

Janez Bregant

Filozofska fakulteta Univerze v Mariboru

Možgani v primežu računalnikov

Članek s pomočjo enostavne dileme zagovarja prepričanje, da smo ljudje oziroma naši možgani v nekem določenem smislu računalnik. Copeland (1993) trdi, da empirične razlike med računalnikom in možgani ne uspejo zavrniti klasičnega pristopa k enačenju možganov in računalnika, t.i. močne simbolne hipoteze, ki zagovarja njuno identifikacijo na osnovi simbolnega predstavljanja znanja o svetu, saj te dokazujejo zgolj to, da možgani niso tradicionalni, von Neumannovski računalnik. Tudi konekcionistični pristop k pojasnjevanju naše kognicije se tukaj ne odreže nič bolje, saj kljub temu, da stavi na nesimbolno predstavljanje informacij o svetu, ki naj bi bilo računalnikom tuje, njihovega enačenja z možgani ne uspe preprečiti. Še vedno smo lahko računalniki, a neke druge vrste, ki ni zgolj načeloma možna, kar dokazuje najbolj razvit hibridni model naše spoznavne ACT-R, ki združuje oba pristopa. In celo če zavoljo argumenta sprejmemo prepričanje, da računalnik kot manipulator s simboli ni združljiv z nevronskimi mrežami, nas to še vedno ne obvaruje pred tem, da ne bi bili računalniki v nekem splošnem smislu.

Ključne besede: močna simbolna hipoteza, konekcionistične mreže, ACT-R, dinamični sistemi, jezik misli

0 Uvod

V 18. stol. so mnogi ljudje menili, da nismo nič drugega kot utelešeni duhovi, tj. duhovi, ki jim telo rabi zgolj kot sredstvo za pridobivanje inteligence, znanja in izkušenj, naši najbližji nečloveški sorodniki pa naj bi bili angeli.

V 19. stol. se je z Darwinom, Freudom in Marxom to spremenilo do te mere, da nismo točno vedeli, kaj smo. Z Darwinovo izdajo prelomne študije *O nastanku vrst* leta 1859, se je zdelo, da smo najbolj razvite živali, ki so do te stopnje prišle tekom evolucije, botroval pa ji je naravni izbor. Da je naš najbližji sorodnik človeku podobna opica, je bila za tisti čas seveda bogokletna trditev, o posledicah katere ni treba izgubljeni besed, segale pa so vse od cerkvenega izobčenja do prikazovanja Darwina kot opice. Pri Freudu se zdi, da nas določata spolni gon – libido in težnja k smrti oziroma samouničenju. Vzroki zanju naj bi se skrivali v naši podzavesti, potlačene misli in občutki pa privrejo na dan, ko smo podvrženi hipnozi. Z Marxovim raziskovanjem ekonomije in njenim povezovanjem s filozofijo pa se je zdelo, da smo subjekti in objekti hkrati, vendar šele potem, ko presežemo

njuno odtujitev. »Dejstvo, da je človek *telesno*, dejansko, čutno, objektivno bitje z naravnimi zmožnostmi, pomeni, da ima za svojo naravo *dejanske, čutne predmete* kot predmete svojega življenjskega izražanja, ali da lahko svoje življenje izrazi zgolj v dejanskih čutnih predmetih« (Marx, 1885/1997: 325).

V 20. stol. pa je bilo že veliko ljudi prepričanih, da smo računalniki, naši najbližji sorodniki pa naj bi bili proizvodi HP-ja, Appli ali Toshibe. Nekateri so se za to možnost glasno izrekli, drugi je niso želeli obešati na veliki zvon, še več pa se jih je s to idejo v obliki fraz, kot je npr. »Vsi smo sprogramirani s strani družbe«, spogledovalo.

Namen tega članka je pokazati, da smo ljudje, čeprav obstajajo empirični in teoretični razlogi za zavrnitev ideje, da so naši možgani računalnik v klasičnem von Neumannovem smislu, kljub vsemu računalnik. Osnovna ideja, ki se za tem skriva, je, da smo v vsakem primeru fizični stroji, vprašanje je zgolj kakšne vrste (Copeland, 1993: 180). Mogoče je to presenetljivo, ampak o tem, da smo ljudje stroji, je pisal že La Mettrie v svojem delu *L'homme machine (Strojni človek)* leta 1748: »Človek je stroj, zgrajen na takšen način, da si je to nemogoče predstavljati in ga zato tudi nemogoče definirati«¹ (La Mettrie: 1748/1996: 5). Zgodovinsko gledano je bilo vedno tako, da so možgane zaradi njihove kompleksnosti enačili s trenutno tehnološko najrazvitejšo napravo, v enem trenutku celo s telegrafom in telefonom. Vendar dejstvo, da znanost o možganih rada sledi »modnim zapovedim« in jih danes skuša razumeti in predstaviti kot računalnike, še ne spodkoplje trditve, da so možgani računalnik, pa čeprav se je v zgodovini analogija med tehnološkimi izumi in delovanjem možganom vedno izkazala za zgrešeno.

