko sama vidi nekaj rdečega. Podobno velja tudi za bolečino. Opis bolečine kot vzburjenosti živčnih vlaken, ki povzroča kričanje in željo po prenehanju, nam o tem, kako je čutiti bolečino, ne pove ničesar. O izkustvu bolečine tako nimamo popolnega znanja, saj nam manjka njena takšnost, ki pa jo dobimo šele takrat, ko nas nekdo npr. udari (Bregant, 2004: 218).

Funkcionalistove muke pa lahko še zaostrimo z naslednjim miselnim

[10] Imenujemo ga lahko tudi *problem dvojnikov*.

eksperimentov, ki ima svoje zametke že pri Descartesu (1637/2007), in njegovem zanikanju, da bi lahko bili avtomati kadarkoli zavestni, razvil pa se je v 70. letih 20. stol. kot orožje za boj proti fizikalizmu nasploh.[11] Predstavljajte si bitje, ki je molekulski duplikat vas samih, tj. s fizičnega vidika je z vami povsem identično, nima pa nobene zavesti oziroma občutkov. Gre za t.i. filozofskega zombija, ki se razlikujejo od filmskega zombija: filmski je neroden, nad svojim obnašanjem nima nadzora, prehranjuje se z ljudmi in ni nujno brez takšnosti, filozofski pa se od nas razlikuje zgolj v tem, da nima fenomenalnega izkustva, kar pomeni, da se tudi obnaša na povsem enak način kot njegov fizični izvirnik. Npr. ko sediva jaz in moj zombi na plaži, se obnašava povsem enako, oba pijeva, kot da bi bila žejna, oba naročiva vino, kot da bi marala njegov okus, oba uživava v razgledu, kot da bi imela vidno izkustvo itn. Razlika je zgolj v tem, da jaz resnično vse to izkušam, moja fizična kopija pa ne, čeprav se v naju, ker sva postavljena v isto okolje, odvijajo enaki fiziološki procesi. Katerikoli dražljaj, ki se pojavi kot vnos, bo moja kopija obdelala na enak način kot jaz, kar pomeni, da bo tudi njeno obnašanje kot iznos enako kot moje. Skratka, vse pri njem je enako kot pri meni razen, da zanj ne obstaja to ‘kako je biti’ v nekem stanju.

Zamisljivost zombijev je tako za funkcionalizem problematična, ker zombiji, tako kot funkcionalizem to zahteva, opravljajo funkcionalno vlogo, opisano v pojmih tipičnih vzrokov in učinkov, npr. ko so žejni, pijejo, ne da bi sploh imeli kakršnokoli mentalno stanje; z drugimi besedami, identifikaciji fenomenalnih pojavov z njihovimi vzročnimi vlogami se izmuzne kvalitativna značilnost, ki jo tipično imajo, zaradi česar funkcionalizem *odsotnosti takšnosti*, kot se ta problem tudi opisuje, ne more pojasniti. Seveda takšni zombiji niso dejansko ali *nomološko* možni, tj. v aktualnem svetu z znanimi naravnimi zakoni jih ni, so pa vsaj zamisljivi in tako logično ali *metafizično* možni.[12]

Ugovor iz obrnjenega spektra

Podobno težavo predstavlja za vse, ki pravijo, da je naš duh zgolj funkcionalna organizacija naših možganov, še en miselni eksperiment, katerega začetki segajo k Locku (1689/2012), orodje za zavračanje funkcionalizma pa je prav tako postal v 70. letih 20. stol.[13] Zamislite si Janeza, ki vidi rdeče, ko vi vidite zeleno in ki vidi zeleno, ko vi vidite rdeče, tj. ko gleda rdeč paradižnik je za njega ‘kako je biti’ tako kot za vas, ko gledate zeleno travo, in ko gleda zeleno travo je za njega ‘kako je biti’ tako kot za vas, ko gledate rdeč paradižnik. To imenujemo *obrnjeni barvni spekter* ali *obrnjena takšnost* (Block in Podor, 1972), zanj pa je značilno, da ga ima Janez od rojstva, da se dogaja v isti jezikovni skupnosti in da se Janez obnaša na enak način kot vi. Z drugimi besedami, Janez in vi sta fenomenalno različna, tj. vajini subjektivni kvalitativni izkustvi sta obrnjeni, vendar

[11] *Glej* npr. Shoemaker (1975), Block (1980c), Chalmers (1996).
[12] Vsi teoretiki se s tem ne strinjajo, za nezamisljivost zombijev *glej* npr. Harnad (1995), Tye (2006), Dennet (1991/2012).
[13] *Glej* npr. Block in Fodor (1972), Block (1980c, 1990), Shoemaker (1982).

funkcionalno enaka, tj. na vprašanje, kakšne barve je paradižnik, oba odgovorita, da rdeče, pa tudi z vrta oba vedno prineseta le zrel paradižnik. To je t.i. klasični/tradicionalni/običajni argument iz obrnjenega spektra, ki za funkcionalista predstavlja težavo, ker isti vzrok v vzročni specifikaciji vidnega izkustva povzroči isti učinek, čeprav imata Janez in vi različni takšnosti, kar kaže na to, da zavestna stanja niso del vzročno-funkcionalnih vlog oziroma, da funkcionalizem obrnjene takšnosti ne more pojasniti. Podobno kot zombiji je tudi obrnjen spekter zgolj metafizično možen, tj. mogoče si ga je zamisliti in to enostavno, ne glede na to, kako težko bi ga bilo v praksi v našem svetu izvesti, tj. da bi postal tudi nomološko možen.

Razen tega bolj lockovsko obarvanega obrnjenega spektra pa obstaja še en, ki si ga je izmislil Block (1990) in ga imenoval *obrnjena Zemlja*. Zamislite si drugo Zemljo, kjer je nebo resnično rumeno, zrel paradižnik zelen in trava resnično rdeča. Njeni prebivalci imajo tako drugačno fenomenalno izkustvo, ko pogledajo v nebo, jedo paradižnik in kosijo travo, kot njihovi sorodniki na normalni Zemlji. Razen tega so tudi vsebine vzročno-funkcionalnih vlog prebivalcev te druge Zemlje obrnjene: mislijo, da je nebo rumeno, paradižnik zelen in trava rdeča. Vse ostalo je na obrnjeni Zemlji prav takšno kot na normalni. Sedaj pa predpostavite, da vas neko noč znanstveniki ugrabijo, opremijo z lečami, ki izničijo ali nevtralizirajo učinek obrnjenega spektra in z normalne Zemlje pošljejo na obrnjeno. Zaradi leč, ki jih nosite, bodo vaše takšnosti enake kot doma, nebo boste videli modro, zrel paradižnik rdeč in travo zeleno, z vidika ‘kako je biti’ za vas ne bo nobene spremembe, v bistvu boste mislili, da ste še vedno doma. Toda ko bo minilo dovolj časa in ko boste posvojili njihovo poimenovaje ter se zlili z njihovo kulturo in fizičnim okoljem, se bodo vsebine vaših vzročno-funkcionalnih vlog začele ujemati z njihovimi, tj. tudi vi boste mislili, da je nebo rumeno, paradižnik zelen in trava rdeča. Skratka, ko se boste dovolj integrirali, se boste obnašali enako kot oni, s čimer se bo vaša funkcionalna vloga v primerjavi z vami, kot ste bili prej, spremenila, čeprav bodo subjektivna izkustva (še vedno nosite leče) ostala ista, zaradi česar gre za, recimo, *nasprotni obrnjeni spekter*. Vendar tudi ta enako kot prejšnji pokaže, da »’kako je biti’ ne more temeljiti na funkcionalnih značilnosti vaših mentalnih stanj« (Tye, 1995: 29).