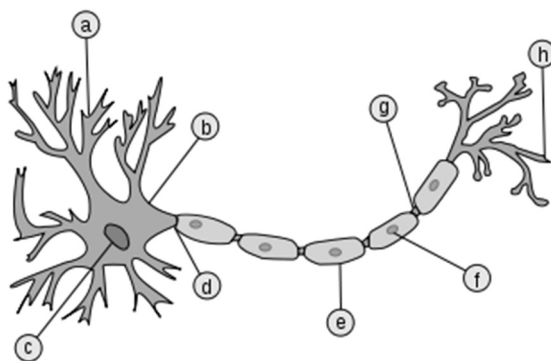
V prvem delu je predstavljen klasični pristop k enačenju možganov in računalnika, t.i. močna simbolna hipoteza, ki zagovarja njuno identifikacijo na osnovi simbolnega predstavljanja znanja o svetu, v drugem pa empirične razlike med računalnikom in možgani, ki naj bi pokazale, da pri tem ni uspešna. V tretjem delu je predstavljen konekcioniistični pristop k pojasnjevanju naše kognicije, ki temelji na nesimbolnem predstavljanju informacij o svetu in tako preprečuje enačenje možganov in računalnika, v petem pa alternativna modela naše kognicije, hibridni, ki združuje najboljše iz simbolnega in konekcioniističnega sveta, ter dinamični, ki poudarja vpetost človeka v okolje in spremembe skozi čas. Peti del predstavlja sklepno dilemo, ki nas zaradi svoje enostavne zgradbe pouči, da smo v vsakem primeru računalniki, res pa je, da zelo splošne sorte.

¹ Pri tem ne pozabi omeniti, da je o tem pisal že Descartes v *Razpravi o metodi* leta 1637: »Nihče ne zanika, da je ta slavn filozof naredil mnogo napak, je pa razumel živalsko naravo in bil prvi, ki je dovršeno pokazal, da so živali stroji« (La Mettrie, 1748/1996: 35).

1 Simbolna hipoteza

Zgradba nevrona

Možgani so sestavljeni iz dveh različnih skupin celic, pri čemer so za razumevanje njihovega delovanja pomembnejši in zanimivejši *nevroni*, katerih funkcija je prenos informacij preko električnih in kemičnih impulzov (manj pomembne in manj zanimive, vsaj kar se tiče možganskega delovanja, so *nevroglijske* celice, katerih funkcija je zgolj oskrba nevronov s hranili in njihova izolacija). Spodnja slika prikazuje (unipolarni) nevron:



Slika 1: Zgradba unipolarnega nevrona

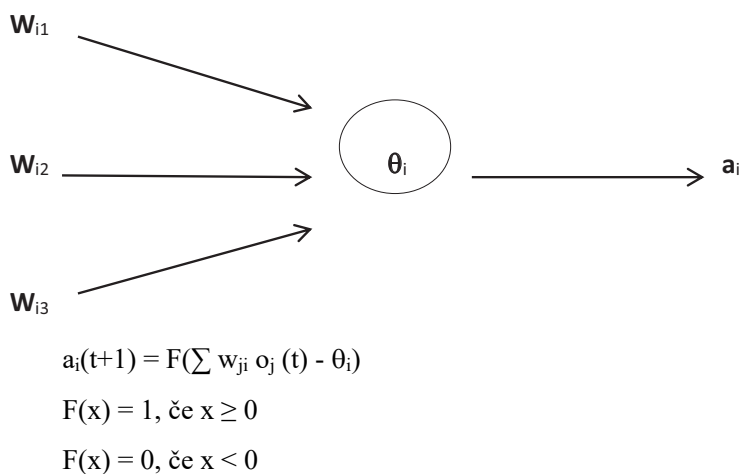
Nevron je sestavljen iz dendritov (a), telesa ali *soma* (b), jedra (c), nevrita ali *aksona* (d), mielinske ovojnice (e), Shwannove celice (f), Ranvierovega zažetka (g) in živčnih končičev (h).

Nevroni so med seboj povezani tako, da se dendriti enega držijo živčnih končičev drugega, zaradi česar tvorijo nepredstavljivo kompleksno celoto oziroma mrežo, ki obsega nekje tisoč dendritov na nevron in 10^{12} nevronov na možgane.

Že La Mettrie je predpostavljal naslednje: »Posebna lastnost našega stroja je, da vsako njegovo vlakno, tudi najmanjše, oscilira. To naravno nihanje je kot ura, ki včasih zamre in se mora potem obnoviti. Če postane šibko, se mora okrepiti, če pa premočno, oslabiliti.« (La Mettrie, 1748/1996: 31) Kljub temu so šele leta 1843 odkrili, da je delovanje živčnega sistema v osnovi sprejemanje in prevajanje električnih impulzov. V somi nevrona se sprosti električna energija, ki potuje po nevritu do sosednjih z njim povezanih nevronov. Do sprostitve električne energije pride v telesu nevrona (celici) takrat, ko je vzdraženih dovolj njegovih dendritov oziroma ko je dosežen ali presežen prag. Takrat »zagori« cel nevron in preko svojih živčnih končičev samodejno vzdraži vse dendrite sosednjih nevronov, s katerimi je povezan. Delovanje nevronov je v nekem smislu krožno: nevron pošilja signale svojim sosedom, da bi ti pošiljali signale svojim sosedom, da bi ti pošiljali signale svojim sosedom itd. Tako nam delovanje nevronov, ker je neinformativ-

no, saj sploh ne vemo, kaj je namen njihovega vzajemnega proženja, ne pove še ničesar o delovanju možganov. No, vsaj tako je bilo v letih med 1840 in 1940, kljub temu da so bile vmes do podrobnosti raziskane kemične značilnosti nevronov in njihov električni potencial.

Potem pa sta leta 1943 McCulloch (nevrofiziolog) in Pitts (matematik) objavila prelomni članek z naslovom »A Logical Calculus of the Ideas Immanent in Nervous Activity«, ki je korenito razburkal in spremenil področje raziskovanja možganov, saj je pokazal, »[k]ako lahko operacije nevronov in njihove povezave z ostalimi nevroni modeliramo s pojmi logike« (Markič, 2010: 30). Dokazovala sta, da je edino, kar je pri nevronih pomembno to, da so sproženi (»prižgani«) ali nesproženi (»neprižgani«), pri čemer lahko prvo označimo z npr. 1 ali »da«, drugo pa z npr. 0 ali »ne«, to, ali so sproženi ali nesproženi, pa je odvisno od tega, ali je dosežen ali presežen njihov aktivacijski prag. Slika njunega idealiziranega nevrone kot osnove za modeliranje operacij pravih nevronov je naslednja:



Slika 2: Delovanje idealiziranega nevrone

Gre za binarno enoto, ki izračuna vsoto vhodnih podatkov iz drugih binarnih enot – $F(\sum w_{ji} o_j(t) - \theta_i)$ – in kot izhodni podatek odda a_i , tj., 1 ali 0, odvisno od tega, ali vsota presega vrednost θ_i , ki predstavlja aktivacijski prag. Utež w_{ji} je moč sinapse, ki povezuje nevron i z nevronom j , lahko pa so vzbujevalne ali zaviralne ali pa so enake nič, kar pomeni, da med nevronoma ni stika. (Markič, 2010: 30)

Tako ni nevron nič drugega kot naprava, ki lahko fizično realizira eno izmed obeh stanj. Vse ostalo, kot je npr. spreminjanje kemične zgradbe nevrone ali intenzivnost električne eksplozije v somi, naj ne bi bilo pomembno. Z drugimi besedami, skupine nevronov delujejo kot neke vrste preprosti manipulatorji s simboli, med seboj pa so povezani tako, da ko se sproži npr. prva dva, se sproži tudi tretji, oziroma ko se pri prvih dveh pojavi znak, da je vključen (1), se pojavi znak, da je

vklučen (1) tudi pri tretjem. »Zakon 'vse ali nič', kateremu je podvržena živčna aktivnost, je zadosten za to, da je lahko aktivnost kateregakoli nevrona predstavljena kot propozicija« (McCulloch, Pitts, 1943/1990: 23–24). To pomeni, da si lahko nevrone zamislimo kot propozicije, ki so resnične (1) ali neresnične (0), obtežene povezave med njimi pa kot logične veznike in tako v pojmihi logike izračunamo vse, kar lahko jezikovno opišemo.

Zanimivo je, da tudi računalnik deluje po istem principu preklapljanja med 1 in 0 kot nevron (deli, ki mu to omogočajo, se imenujejo »logična vrata«). Recimo, če v računalnik preko tipkovnice vnesem »mačka« (vhod ali vnos), računalnik preklaplja med 0 in 1, odvisno od kode, ki jo vsaka črka ima, dokler ne pride do binarnega zapisa, npr. 1000011 1000001 1000111 1001111 1000001, ki pomeni izpis »mačka« na mojem ekranu (izhod ali iznos). Tako tudi računalnik ni nič drugega kot naprava, ki lahko fizično realizira eno od obeh stanj. Ker je tako tudi sam podobno kot nevron manipulator s simboli, so ljudje, ko so se prvi računalniki pojavili, te povezali z modelom nevrona, ki sta ga razvila McCulloch in Pitts, ter prišli do sklepa, da so možgani (in s tem tudi ljudje) računalniki. To, kako smo pri enačenju možganov z napravami napredovali, pa lepo ponazarja naslednji odlomek iz sredine 18. stoletja: »Ne motim se; človeško telo je ura /.../ ogromna in inteligentno zgrajena /.../, njen izdelovalec pa je njen nov prebavni sok.« (La Mettrie, 1748/1996: 34, 31)