Kakorkoli, funkcionalistična ideja, da se danes pojem mentalnih stanj nanaša na formalne vzorce, ne pa na njihove materialne elemente, je postala osnova za govorjenje in ustvarjanje inteligentnih strojev, o obstoju katerih pa si vsi teoretiki še zdaleč niso edini, kar si bomo ogledali v nadaljevanju.

2 »Kitajska soba«

V 90. letih 20. stol. je Edward Fredkin, ki je bil takrat direktor laboratorija za umetno inteligenco (UI) na MIT-ju, menil, da so se v zgodovini zgodili trije pomembni dogodki. Prvi je bil nastanek vesolja, drugi nastanek življenja. Tretji, za katerega je trdil, da je enako pomemben, pa je bil nastanek umetne inteligence

(Copeland, 1993: 1). S pojmom ‘inteligenca’ mislimo na izvajanje umskih/spoznavnih operacij, s pojmom ‘umetna’ pa na to, da jih ne izvaja človek, ampak stroj/sistem (tj. nekaj, kar ni rezultat narave, ampak človekovega ali strojevega dela). Odgovora na to, kaj ga je pripeljalo do tega, da je nastanek umetne inteligence postavil ob bok nastanku vesolja in življenja, ni težko najti, saj gre za obliko življenja, ki je povsem drugačna od običajne in s potencialom, ki si ga težko predstavljamo.

Umetno inteligenco lahko najenostavneje definiramo kot znanost o izdelavi mislečih strojev, ki so sposobni reproducirati različne zmožnosti našega duha. Način, kako jim to uspe, ni pomemben, zaradi česar je za to početje zelo prikladna osnovna funkcionalistična ideja o tem, da so mentalna stanja stanja Turingovega stroja. Ker si ni težko zamisliti, da se bodo stroji razvijali, s tem postajali pametnejši in nekega dne samostojno načrtovali ter izdelovali druge stroje, je povsem na mestu vprašanje, kaj se bo zgodilo s svetom, v katerem ima prvenstvo človek kot najbolj inteligentno bitje? Strah, da bo to prevlado, če že ne bo prišlo do uničenja sveta, kot ga poznamo iz filmov, izgubil, izhaja iz tega, da si je težko zamisliti stroj, ki je pametnejši od najpametnejšega človeka, pa kljub temu dela zgolj tisto, kar od njega zahtevamo. Lažje si je predstavljati situacijo, kjer se takšen stroj poniža do te mere, da se z nami igra, pogovarja in druži, podobno, kot se mi obnašamo do naših domačih živali, ki nam krajšajo čas.

Iz tega pa izhaja kopica filozofskih vprašanj, izmed katerih so najbolj pomenljiva naslednja: (i) je misel biološki pojav in tako umetnim strojem nedosegljiva, podobno, kot to velja za fotosintezo (produkcija kisika) ali laktacijo (izločanje mleka), ali pa sta mišljenje in zaznavanje bolj podobna letenju, nečemu, kar lahko delajo tako biološki organizmi kot tudi umetni sistemi; (ii) ali bi lahko postal računalnik kadarkoli pametnejši od ljudi, ki so ga programirali; (iii) ali je lahko stroj v svojih odločitvah svoboden, tj. ali so njegova dejanja posledica njegove lastne volje; (iv) ali bi računalnik lahko imel zavest, tj. ali bi lahko imel subjektivna kvalitativna izkustva ‘kako je biti’; (v) ali nismo tudi mi zgolj neke vrste stroji (Copeland, 1993: 1–2).

V nadaljevanju se bomo osredotočili zgolj na prvo, pri čemer si bomo ogledali slavni Searlov argument iz »kitajske sobe«, ki so ga leta 1980 ob izidu njegovega članka *Duhovi, možgani in programi*, nasprotniki inteligentnih strojev označevali kot »popolno zmagoslavje«, njihovi zagovorniki pa kot nekaj »preprosto smešnega«. Pri tem ni odveč omeniti, da so od rojstva umetne inteligence,[14] leta 1956 na slavni darthmouthski konferenci, mislečim strojem vedno bili bolj naklonjeni računalničarji (programerji), kot pa filozofi. Turing je npr. že leta 1950 v svojem članku *Stroji, ki računajo in umetna inteligenca* menil, da bodo danes inteligentni stroji nekaj običajnega.[15] Ali lahko stroj misli? To vprašanje se ni prvič pojavilo šele v 20. stol., ko so bili računalniki zreli za masovno

[14] Prvič naj bi se ideja o umetni inteligenci sicer pojavila v McCulloch in Pitts (1943).
[15] Za to *glej* Bregant (2012).

proizvodnjo, ampak že daljnega leta 1637, ko je Descartes trdil, da noben stroj nikoli ne bo sposoben simbolne manipulacije, ki bi mu omogočala tvorjenje smiselnih stavkov, na osnovi česar bi mu lahko pripisali mišljenje (Bregant, 2010:55). Šele po stoletjih čakanja s prihodom prvih elektronskih računalnikov v 50. letih 20. stol. je ideja o mislečem stroju dobila nov zagon. Od takrat je sicer doživela vzpone in padce, hiter napredek informacijskih tehnologij pa jo še vedno, mogoče celo bolj kot kdajkoli prej, dela živo.

Od začetka obstajata v splošnem dve glavni pojmovanji/razumevanji umetne inteligence. Po prvem je njen cilj zgraditi stroj, ki ima vse običajne (človeške) psihične lastnosti, kot so mišljenje, razumevanje in sklepanje, še več, celo zavest, čustva in občutke, skratka, *ustrezno programiran računalnik bi resnično imel duh* v *istem smislu kot ga imate vi ali jaz*, kar imenujemo *močna umetna inteligenca*. Po drugem, manj skrajnem, pa je njen cilj zgraditi stroj, ki lahko opravi enake stvari kot človek, s to razliko, da človeku, ko jih naredi, pripišemo psihične lastnosti, stroju pa ne, skratka, *programirani računalniki zgolj simulirajo mišljenje*,[16] kar imenujemo *šibka umetna inteligenca*. Searlov argument iz »kitajske sobe« je uperjen proti močni umetni inteligenci, katere osrednjo trditev sam opredeli takole: »[Tukaj] računalnik ni le orodje za raziskavo duha; tu je ustrezno programiran računalnik *sam* duh, v tem smislu, da za računalnike, ki jih ustrezno programiramo, lahko rečemo, da dejansko razumejo in imajo še druga spoznavna stanja« (Searle, 1980/1990:361).

Vprašanje je, ali sploh obstaja kdo, ki je to pripravljen zagovarjati? Dvom v to ni na mestu, saj se zdi, da to sprejemajo vsi ustanovitelji umetne inteligence vključno s Turingom,[17] kasneje pa tudi npr. Newell (1980), Dennett (1985), Kurzweil (1998) itd. Zdi se, da »je *kdorkoli*, ki *resno* opisuje elektronski digitalni računalnik v psiholoških pojmih ali predlaga reprodukcijo psihičnih pojavov v takšnih računalnikih, zavezan močni UI« (Preston, 2002: 16). Enako naj bi veljalo tudi za Schanka in Abelsona, ki sta v 70. letih 20. stol. na Yaleu razvila program z imenom *Sam*, ki je bil sposoben brati med vrsticami, tj. bil je sposoben v zgodbi zapolniti luknje oziroma prepoznati manjkajoče podrobnosti. To se zdi nujno za *razumevanje* kateregakoli teksta in predstavlja nekaj, s čimer ljudje, npr. v novinarskih člankih, kjer se to običajno dogaja zaradi varčevanja s prostorom, nimamo težav. Na vprašanje znamo odgovoriti pravilno tudi takrat, ko odgovor v zgodbi ni izrecno naveden. Enako sposobnost je imel tudi Sam, ki je imel vgrajene različne scenarije/besedila za različne situacije, t.i. *skripte*. Z njimi je razširil pomanjkljivo zgodbo in potem brez težav zapolnil manjkajoče podrobnosti v njej, kar mu je omogočalo, da je pravilno odgovoril na vsa vprašanja v zvezi z njo in tako prestal preizkus razumevanja. Z drugimi besedami, stroj, na katerem bi tekel takšen program, bi *razumel* zgodbo na enak

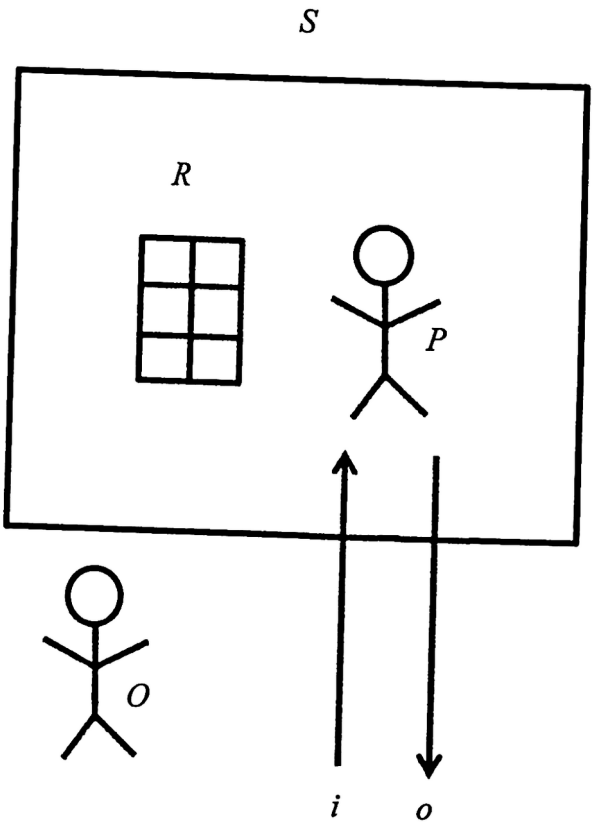
[16] Ta simulacija je sicer koristna za spoznavo tega, kako duh deluje, vendar iz tega še ne sledi, da je tudi sam računalnik duh.

[17] Turing sicer ne velja za očeta umetne inteligence, ampak za začetnika filozofije umetne inteligence (Copeland, 1993).

način, kot jo razumemo sami, zaradi česar bi morali razumevanje pripisati obema ali nikomur.[18] Searlov argument iz »kitajske sobe« naj bi pokazal, da je takšno prepričanje zmotno.

V svoji izvirni različici poteka njegov miselni eksperiment takole. Zamislite si Searla, ki ne obvlada nobenega drugega jezika razen angleščine, zaprtega v sobi z odprtino, skozi katero dobi tri svežnje kitajskega besedila: prvi predstavlja t.i. *skripte*, tj. scenarije za različne situacije, drugi predstavlja *zgodbo*, vključuje pa še pravila v angleščini, ki Searlu omogočajo, da poveže formalne simbole iz skriptov s temi iz zgodbe, pri čemer 'formalen' pomeni, da jih lahko spozna izključno po njihovih oblikah, tretji pa predstavlja *vprašanja*, na katera mora odgovoriti in prav tako vključuje pravila v angleščini, ki pa mu tukaj omogočajo, da poveže formalne simbole iz skriptov in zgodbe s temi iz vprašanj. Torej, Searle ima v sobi na voljo tri svežnje kitajskih pismenk, tj. skripte, zgodbo in vprašanja ter pravila za njihovo povezovanje. Ljudje zunaj sobe sedaj od njega pričakujejo, da jim skozi odprtino posreduje odgovore na vprašanja, ki jih je dobil, zamislite pa si, da se Searle v rokovanju s simboli sčasoma tako izuri, da se ob odličnih programih, ki jih pri tem uporablja, njegovi odgovori v kitajščini na dana vprašanja za nekoga, ki stoji zunaj sobe, ne razlikujejo od odgovorov rojenih Kitajcev. Slikovno lahko to prikažemo takole (Bringsjord in Noel, 2002:148): S

[18] Vprašanje je, ali so res filozofi poklicani za to, da o tem sodijo? Res je sicer, da nimajo vpogleda v tehnične podrobnosti računalnikov in ne vedo, česa vsega so sposobni in kdaj, kljub temu pa lahko ocenjujejo, kateri je najboljši način za opisovanje psihičnih sposobnosti v ne-tehničnih pojmih in z miselnimi eksperimenti zavračajo tiste, ki se izkažejo za neprimerne (Preston, 2002: 16).



Slika 1: *S* označuje sistem, *R* pravila, *P* osebo, tj. Searla, *O* opazovalce, tj. rojene Kitajce, *i* vnose in *o* iznose.

Vprašanje je, ali Searle, ki deluje kot računalniški program, tj. izvaja operacije na formalno opredeljenih elementih oziroma proizvaja kitajske odgovore tako, da uporablja neraztolmačene formalne simbole, razume kitajsko? Odgovor je slej ko prej *ne*. Čeprav se njegovi vnosi in iznosi ne razlikujejo od tistih rojenega Kitajca in čeprav lahko izvede katerikoli naložen program, je in bo med njima vedno velika razlika: Searle ne razume, rojeni Kitajec pa razume. »Iz istega razloga tudi Schankov računalnik ne razume ničesar iz nobene zgodbe, pa najsi bo ta v angleščini, kitajščini ali kateremkoli drugem jeziku, saj je v kitajskem primeru računalnik jaz, v primerih pa, ko ni jaz, računalnik nima ničesar več, kot imam jaz v primeru, ko ne razumem ničesar« (Searle, 1980/1990: 363–364). (To, mimogrede, predstavlja ugovor funkcionalizmu, saj imamo v tem primeru dva sistema, Searla in rojenega Kitajca, ki sta funkcionalno enaka, pa kljub vsemu psihološko različna, tj. eden razume, drugi pa ne, kar je s funkcionalizmom, ki pravi, da iz funkcionalnega izomorfizma sledi tudi psihološki, nezdružljivo: opravljati prave funkcije, vsaj kadar so določene zgolj z vnosi in iznosi, še ni dovolj za posedovanje mentalnih stanj).

Skratka, Searlov miselni eksperiment pokaže, da obvladovanje slovnice/sintakse,

tj. poznavanje niza pravil za simbolno manipulacijo, še ne pomeni obvladovanja pomena/semantike, tj. poznavanje tega, kar simboli dejansko pomenijo. To pa je v nasprotju z močno umetno inteligenco, ki pravi, da ustrezno programirani računalniki tudi razumejo, zaradi česar jih lahko prištevamo med inteligentne stroje, ki so sposobni izvajati tudi druge kognitivne operacije in imajo poleg intencionalnih stanj tudi fenomenalna: »/.../ *v* dobesednem pomenu programirani računalnik razume toliko, kot razumeta avto in seštevalni stroj, to pa je natanko nič. Računalnikovo razumevanje ni niti delno niti nepopolno (kot je moje razumevanje nemščine); je povsem nično« (Searle, 1980/1990: 366).