Močna in šibka simbolna hipoteza

V splošnem simbolna hipoteza pravi, da je mišljenje računanje, ki vključuje manipulacijo s simboli, ta pa je lahko realizirana v sistemih, ki jo izvajajo s preklapljanjem med 0 in 1 kot elementoma binarnega sistema. Če mislimo, da je ta ideja nova, se motimo, »Mišljenje ni nič drugega kot računanje« pravi že Hobbes leta 1651 v *Leviathanu* (Hobbes, 1651/2006: 32). Njegove besede, kot sem dejal že drugje,² vsebujejo dve domnevi: (i) mišljenje je mentalni proces, ki obsega manipulacijo s simboli, ki se dogaja v možganih, zaradi česar so primerki misli v bistvu primerki možganov, ali, kot jih imenuje Hobbes, »fantazme«. »Ko človek razmišlja, ne dela ničesar drugega kot, da si zamišlja končno vsoto, ki jo dobi s seštevanjem delov, ali ostanek, ki ga dobi z odštevanjem ene vsote od druge. /.../ Te operacije se ne dogajajo samo pri številih, ampak pri vseh načinih stvari, ki jih lahko združujemo ali odvzemamo« (Hobbes, 1651/2006: 32). In (ii) mišljenje je najbolj čisto, ko zvesto sledi metodičnim pravilom, zaradi česar lahko rečemo, da je mišljenje mehanski proces, je nekaj takega kot upravljanje mentalnega abaka: fantazme premetavamo sem in tja skladno s pravili razuma, podobno kot na pravem abaku skladno z računskimi pravili sem in tja premikamo kroglice (Haugeland, 1986: 23).

² Glej Bregant (2010).

Vendar danes znotraj simbolne hipoteze ločimo dve smeri, prva, imenujmo jo *šibka simbolna hipoteza*, je predpostavka, da *je načeloma možno zgraditi univerzalni simbolni sistem (računalnik), ki misli*. Recept, kateremu moramo slediti, vsebuje pet točk: (i) uporabo primerne kode za predstavitev stvarnih predmetov; (ii) kreacijo spomina, v katerem je shranjeno znanje o svetu; (iii) izdelavo ustreznih vhodnih naprav, ki pretvorijo dražljaje iz okolja v simbole; (iv) razvoj naprav, ki omogočajo uporabo osnovnih operacij na proizvedenih in shranjenih simbolnih strukturah, in (v) proizvodnjo izhoda v obliki simbolne predstave ustreznega (tudi govornega) obnašanja (Copeland, 1993: 80).

Druga, imenujemo jo, *močna simbolna hipoteza*, pa je predpostavka, da *so zgolj univerzalni simbolni sistemi (računalniki) zmožni misliti*. Krila ji daje ideja o *večvrstni realizaciji simbolov*, ki vsebuje domnevo, da so lahko simboli oziroma njihovi tipi realizirani na različne načine: z barvo, ogljem, svinčnikom, elektromagnetnimi stikali (releji), elektronkami, silikonskimi polprevodniki (tranzistorji) itd. Že Turing je menil, da večvrstna realizacija simbolov dokazuje, da računalnik v svojem bistvu ni električna naprava:³ »Pogosto se misli, da je pomembno, da so sodobni digitalni računalniki električni in da je tudi živčni sistem električen. /.../ ker so vsi digitalni računalniki v nekem smislu enakovredni, vidimo, da uporaba elektrike z vidika teorije ni pomembna« (Turing, 1950: 439).

Od razumevanja računalnikov na ta način pa ni več daleč do njihove izenačitve z možgani. Skladno z močno simbolno hipotezo, namreč noben umetni sistem, ki ni računalnik, ne more misliti. Še več, vsako naravno bitje, kateremu pripisujemo sposobnost mišljenja, se po natančni preiskavi izkaže za računalnik. Res je, da so pri računalnikih in naših možganih manipulacije s simboli verjetno realizirane z bistveno drugačnimi fizičnimi operacijami, vendar to, kot smo videli, pri tem, kaj smo, nima nobene vloge. Tako ravno močna simbolna hipoteza pooseblja idejo, ki se je pojavila, ko so nekateri avtorji povezali McCullohovo in Pittsovo model nevronskega računalnika, da smo ljudje (oziroma naši možgani) računalniki. Vnet zagovornik močne simbolne hipoteze je Pylyshyn: »Semantična vsebina [prepričanj in namer] je kodirana s strani možganskih lastnosti na isti splošni način, na katerega so semantične vsebine računalniških predstav kodirane s strani fizično uprimerjanih simbolnih struktur« (Pylyshyn, 1984: 258). Meni, da so naše mentalne operacije v bistvu manipulacije s simbolnimi reprezentacijami, ki so zgolj podobne običajnim stavkom kateregakoli govorjenega jezika: »/.../ [simbolne reprezentacije] so simbolni izrazi v notranjem, fizično uprimerjanem simbolnem sistemu, ki se včasih imenuje 'mentalese' ali 'jezik misli'«⁴ (Pylyshyn, 1984: 194). Poudariti velja, da kljub enačenju možganov in računalnika, med njimi skorajda ne

³ Analitični stroj Charlesa Babbaga, razvit med leti 1823 in 1833, ki je bil prvi programiran računalnik, je bil denimo »narejen« (v resnici ni nikoli zaživel, ker bi bila za tisti čas njegova izdelava tehnološka prezapletena) izključno iz mehanskih delov: koles in zobnikov.

⁴ Za jezik misli glej tudi Fodor (1975), Markič (2010), Bregant (2013).

moremo pričakovati drugih podobnosti razen manipulacije s simboli, pa še ta lahko poteka na povsem različne načine, in jih močna simbolna hipoteza tudi ne predpostavlja, kot bomo videli v nadaljevanju.

Kakorkoli, z vidika identifikacije možganov z računalnikom imamo na voljo tri možnosti: (i) da jo vzamemo dobesedno, in sicer v smislu, da smo zgolj manipulatorji s simboli (kot v primeru močne simbolne hipoteze), (ii) da jo vzamemo kot koristno metaforo ali provokacijo in (iii) da jo vzamemo kot resnično v smislu, da sicer smo (neke vrste) računalniki, nismo pa manipulatorji s simboli.

2 Možgani in računalnik: empirične razlike

Vprašanje je, zakaj že nekateri pionirji umetne inteligence niso verjeli tega, da je človek računalnik, npr. von Neumann (1958), odgovor na to pa je sedaj verjetno še bolj na dlani kot v 50. in 60. letih 20. stoletja. Danes vemo o možganih več kot kdajkoli prej, ta nevrofiziološki napredek pa postavlja identifikacijo človeka z računalnikom, vsaj v smislu močne simbolne hipoteze, po kateri smo oboji zgolj manipulatorji s simboli, pod velik vprašaj. Empirične razlike med možgani in računalniki lahko strnemo v sedem točk:⁵

(i) možgani so v primerjavi z računalnikom katastrofalno počasni: čas, ki poteče od vzdraženja dendritov do sprostitve elektrike v njegovih živčnih končičih (prehod iz stanja 0 v stanje 1), je tisočinka sekunde, kar je milijonkrat dlje od časa, ki ga porabi računalnik, da v logičnih vratih preklopi med 0 in 1;

(ii) nevroni so preko dendritov in živčnih končičev povezani z ogromnim številom drugih nevronov: tudi do 100.000 jih vpliva na to ali »gorijo« ali ne »gorijo«, kar močno prekaša povezave med deli računalnika, ki so odgovorni za preklapljanje med 0 in 1, preštejemo jih lahko na prste ene roke;

(iii) možgani so v nekem smislu odporni na poškodbe: njihova zmogljivost upada postopoma, tj. glede na količino možganskega tkiva, ki ga izgubimo; tudi huda poškodba, npr., ne pomeni nujno tega, da možgani ne zmorejo več opravljati svojih bistvenih funkcij, medtem ko računalnik povsem odpove že ob najmanjši fizični poškodbi elektronskih delov procesorja;

(iv) v možganih obstaja več različnih nevronov: zdi se, da je njihova oblika, npr. piramidni, zvezdnati, vretenasti itd., odvisna od njihove funkcije, pa tudi določene oblike se vedno pojavljajo na istih mestih, tako da tega ne moremo označiti kot nekaj nepomembnega, medtem ko je računalniški procesor sestavljen iz omejenega števila istih elementov;

⁵ Glej npr. Churchland (1986), Copeland (1993).

(v) možgane obravnavamo kot paralelne (vzporedne) procesorje, tj. takšne, ki hkrati izvajajo različne operacije, računalnike pa kot sekvenčne (zaporedne) procesorje, tj. takšne, ki izvajajo operacije eno za drugo, kar pomeni, da možgani zaradi velikega števila med seboj povezanih nevronov v tisočinki sekunde izvedejo toliko različnih operacij, kot bi jih računalnik kljub svoji hitrosti v nekaj letih ali stoletjih;

(vi) človeški spomin deluje po načelu dostopa preko vsebine, tj. glede na vsebino/opis vprašanja se bomo spomnili odgovora nanj, računalniški spomin pa po načelu dostopa preko naslova (pri tem je razmerje med številko in vsebino podobno razmerju med hišno številko in njeno notranjo opremo): glede na številko naslova v spominu, se bo računalnik »spomnil« odgovora na vprašanje;

(vii) možgani ne shranjujejo vsakega podatka na svojem mestu, ampak je to porazdeljeno med določeno število nevronov (domneva se, da pri shranjevanju posameznega podatka sodeluje od 700.000 do 7 milijonov nevronov), zaradi česar npr. v poskusih podgane, ki jim, potem ko so se naučile pot skozi labirint, odvzamemo del možganov (do 10% in kjerkoli), pot do cilja kljub vsemu najdejo,⁶ medtem ko se v računalniku vsak drobec informacij nahaja na točno določenem in njemu lastnem mestu.