3 Ugovori argumentu iz »kitajske sobe«

Ko želimo govoriti o »kitajski sobi« kot o argumentu, takoj naletimo na dve težavi, prva je, da Searle svojega miselnega eksperimenta nikoli ne zapiše v obliki argumenta, druga pa, da zaradi tega ni povsem jasno, proti čemu je uperjen. Če bi »kitajsko sobo« skušali predstaviti kot argument, bi ta verjetno bil videti takole (Preston, 2002: 26):

1. Oseba v sobi ima dostop zgolj do formalnih, sintaktičnih značilnosti simbolov, ki so ji posredovani.

2. Da bi razumela kitajske iznose, bi morala imeti dostop do semantičnih značilnosti njihovih vnosov.

3. Noben niz formalnih ali sintaktičnih načel ni zadosten za razumevanje.

/. 4. Oseba v sobi ne razume kitajskih vnosov.[19]

Problem je, da takšen sklep, ker se nanaša zgolj na osebo v sobi, ne zavrača močne umetne inteligence, saj ta ne trdi, da razumejo vsi deli ustrezno programiranega računalnika, ampak takšen stroj kot celota, zaradi česar zagreši *zmoto sestave*, kar bomo podrobneje predstavili v nadaljevanju.

Kakorkoli, glavna tarča argumenta iz »kitajske sobe« je nedvomno močna umetna inteligenca,[20] vprašanje je samo, v kakšnem smislu. Searle namreč ne zavrača splošne ideje, da lahko stroji mislijo, saj je prepričan, da smo tudi ljudje zgolj neke vrste stroji: »Ali lahko stroj misli? Odgovor je očitno – da. Mi sami smo natanko takšni stroji« (Searle, 1980/1990: 375). Razen tega meni, da lahko mislijo celo umetno narejeni stroji, tj. takšni, ki jih je naredila človeška roka, če bi bili z vidika živčne zgradbe dovolj podobni nam. Sprejema celo predstavo, da lahko mislijo tudi računalniki, saj smo tudi mi samo neke vrste računalniki: »/…/

[19] Kasneje Searle (npr. 1997) svoj argument iz »kitajske sobe«, t.i. »brutalno preprost argument«, predstavlja takole:

1. Programi so povsem formalni (sintaktični).
2. Duh (vsaj človeški) ima semantiko, tj. mentalno vsebino.
3. Sintaksa sama po sebi ni niti ista kot semantika niti ni zanjo zadostna.

∴ 4. Programi sami po sebi niso niti bistveni niti zadostni za duh.

[20] Razen tega je uperjen še proti Turingovemu testu (*glej* npr. Harnad, 2002, Coulter in Sharrock, 2002) in kot smo že na kratko omenili, tudi funkcionalizmu (*glej* npr. Rey, 2002, Block, 2002).

ali lahko misli digitalni računalnik? Če z ‘digitalnim računalnikom’ mislimo karkoli, kar lahko opišemo kot izvedbo računalniškega programa, potem je odgovor seveda spet pritrdilen, saj smo tudi mi sami ‘realizacija’ neke vrste računalniškega programa in mi že lahko mislimo« (Searle, 1980/1990: 375).

Tisto, čemur Searle nasprotuje, je, da ima lahko nekaj psihične lastnosti *izključno* zaradi tega, ker je programirano na ustrezen način oziroma, da je biti sposoben izpeljati ali izvesti primeren program, zadostni pogoj/razlog za pripisovanje inteligence. »Toda ali lahko nekaj misli, razumeva in tako naprej *že samo* zaradi tega, ker je računalnik s pravo vrsto programa? Ali je lahko izvedba programa, seveda če je ta pravi, že sama po sebi zadosten pogoj za razumevanje? To pa, mislim, je tisto vprašanje, ki si ga je treba zastaviti, čeprav ga običajno mešajo s prejšnjimi vprašanji, odgovor nanj pa je – ne« (Searle, 1980/1990: 375). Razlog za njegovo nestrinjanje s tem smo že omenili, gre pa za to, da formalni program nima nobene intencionalnosti, je vsebinsko prazen in brez smisla, zaradi česar sistemu ne more zagotoviti razumevanja, če ga ta ni imel že prej: »Intencionalnost, kakršno naj bi imeli računalniki, obstaja samo v glavah tistih, ki jih programirajo, in tistih, ki jih uporabljajo, tistih, ki vnašajo input in tistih, ki tolmačijo output« (Searle, 1980/1990:376).

Sistemski ugovor

Čeprav Searle »kitajske sobe« kot argumenta ni zapisal v prej predstavljeni obliki, pa nekateri avtorji menijo, da v miselnem eksperimentu oseba v sobi ni analogna računalniku (Copeland, 1993, Block, 2002, Haugeland, 2002), kot Searle običajno trdi, ampak zgolj *procesorju*,[21] tj. osrednjemu delu računalnika, ki obdeluje (procesira) podatke ter nadzoruje in upravlja ostale enote. Če to drži, argument iz »kitajske sobe« zgreši svojo tarčo, saj se trditve močne umetne inteligence nanašajo na računalnik kot celoto, ne pa samo na enega izmed njegovih delov, če pa vztrajamo, da se nanaša na celoten sistem, zagrešimo, kot smo že omenili, zmoto sestave, kar se lepo vidi iz njegovega standardnega zapisa:

1. Nobena simbolna manipulacija, ki jo oseba v sobi opravi, ji ne bo pomagala, da bi razumela kitajski vnos.

/. 2. Nobena simbolna manipulacija, ki jo oseba v sobi opravi, ne bo širšemu sistemu, katerega del je, pomagala, da bi razumel kitajski vnos (Copeland, 1993: 125).

Ker v sklepu nastopa pojem ‘sistem’, o katerem ni v premisah ne duha ne sluha, je argument očitno neveljaven, tj. iz premis ta sklep preprosto ne sledi. »Logični odgovor«, kot sistemski ugovor imenuje Copeland (1993), Searlu očita osredotočenje na napačnega akterja, vprašanje ni, ali razume procesor, ampak, ali razume računalnik. Sistemski ugovor Searle povzame kot »Res je sicer, da oseba,

[21] Centralno procesna enota (CPE).

ki je zaklenjena v sobi, ne razume zgodbe, vendar je res tudi to, da je sama le del celotnega sistema, ta sistem pa zgodbo razume« (Searle, 1980/1990: 366). Vprašanje je, ali zagovorniki sistemskega ugovora to tudi resnično trdijo?

Preden na to odgovorimo, si oglejmo Searlovo zavrnitev sistemskega ugovora. Vsebuje tri pripombe, vsaka izmed njih pa naj bi dokazovala, da ne drži, (i) Pr... pravi, da nima sistem za razumevanje kitajskih simbolov na voljo nič takšnega, česar ne bi imela na voljo tudi oseba v sobi, tj. drugi deli sistema – knjige, papir, svinčniki itd., ki jih pri računanju uporablja, ne prispevajo k tolmačenju formalnih simbolov nič novega, zaradi česar bi celotni sistem razumel, njegov del (procesor) pa ne. (ii) Druga trdi, da je od sistema nesmiselno pričakovati, da se bo samo zaradi konjunkcije osebe in ostalih delov »kitajske sobe« (pravil, svinčnikov, papirja itd.) iz nečesa, kar ne razume, takoj prelevil v nekaj, kar razume. »Šlo naj bi za to, da oseba sicer resda ne razume kitajščine, vendar naj bi jo nekako razumela *zveza* te osebe in kosov papirja. Le stežka si predstavljam, da se lahko ta misel komurkoli, ki ni v primežu ideologije, sploh zdi verjetna« (Searle, 1980/1990: 366). (iii) Tretja pa pravi, da vodi sprejetje sistemskega ugovora v absurd, saj bi očitno nementalni deli sistema, kot npr. ledvice, jetra in želodec, naenkrat postali mentalni: res je, da obstaja (funkcionalna) raven opisa mojih jeter, kjer te obdelujejo vnosne podatke in izvajajo iznosne operacije, vendar to še ne pomeni, da karkoli razumejo.