Na kaj lahko iz navedenih razlik sklepamo? Churchland meni, da iz tega preprosto sledi, da ljudje nismo računalnik: »/.../ te razlike povzročijo, da metafora, ki identificira možgane z računalnikom, izpuhti« (Churchland, 1986: 459). Copeland (1993) se z njo ne strinja in trdi, da je največ, kar lahko iz tega izpeljemo, da možgani niso to, kar se razume pod pojmom von Neumannovski računalnik. Seznam lastnosti, ki jih je von Neumann predpisal za to, da bi naprava štela kot računalnik, se namreč povsem sklada z opisanimi razlikami med njim in možgani. Posplošeno so zahtevane lastnosti štiri: (i) naprava je zaporedni procesor: ukaze izpolnjuje enega za drugim, (ii) vsak niz simbolov v njej je shranjen na svojem posebnem mestu v spominu, (iii) dostop do shranjenih podatkov poteka preko njihovih oštevilčenih naslovov in (iv) nadzor se izvaja preko zgolj ene, t.i. centralno procesne enote (CPE) (Copeland, 1993: 193). Res je, da je njegov seznam še danes osnova vsakega računalnika, vendar gre kljub temu zgolj za en način, kako lahko izdelamo univerzalni simbolni sistem, iz česar pa še ne sledi, da bi takšni morali biti možgani, pa čeprav bi se izkazalo, da so resnično zgolj manipulatorji s simboli. In močna simbolna hipoteza načina zgradbe univerzalnega simbolnega sistema, ki je edini sposoben misliti, niti ne določa: zaporednost in vzporednost, način shranjevanja in dostopa do podatkov, izvedba kontrole nad izvajanjem operacij itd. na to, ali je naprava univerzalni manipulator s simboli, nimajo nobenega vpliva in tako močne simbolne hipoteze niti ne zavrnejo niti ne potrdijo.

⁶ Glej Lashley (1960).

Vprašanje, na katerega moramo odgovoriti, da bi prišli do tega, ali je močna simbolna hipoteza resnična ali ne, je, ali je »mentalese«, ki naredi možgane za manipulator s simboli, resnično edini način, s katerim lahko razložimo našo kognicijo. Tipično se ji pripisujejo štiri lastnosti, *produktivnost*, *sistematičnost*, *kompozicionalnost* in *sklepalna koherentnost*. Produktivnost misli pomeni, da lahko iz končnega števila propozicij tvorimo neomejeno število misli, npr. vedno bi lahko končnemu naboru vseh misli, ki jih proizvedemo, dodali še eno in tako dalje v neskončnost (če nas ne bi prej prehitela smrt). Sistematičnost misli pomeni, da če razumemo stavek, ki ima določeno obliko (je zgrajen iz danih elementov), potem razumemo tudi stavek, ki ima podobno obliko (je zgrajen iz istih, a drugače povezanih elementov), npr., če razumemo stavek »Peter ljubi Mojco«, razumemo tudi stavek »Mojca ljubi Petra«. Kompozicionalnost misli pomeni, da sintaksa, tj. elementi stavkov in slovnična pravila, določa pomen, ker so, npr., besede »Peter«, »ljubiti« in »Mojca« neodvisne od konteksta, lahko iz tega, kako so povezane, brez da bi morali vedeti karkoli drugega, takoj izpeljemo pomen danega stavka, v katerem nastopajo. Sklepalna koherentnost ali sistematičnost sklepanja pa pomeni, da če lahko izvajamo sklepanje, ki je določene oblike, potem lahko izvajamo tudi sklepanje, ki je podobne oblike, npr. iz »Peter ljubi Mojco, stanuje v Mariboru, dela na Filozofski fakulteti in rad smuča«, lahko izpeljemo »Peter ljubi Mojco«, to pa lahko izpeljemo npr. tudi iz »Peter ljubi Mojco in stanuje v Mariboru«⁷ (Fodor, Phylyshyn, 1988: 33–50).

Fodor in Pylyshyn (1988) trdita, da je pojasnitev opisanih štirih lastnosti kognicije, ki jo lahko imenujemo *razlagalna naloga* (Markič, 2010), možna zgolj v primeru, ko možgane razumemo kot manipulator s simboli, ki temelji na fizičnem simbolnem sistemu, imenovanem jezik misli, zaradi česar je močna simbolna hipoteza resnična. Takšen *simbolni model* naše duševnosti, ki ga Beer opredeli kot »/.../program, katerega vhod je simbolni opis problema, ki ga mora rešiti, kar stori tako, da preko svojega (simbolno predstavljenega) znanja z vhodnim simbolnim opisom tako dolgo manipulira (pri čemer se drži izključno sintaktičnih pravil), dokler ne pride do izhodne rešitve« (Beer, 2010: 96), in ki ga imenujemo tudi *kognitivizem*, je, kot sta se slikovito izrazila, »edina igra v mestu«. Zanimivo je, da je resničnost njune izjave postala vprašljiva tako rekoč takoj zatem, ko sta jo izrekla, saj so se že v sredini 80. let 20. stoletja pojavili prvi *konekcionistični modeli kognicije*, ki so svoj navdih prav tako našli v že omenjenem prelomnem članku McCullocha in Pittsa. Tako se naenkrat pojavi nov igralec, ki je prav tako sposoben opraviti razlagalno nalogo in to brez sklicevanja na jezik misli kot simbolni sistem, kar postavlja resničnost močne simbolne hipoteze pod velik vprašaj in s tem spet oživlja idejo, da možgani kljub vsemu niso računalniki oziroma vsaj ne

⁷ Za povzetek njune razlage štirih lastnosti kognicije in možnosti njihove realizacije v računalnikih glej Markič, 2010: 59–65.

takšni, kot se je mislilo, tj. manipulatorji s simboli. Če se lahko dostojno zoperstavi simbolnemu modelu kognicije, pa bomo videli v nadaljevanju.