Searle (1993) sicer meni, da je najmočnejši dokaz za to, da je sistemski ugovor ničvreden, prva, vprašanje pa je, ali je prepričljiv sploh katerakoli izmed naštetih. Copeland (1993) pravi, da je tarča Searlovih odgovorov nekaj, česar sistemski ugovor sploh ne trdi, in to je, da del sistema sicer ne razume, razume pa sistem kot celota. Poanta »logičnega odgovora« je zgolj ta, da je argument iz »kitajske sobe« neveljaven, tj. da iz dane premise ne sledi navedeni sklep. O njegovi resničnosti ne pove ničesar, še najmanj to, kar mu pripisuje Searle, zaradi česar njegovi odgovori niso uperjeni proti zagovornikom sistemskega ugovora, ampak proti nekomu drugemu. V tej luči je tudi njegova opazka o krožnosti »logičnega odgovora« neumestna, saj bi to morda držalo samo v primeru, če bi Searlu, ki skuša dokazati, da delujoči program ni sposoben razumeti, ugovarjali s tem, da oseba v sobi sicer ne razume kitajsko, razume pa sistem, katerega del je (kar Searle sicer neupravičeno domneva), saj bi tako res že na začetku predpostavili tisto, kar bi morali šele dokazati. Ker pa sistemski ugovor tega ne sprejema kot izhodišče, ni jasno, kako naj bi navedene tri pripombe predstavljale orožje za njegovo zavrnitev.

Poleg tega pa se zdi, da lahko Searla obtožimo tudi *zmote delitve*, ki se kaže v njegovi zadnji pripombi na »logični odgovor«. Če je bilo prej sporno pripisovanje nerazumevanja celoti na osnovi pripisovanja nerazumevanja njenemu delu, pa je sedaj zgrešena njegova trditev, da vodi sistemski ugovor v absurden nastanek mentalnih jeter, ledvic ali želodca. Iz tega, da je sistem mentalen, namreč še ne sledi, da so mentalni tudi vsi njegovi deli, podobno kot ne velja za vsako rimsko legijo, da je bila nepremagljiva, če je bila nepremagljiva

rimska armada. Ker je celota več kot zgolj seštevek njenih delov, je možno, da ostanejo »notranji organi« sistema nementalni, njihova specifična organizacija pa zagotavlja celoti novo lastnost, npr. mentalnost, ki je sami nimajo. V končni fazi med nami in računalniki z vidika osnovnih sestavnih delov, tj. kemičnih elementov, ni razlik, da se po prevladujočem mnenju v nas pojavi duh, v njih pa ne, mora to očitno biti posledica razlike v urejenosti med nami in njimi, saj bi v nasprotnem primeru bili primorani razmišljati o obstoju neke posebne substance v slogu Descartesovih nematerialnih duhov, ki so se po nekem čudnem naključju (ali morda božji intervenciji) znašli v človeških telesih, ne pa tudi v računalniških ohišjih.[22]

Ugovor z robotom

Kaj pa, če »kitajska soba« ne razume, ker gre preprosto za sistem (ki v slogu enega izmed ugovorov Turingovemu testu, in sicer *ugovora iz čutil*, ki pravi, da se ta omejuje zgolj na strojevo sposobnost dajanja verbalnih odgovorov, ne pa tudi na njegovo zmožnost povezati besede s stvarmi, ki jih označujejo), ki ni sposoben proizvesti vzročnih relacij med simboli in stvarmi, na katere se nanašajo. Če pa bi npr. Schankov računalnik dali v robota, kjer bi predstavljal njegove ‘možgane’, bi se takšno bitje odzivalo na dražljaje, odločalo glede na značilnosti okolja in obnašalo skladno z njimi, tj. opremljeno s kamero, ki zaznava zunanji svet, bi delovalo po načelu vzroka in učinka, zaradi česar bi mu lahko pripisali tudi pravo razumevanje. Searle ugovor z robotom povzame takole: »Recimo, da damo v računalnik robota, in ta računalnik ne bo le sprejemal formalnih simbolov skoz input in jih nato oddajal skoz output, temveč bo dejansko vodil robota, tako da bo ta opravljal funkcije, ki bodo zelo podobne zaznavanju, hoji, pomikanju, zabijanju žebljev, jedenju, pitju – kar pač že hočete. Ta robot bi recimo na sebi imel pritrjeno televizijsko kamero, s katero bi lahko videl, imel bi roke in noge, s katerimi bi lahko ‘nastopal’, in vse to bi nadzorovali njegovi računalniški ‘možgani’. Tak robot bi imel, za razliko od Schankovega računalnika, pravo zmožnost razumevanja in druga mentalna stanja« (Searle, 1980/1990: 369).

Searle meni, da je ugovor z robotom kontraproduktiven, saj s tem, ko od sistema za razumevanje poleg sintakse zahteva še vzročne relacije z zunanjim svetom, priznava, da sintaksa sama po sebi še ni dovolj za mentalno vsebino in tako postavlja pod vprašaj močno inteligenco: »Prvo, kar opazimo pri odgovoru z robotom, je, da molče priznava, da spoznavanje ni zgolj stvar manipulacije s formalnimi simboli, saj dodaja kup vzročnih zvez z zunanjim svetom /.../« (Searle, 1980/1990: 369). Komentatorji menijo (Preston, 2002), da Searle ugovoru z robotom ni naklonjen že zato, ker temelji na redukciji semantike na (vzročne) naravne pojave, ki so brez vsakršne intencionalnosti, kar pa zanj, ki je zagovornik t.i. *internalistične teorije semantike*, ni sprejemljivo, saj naj bi se tako

[22] Za Searlovo možno rešitev iz zagate *glej* Preston, 2002: 30, vprašanje pa ni zgolj, ali bi se z njo strinjal, ampak tudi, ali bi navedeno sploh (upravičeno ali neupravičeno) sprejel kot problem.

izgubila *subjektivnost*, kije bistvo mentalne vsebine.

Razen tega Searle pravi, da »takšne ‘zaznavne’ in ‘motorne’ zmožnosti v ničemer ne povečujejo razumevanja oziroma intencionalnosti prvotnega Schankovega programa« (Searle, 1980/1990: 369–370). Ne glede na to, kako bogato opremimo našega robota, in ne glede na to, kako zahteven Turingov test bo potem lahko opravil, še vedno bo zgolj manipuliral s simboli (noben še tako sofisticiran tehnološki dodatek ga ne bo pripeljal do razumevanja), zaradi česar opravljen Turingov test nikoli ne more biti zagotovilo za to, da nekaj ima duševnost.

Harnad (1993) na to odgovarja, da kakor hitro v igro pripeljemo tako zaznavno in motorno razvitega robota, argument iz »kitajske sobe« pade, saj moramo v ten primeru *klasični Turingov test, v* katerem računalnik zgolj v grobem in iz druge roke opisuje svoje občutke, ko npr. igra tenis (ki ga v resnici seveda ne igra), zamenjati z *absolutnim Turingovim testom*, kjer jih opisuje iz prve roke (zdaj pa tenis resnično igra). To pa vodi Searla v naslednjo (konstruktivno) dilemo (po analogiji je tukaj razumevanje zamenjano z videnjem), ki je kombinacija sistemskega ugovora in ugovora z robotom, z rogovoma, ki zanj nista sprejemljiva:

1. Ali robot (prej Searle) zgolj sprejema iznos optičnega aparata (prej iznos možganske kalkulacije) ali pa robot dejansko gleda, npr. kip Bude.

2. Če robot zgolj sprejema iznos optičnega aparata, potem ne implementira celotnega sistema (ampak le njegov del), torej velja sistemski ugovor. 3. Če robot dejansko gleda kip Bude, potem implementira celoten sistem (ne zgolj njegovega dela), torej resnično vidi.