3 Konekcionistični model

Analogija med računalnikom in možgani temelji na kognitivistični predpostavki, da poteka pojasnjevanje obnašanja stroja in človeka na treh istih ravneh, ki jih je prvi razdelal Marr (1982), danes pa so znane pod imeni (i) semantična (vsebinska), (ii) sintaktična (algoritemska) in (iii) strukturna (fizična). Prva se ukvarja s tem, kaj je cilj računanja z mentalnimi predstavami, druga s tem, kako to računanje poteka, tj. z odkrivanjem algoritmov, s katerimi se iz vhodnih podatkov dobijo izhodni, tretja pa s tem, kako to računanje fizično realizirati. Videli smo, da močna simbolna hipoteza poudarja prvi dve ravni, tj. semantično in sintaktično, zanemarja pa tretjo, tj. strukturno raven, s čimer se spogleduje s teorijo duha, znano kot *funkcionalizem*.⁸ Ta poudarja, da sta za pojasnitev našega obnašanja, t.i. psihološko razlago, pomembni zgolj prvi dve ravni, ne pa tudi tretja, saj razume mentalno kot »[m]nožico duševnih reprezentacij, nad katerimi potekajo sintaktično določene transformacije« (Markič, 2010: 39). Ni čudno torej, da sta močna simbolna hipoteza in funkcionalizem zaveznika, kar se ne nazadnje kaže tudi v tem, da so zagovorniki simbolnega modela kognicije ravno filozofi, ki prisegajo na formulacijo duha v pojmih njegovih tipičnih vzrokov in učinkov, tj. njegove funkcije.

Posledica prepričanja, da takšno zanemarjanje tretje, tj. strukturne, ravni pri pojasnjevanju mentalnih pojavov ni na mestu, pa je bil vzpon novega modela naše kognicije, ki na razlagalno hipotezo ne odgovarja v pojmih manipulacije s simboli, pa tudi načina, kako je obdelava podatkov izvedena, povsem ne zanemarja. Njegov vzpon se je začel v sredini 80. let 20. stoletja, ko so simbolni modeli padli v krizo, saj je postajalo vse bolj očitno, da lahko z njimi dobro ponazorimo zgolj višje kognitivne procese,⁹ npr. izpeljevanje, kjer iz pridobljenega znanja sklepamo na dogodke v svetu, nižjih kognitivnih procesov,¹⁰ kot je npr. zaznavanje, kjer čutni podatki zgolj opisujejo dogodke v svetu, pa ne. Nov spoznavni model, ki naj bi to odpravil, sicer ni povsem nov, saj ima svoje korenine v modeliranju nevronskih mrež že omenjenih McCullocha in Pittsa, se je pa v času razcveta simbolnega modela nanj gledalo z omalovaževanjem in se ga skorajda povsem ignoriralo.¹¹ Zgleduje se po tem, kako so zgrajeni in kako delujejo naši možgani, ter sku-

⁸ Zanj glej npr. Block (1980/2007), Kim (1996), Bregant (2013).

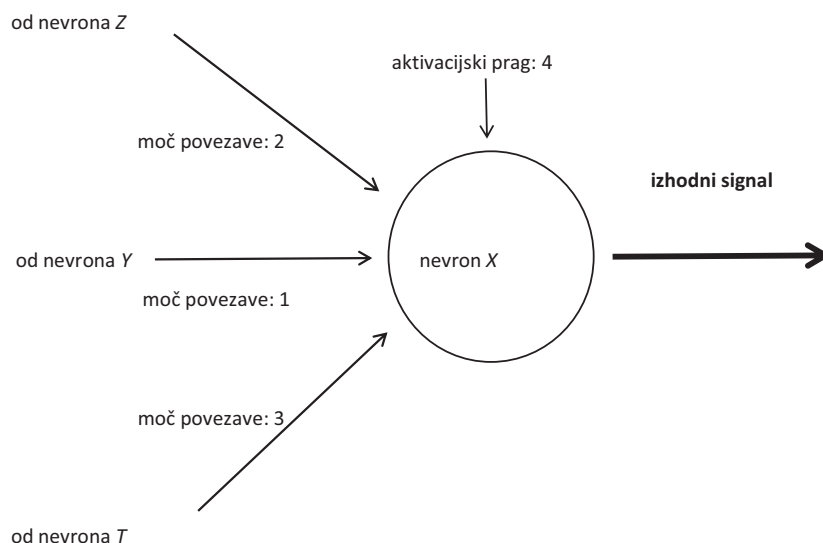
⁹ Ang. *top-down*.