/. 4. Ali velja sistemski ugovor ali pa robot resnično vidi (Harnad, 1993: 17).

Dejstvo je, da mora Searle takšen modificiran ugovor z robotom zavrniti, saj v nasprotnem primeru njegov argument iz »kitajske sobe« ne deluje več: prvi rog ga sili v ubadanje z že predstavljenim »logičnim odgovorom«, ki mu očita zmoto sestave in zmoto delitve, drugi rog pa po analogiji s takšnim sofisticiranim računalnikom v kombinaciji z robotom pripisuje dejansko razumevanje, kar je hotel ovreči.[23]

Ugovor z možganskim simulatorjem

Zamislimo si računalnik, ki je s pomočjo več programov, ki delujejo vzporedno, sposoben simulirati možgane rojenih Kitajcev, ko obdelujejo formalne elemente kitajskih vnosov in iznosov: na ravni nevronov in sinaps, ki pri reševanju zahtevanih formalnih nalog tvorijo dobro delujočo mrežo, med računalniškim programom in programom možganov rojenih Kitajcev ni nobene razlike. Ugovor z možganskim simulatorjem Searle povzame kot »Ta stroj sprejema kot input kitajske zgodbe in vprašanja o njih, simulira formalno strukturo dejanskih

[23] Za zagovor argumenta iz »kitajske sobe« proti Harnadovi konstruktivni dilemi *glej* Bringsjord in Noel (2002).

kitajskih možganov, ko obdelujejo te zgodbe in oddaja kot output kitajske odgovore. Zamišljamo si lahko celo, da stroj ne deluje samo z enim serijskim programom, temveč s celim nizom programov paralelno, tako kot domnevno delujejo tudi človeški možgani, ko obdelujejo materni jezik« (Searle, 1980/1990: 370). Vprašanje je, ali ne bi morali v tem primeru reči, da računalnik razume, saj bi v nasprotnem primeru morali odreči razumevanje tudi ljudem?

Searlov odgovor je, kajpak nikalen, začne pa ga z naslednjo ironično pripombo: »/.../ mislil sem /.../, da je poglavitna misel UI ta, da nam ni treba vedeti, kako delujejo možgani, če hočemo zvedeti, kako deluje duh /…/ Če bi morali vedeti, kako delujejo možgani, da bi razvili UI, se pač ne bi ukvarjali z UI« (Searle, 1980/1990: 370). Motivacija za takšno razmišljanje se skriva v tem, da je bil v času nastanka argumenta iz »kitajske sobe« prevladujoč pristop na področju umetne inteligence t.i. *simbolna hipoteza*, ki pravi, da naš silni duh ni nič več kot univerzalni sistem simbolov, zaradi česar je tudi človeško mišljenje zgolj manipulacija z njimi.

Kakor hitro pa v igro pripeljemo antropomorfistično zahtevo po poznavanju zgradbe in delovanja človeških možganov za to, da bi stroj mislil, se oddaljimo od ideje močne umetne inteligence, po kateri je *»/.../* duh nasproti možganom toliko, kot je program nasproti hardveru, zaradi česar lahko razumemo duha, ne da bi se poglabljali v nevrofiziologijo« (Searle, 1980/1990: 371).

Ugovor z možganskim simulatorjem pa z vidika konstrukcije mislečega računalnika po njegovem mnenju na neuspeh obsodi tudi takrat manjšinski pristop k umetni inteligenci – *nevronske mreže*, ki trdi, da ima ustrezno sestavljena (iz enot, ki predstavljajo nevrone, in povezav med njimi, ki predstavljajo sinapse) in vrhunsko izurjena konekcionistična mreža duh. Zagovorniki tega t.i. *konekcionizma* predpostavljajo, da ima tako zgrajen sistem pravo duševnost izključno zaradi primerne (našim možganom podobne) konfiguracije in izvežbanosti nevronske mreže, kar pa za izdelavo inteligentnega stroja spet zahteva antropomorfizem in tako spodbija zamisel močne inteligence, ki v duhu računskega funkcionalizma pravi, da *»/.../* obstaja raven mentalnih operacij, ki sestojijo iz računskih procesov nad formalnimi elementi, ki sestavljajo bistvo mentalnega in se jih da uresničiti v najrazličnejših možganskih procesih, prav tako kot lahko vsakteri računalniški program uresničnimo v različnih računalniških hardverih */.../«* (Searle, 1980/1990:371).

Za primer, da bi se konekcionisti iz Searlovega primeža skušali izviti z očitkom, da njihovo pojmovanje močne inteligence tega, kar je navedeno v citatu, ne zahteva, saj pravi zgolj, da »ima ustrezno sestavljena in izurjena nevronska mreža duh izključno zaradi svoje konfiguracije in izvežbanosti« (Preston, 2002: 33), pa Searle (1990) razvije t.i. *miselni eksperiment iz »kitajske telovadnice«*. Zamislite si telovadnico namesto sobe, v kateri imamo veliko samo angleško govorečih ljudi, ki izvajajo iste operacije kot nevroni in sinapse v konekcionistični mreži. Rezultat je na koncu enak kot v sobi, kjer s kitajskimi simboli s pomočjo knjig s

pravili manipulira zgolj en človek. Nihče ne govori kitajsko, čeprav se zunanjim opazovalcem tako zdi in ne obstaja noben način, da bi se celota naučila pomena kitajskih besed (Searle, 1990: 22). Če prej ni razumel en človek, sedaj ne razume več ljudi, in če prej ni razumel sistem z enim procesorjem, ki je svoje operacije izvajal zaporedno, sedaj ne razume sistem z več procesorji, ki svoje operacije izvajajo vzporedno; druge razlike ni.

Copeland se s tem ne strinja in pravi, da je »sistemski ugovor« dovolj za zavrnitev »kitajske telovadnice«: posamezni igralci v mreži sicer res ne razumejo kitajsko, vendar iz tega še ne sledi, da kitajsko ne razume celoten sistem. »Napaka, ki je posledica sklepanja iz dela na celoto, je tukaj celo bolj očitna kot v izvirni različici argumenta iz kitajske sobe« (Copeland, 2002: 116).

Ugovor »drugih duhov«

Vprašajmo se, kako sploh vemo, da npr. naš prijatelj Janez, s katerim redno razpravljamo o aktualnih družbenih, gospodarskih in kulturnih problemih, karkoli od tega, o čemer se z njim pogovarjamo, sploh razume? Zdi se, da razen njegovih primernih in smiselnih verbalnih odgovorov na vprašanja, ki se v danem kontekstu porajajo, nimamo na voljo nobenega drugega dokaza, da to, o čemer teče beseda, razume. Če pa to drži, bi morali tudi računalnik, ki se je v danem okolju sposoben obnašati na ustrezen način, bodisi verbalno bodisi fizično, obravnavati kot bitje ali sistem, ki to, kar je vzrok za njegovo takšno ali drugačno delovanje, razume. V nasprotnem primeru bi nas lahko nekdo pri »merjenju« duševnosti obtožil uporabe različnih vatlov za različne organizme oziroma t.i. šovinizma vrst, ki favorizira ljudi pred stroji, ko za to ni nobene osnove. Searle ugovor drugih duhov povzame takole: »Kako pa veste, da drugi ljudje zares razumejo kitajsko ali karkoli drugega? Samo po njihovem vedenju. Računalnik lahko ravno tako uspešno opravi vedenjske teste kot oni (v načelu); če torej pripisujemo drugim ljudem spoznavnost, jo morate načeloma pripisati tudi računalnikom« (Searle, 1980/1990:373).

Zanimivo je, da z njim opravi na kratko, meni namreč, da »zgreši poanto«, saj pri argumentu iz »kitajske sobe« ne gre za to, *kako* vemo, da imajo druga bitja duh, ampak zato, *kaj* jim pripisujemo, ko pravimo, da ga imajo. »Ost dokazovanja je v tem, da to ne morejo biti zgolj računski procesi in njihov output, saj lahko računski procesi in njihov output obstajajo brez kognitivnega stanja« (Searle, 1980/1990:373).

Zdi se mi, da si ugovor drugih duhov kljub vsemu zasluži več pozornosti, posebej v luči (i) Searlovega ostrega razlikovanje med sintakso in semantiko in (ii) njegove splošne teorije duha. Searle Turingu očita, da se ni nikoli vprašal, *kaj* je tisto, zaradi česar ima neka naprava mentalno vsebino, tj. semantične značilnosti. S tem je zanemaril glavno razliko med strojem in človekom – slednji je po njegovem mnenju pač edini organizem, ki ima dostop tako do sintakse kot tudi do semantik, in hkrati s tem tudi glavni razlog, zakaj mi razumemo ali imamo

intencionalnost, računalniki pa ne. V svojih odgovorih na kritike argumenta iz »kitajske sobe«, kot smo videli, vedno poudarja, da niti celoten sistem niti sofisticiran robot niti možganski simulator nimajo na voljo ničesar, da bi simbolom pripisali pomene, edino, kar delajo je, da s pomeni, ki jih simboli imajo, manipulirajo. Še tako izbrana naprava (od znotraj) simbolom ne more dati pomena, to lahko stori zgolj nekdo od zunaj, človek, ki pa še zdaleč ne rabi biti tako sofisticiran. To, da sintaksa ni zadostna za semantiko, je za Searla zelo *preprosta logična resnica* (Searle, 1992: 200).

Ljudje pa se zavedamo, da v duhu obstaja vsebina, ki nam omogoča vednost o tem, da so naše misli o npr. o predmetih ali simbolih, ki nekaj predstavljajo in imajo pomen in da je to več od zgolj manipulacije s formalnimi simboli. Vprašanje je, zakaj mi to vemo, stroji pa ne? Odgovor na to se skriva v njegovi teoriji duha, ki pravi, da so mentalni pojavi *biološki* pojavi, kar bolj konkretno pomeni, npr., daje vse, kar je zavestno, v resnici biološko, ki jo imenuje *biološki naturalizem*.[24] Natančneje pravi, da so *možgani oboje, vzrok in realizator zavesti*, podobno, kot je voda oboje, vzrok in realizator tekočnosti. Razlikuje med *mikrolastnostmi* sistema, tj. njegovimi sestavnimi deli, kisikom in vodikom ter relacijami med njima pri vodi oziroma nevroni in povezavami med njimi pri možganih, ter *makrolastnostmi* sistema, tj. njegovimi vidnimi lastnostmi, npr. tekočnostjo pri vodi oziroma zavestjo pri možganih. Poanta je, da mikrolastnosti povzročajo makrolastnosti, kisik in vodik ter odnos med njima tekočnost vode oziroma nevroni in interakcije med njimi zavest možganov, pri čemer sistem, voda oziroma možgani, predstavljata okvir, v katerem se to realizira. To vodi do tega, da Searle zavest pripisuje možganom – sistemu (makroraven), ki v emergentističnem slogu vznikne iz nevronov – sestavnih delov (mikroraven), ne pripisuje pa je tudi njim samim (kot je dokazoval, da mora sprejeti zagovornik sistemskega ugovora, če predpostavi, da je sistem zavesten, saj so potem zavestni tudi njegovi deli, npr. jetra, ledvice in želodec, kar pa je, kot pravi, zmotno). Skratka, ljudje se zavedamo mentalne vsebine, ki nam omogoča napredovanje od sintakse do semantike zato, ker imamo ustrezno zgrajene biološke možgane, ki mentalno vsebino hkrati povzročajo in realizirajo, stroji pa ne.

Warwick (2002) dokazuje, da nas argument iz »kitajske sobe« vodi v omenjeni šovinizem vrst, in da pade, če nam uspe pokazati, da ima lahko tudi stroj, računalnik ali robot zavest, pri čemer za izhodišče uporabi Searlovo trditev, da je zavest lastnost možganov. Če to drži, potem jo težko odrečemo umetnim silikonskim možganom ali umetnim nevronskim možganom; smiselno pa je predpostaviti, da gre za različne vrste zavesti, saj je vedno realizirana v različnih vrstah možganov. Neumestno pa je trditi, da se lahko zavest pojavi zgolj v bioloških sistemih, saj je po definiciji tega, kaj je življenje, kmalu jasno, da pogoje zanj, ki segajo od samoorganizacije, neodvisnosti in odgovornosti do reprodukcije, evolucije in metabolizma, izpolnjujejo tudi številni stroji (Warwick,

[24] Za kritiko tega stališča *glej* npr. Penrose (2002).

2002). Zavest, ki jo ima v mislih, je *sposobnost opravljanja takšnih in drugačnih operacij*, v primeru z roboti *Sedem palčkov* je to npr. gibanje naprej brez zadevanja predmetov iz okolice. Po analogiji s človekom, kjer nadzor njegovega telesa s strani človeških možganov pomeni uprimeritev človeške zavesti, v primeru stroja nadzor njegovega telesa s strani strojnih možganov pomeni uprimeritev strojne zavesti. Poanta njegovega odgovora je tako, da so tudi stroji (v okviru svoje vrste) zavestni in sposobni učenja, čeprav nimajo bioloških možganov, ampak umetne, zaradi česar je možno razumevanje (preskok od sintakse do semantike) načeloma pripisati tudi njim.

Kakorkoli, Searle s svojim miselnim eksperimentom iz »kitajske sobe« namiguje na to, da vprašanje, ali stroj lahko misli, ni empirično, ampak miselno, zaradi česar nam pri iskanju odgovora nanj noben tehnološki napredek, pa naj bo še tako velik, ne pomaga. Kljub dosežkom umetne inteligence (včeraj, danes ali jutri) lahko z argumentom iz »kitajske sobe« takoj dokažemo, da inteligentni stroji ne glede na to, kako bodo realizirani, nikoli ne bodo stvarnost. Videli pa smo, da vsaj sistemski ugovor z opozarjanjem na njegovo logično neveljavnost in ugovor drugih duhov v kombinaciji s kritiko biološkega naturalizma, to napoved postavljata pod vprašaj.

FUNCTIONALISM, COMPUTERS, AND »CHINESE ROOM«

The article introduces a relation among functionalism which considers mental states as functional states, computers that are according to its computational version able to think and the Chinese Room Argument which is supposed to prove, once and for all, that machines will never be intelligent. The first part shows the fundamental idea and motivation for functionalism together with the main objections to it, the second part of the article represents the Chinese Room Thought Experiment which, formulated as an argument, rejects the idea of thinking machines on the grounds that a manipulation with symbols is not enough for understanding their contents, and the third part summarizes the key objections to the Chinese Room Argument as well as principle answers to them bringing about thereby some new ideas in order to defend the belief that intelligent computers are, in principle, possible.

Keywords: the Chinese Room Argument, artificial intelligence, the symbol system

hypothesis, the language of thought hypothesis, computation-representation functionalism.

Literatura

Armstrong, D. M. (1999). *The Mind-Body Problem*. Boulder: Westview Press.

Aristotel (1993). *O duši*. Ljubljana: Slovenska matica.

Block, N. (ur.) (1980a). *Readings in Philosophy of Psychology, vol. 1*. Cambridge: Harvard University Press.

Block, N. (1980b/2007). »Uvod: Kaj je funkcionalizem?«. V Markič, O. in Bregant, J. (ur.), *Narava mentalnih pojavov*, Maribor: Aristej, str. 147–166.

Block, N. (1980c). »Troubles with Functionalism«. V Block, N. (ur.), *Readings in Philosophy of Psychology, vol. 1*, Cambridge: Harvard University Press, str. 268–305.

Block, N. (1990). »Inverted Earth«. *Philosophical Perspectives*, 4, str. 53–79.

Block, N. (2002). »Searle's Arguments against Cognitive Science«. V Preston, J. in Bishop, M. (ur.), *Views into the Chinese Room*, str. 70–79.

Block, N. in Fodor, J. (1972/1980). »What psychological states are not«. V Block, N. (ur.), *Readings in Philosophy of Psychology, vol. 1*, Cambridge: Harvard University Press, str. 237–250.

Bregant, J. (2004). *Misel kot vzrok*. Maribor: Pedagoška fakulteta.

Bregant, J. (2010). »Ali lahko stroj misli?«, *Analiza*, 4, str. 55–72.

Bregant, J. (2012). »Turing se je motil«, *Analiza*, 1/2, str. 5–31.

Bringsjord, S. in Noel, R. (2002). »Real Robots and the Missing Thought-Experiment in the Chinese Room Dialectic. V Preston, J. in Bishop, M. (ur.), *Views into the Chinese Room*, str. 144–166.

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

Churchland, P. (2011). *Braintrust*. Princeton: Princeton University Press.

Copeland, J. (1993). *Artificial Intelligence: A Philosophical Introduction*. Oxford: Blackwell.

Copeland, J. (2002). »The Chinese Room from a Logical Point of View«. V Preston, J. in Bishop, M. (ur.), *Views into the Chinese Room*, str. 109–122.

Coulter, J. in Sharrock W. (2002). »The Hinterland of the Chinese Room«. V Preston, J. in Bishop, M. (ur.), *Views into the Chinese Room*, str. 181–200.

Dennett, D. C. (1985). »Can Machines Think«. V Shafto, M. (ur.), *How we Know*, San Francisco: Harper & Row, str. 121–145.

Dennett, D. C. (1991/2012). *Pojasnjena zavest*. Ljubljana: Krtina.

Descartes, R. (1637/2007). *Razprava o metodi*. Ljubljana: ZRC SAZU.

Fodor, J. (1975). *The Language of Thought*. Cambridge: Harvard University Press.

Fodor, J. (1981). »The Mind-Body Problem«, *Scientific American*, 244, str. 124–132.

Fodor, J. (1987). *Psychosemantics*. Cambridge: The MIT Press.

Fodor, J. A. in Pylyshyn, Z. W. (1988). »Connectionism and Cognitive Architecture: A Critical Analysis«, *Cognition*, 28, str. 3–71.

Gendron, B. (1971). »On the Relation of Neurological and Psychological Theories. A Critique of the Hardware Thesis«. V Buck, R. C. in Cohen, R. S. (ur.), *In Memory of Rudolf Carnap*, Dordrecht: Reidel, str. 483–495.

Harnad, S. (1993). »Grounding Symbols in the Analog World with Neural Nets: A Hybrid Model«, *Think*, 2, str. 12–20.

Harnad, S. (1995). »Why and How We Are Not Zombies«. *Journal of Consciousness Studies*, 1, str. 164–167.

Harnad, S. (2002). »Minds, Machines, and Searle2: What's Right and Wrong aboout the Chinese Room Argument«. V Preston, J. in Bishop, M. (ur.), *Views into the Chinese Room*, str. 294–307.

Haugeland, J. (2002). »Syntax, Semantics, Physics«. V Preston, J. in Bishop, M. (ur.), *Views into the Chinese Room*, str. 379–392.

Jackson, F. (1995). »What Mary Didn't Know«. V Moser P. K. in Trout, J. D. (ur.), *Contemporary Materialism*, London: Routledge, str. 180–189.

Kalke, W. (1969). »What is Wrong with Fodor and Putnam's Functionalism«. *Nous*, 3, str. 83–94.

Locke, J. (1689/2012). *An Essay Concerning Human Understanding*. Verzija je dosegljiva na www.earlymoderntexts.com.

Kurzweil, R. (1998). *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*. New York: Viking.

Markič, O. (2010). *Kognitivna znanost*. Maribor: Aristej.

Miščević, N. (2001). »Uvod«. V Miščević, N. in Smokrović, N. (ur.), *Računala, mozak in ljudski um*, Rijeka: Izdavački centar Rijeka, str. 9–42.

Newell, A. (1980). »Physical Symbol Systems«. *Cognitive Science*, 4, str. 135–183.

McCulloch, W. in Pitts, W. (1943). »A Logical Calculs of the Ideas Immanent in Nervous Activity«, *Bulletin of Mathematical Biophysics*, 5, str. 115–133.

Preston, J. (2002). »Introduction«. V Preston, J. in Bishop, M. (ur.), *Views into the Chinese Room*, str. 1–50.

Putnam, H. (1960). »Minds and Machines«. V Hook, S. (ur.), *Dimensions of Mind*, New York: New York University Press, str. 148–179.

Putnam, H. (1966/1975). »The mental life of some machines«. V *Mind, Language and Reality*, Cambridge: Cambridge Unviersity Press, str. 408–428.

Putnam, H. (1963/1980a). »Brains and Behaviour«. V Block, N. (ur.), *Readings in Philosophy of Psychology, vol. 1*, Cambridge: Harvard University Press, str. 24–36.

Rey, G. (2002). »Searle's Misunderstanding of Functionalism and Strong AI«. V Preston, J. in Bishop, M. (ur.), *Views into the Chinese Room*, str. 201–225.

Searle, J. (1980/1990). »Duhovi, možgani in programi«. V Hofstadter, D. R. in Dennet, D. C. (ur.), *Oko duha*, Ljubljana: Mladinska knjiga, str. 361–379.

Searle, J. (1990). »Is the Brain's Mind a Computer Program?«, *Scientific American*, 262, str. 20–25.

Searle, J. (1992). *The Rediscovery of the Mind*. Cambridge. The MIT Press.

Searle, J. (1993). »The Failures of Computationalism«. *Think*, 2, str. 68–71.

Searle, J. (1997). *The Mystery of Consciousness*. New York: A New York Review Book.

Sesardić, N. (1984). *Fizikalizam*. Novi Beograd: Istraživačko-izdavački centar SSO Srbije.

Shoemaker, S. (1975). »Functionalism and Qualia«. *Philosophical Studies*, 27, str. 291–315.

Shoemaker, S. (1982). »The Inverted Spectrum«. *Journal of Philosophy*, 79, str. 357–381

Smart, J. J. C. (1959/2007). »Občutki in možganski procesi«. V Markič, O. in Bregant, J. (ur.), *Narava mentalnih pojavov*, Maribor: Aristej, str. 79–93.

Turing, A. (1936). »On Computable Numbers, with an Application to the *Entscheidungsproblem*«, *Proceedings of the London Mathematical Society*, 42, str. 230–265.

Turing, A. (1990). »Stroji, ki računajo, in inteligenca«. V Hofstadter, D. R. in Dennet, D. C. (ur.), *Oko duha*, Ljubljana: Mladinska knjiga, str. 61–74.

Tye, M. (1995). *Ten Problems of Consciousness*. Cambridge: The MIT Press.

Tye, M. (2006). »Absent Qualia and the Mind-Body Problem«. *Philosophical*

Review, 115, str. 139–168.

Warwick, K. (2002). »Alien Encounters«. V Preston, J. in Bishop, M. (ur.), *Views into the Chinese Room*, str. 308–318.