¹⁰ Ang. *bottom-up*.

¹¹ Rosenblatt (1962) se takšnemu pristopu ni nikoli odrekel, zaradi česar je bil v zlati dobi simbolnih modelov (60. in 70. leta 20. stoletja) tarča pogostih kritik.

ša medsebojno povezanost njihovih osnovnih enot, nevronov, ki tvorijo prepleteno mrežo povezav, prenesti na bolj abstraktno raven. Z idealiziranim opisom nevronov v smislu njihovega prepletanja in delovanja nam ponudi novo orodje za modeliranje mentalnih procesov oziroma odgovarjanje na razlagalno nalogo, ki pa računalnika ne dojema več kot manipulatorja s simboli, ampak kot virtualno mrežo medsebojno povezanih enot. Takšen model naše kognicije, ki za vzor sicer jemlje zgradbo naših možganov, zanemarija pa podrobnosti delovanja nevronov in živčnega sistema, je dobil ime *konekcionistični model* (včasih se uporablja izraz nevronska mreža, vendar lahko ta pomeni tudi model na ravni dejanske, možganske zgradbe in ne abstraktne). Ker zagovarja nesimbolno procesiranje informacij, tako spodkopava idejo, da smo računalniki, vsaj v primeru, če slednje razumemo v klasičnem smislu, tj. kot napravo za obdelovanje predstav, prikazanih s simboli.

Konekcionistični model kognicije jemlje za vzor, kot smo omenili, delovanje dejanskega nevrona, zanj pa so z vidika delovanja bistvene naslednje štiri stvari: (a) da je povezan z drugimi nevroni, (b) da so te povezave različno obtežene (močne), (c) da obstaja prag, ki določa, ali je aktiviran ali ne in (d) da je tako »izključen« (0) ali »vključen« (1). Skratka, biološki nevroni tvorijo prepleteno mrežo povezav, ki so različno močne, slednje je deloma odvisno od premera vlaken, ki so povezana, deloma pa od kemične zgradbe stika (sinaps). Ko vsota vhodnih signalov določenega nevrona doseže ali preseže njegov aktivacijski prag, nevron »zagori« (se sproži) ter posreduje izhodni signal naprej svojemu najbližjemu sosedu. Oglejmo si to na naslednji poenostavljeni sliki dejanskega nevrona (Copeland, 1993: 209):



Slika 3: Delovanje dejanskega nevrona

Nevron X posreduje elektrokemične signale sosednjemu nevronu, tj. je »vključen« oziroma v stanju 1, ko vsota njegovih vhodnih signalov doseže ali preseže njegov aktivacijski prag, ki je 4, kar se lahko zgodi, npr., ko sta sprožena nevrona Y in T (skupna moč vhodnega signala je potem točno 4) ali ko so sproženi vsi trije nevroni, Y , T in Z (vsota moči vhodnega signala je takrat 6). Vse opisane nevronske povezave, ki prispevajo k vhodnemu signalu nevrona X , so bile v tem primeru vzbujevalne ali *ekscitirajoče*, tj. nevron X so spodbujale k proženju. Lahko pa so vhodne nevronske povezave tudi takšne, da sosednji nevron odvrtaajo od proženja, zaradi česar jih imenujemo zaviralne ali *inhibirajoče*, npr. če so sproženi vsi trije nevroni, Y , T in Z , vendar povezava med nevronoma X in Z ni vzbujevalna, ampak zaviralna (ne več +2, pač pa -2), nevron X ne bo sprožen, saj skupna moč vhodnega signala ne bo več 6, temveč 2, s čimer ne bo dosežen ali presežen njegov aktivacijski prag.

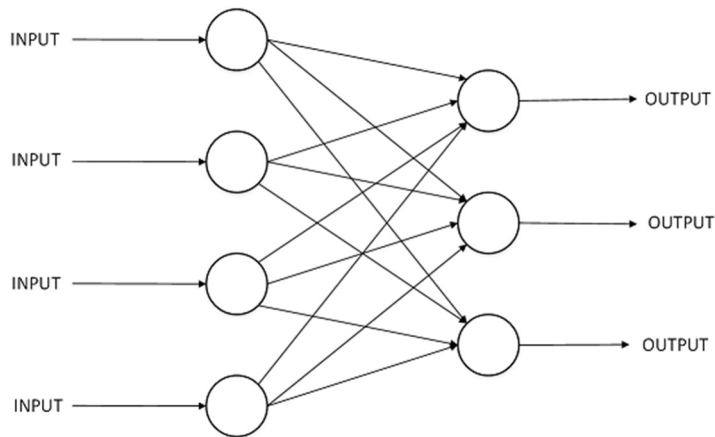
Zgradba konekcionističnih mrež

Podobno je zgrajen tudi konekcionistični model, pri čemer so njegovi osnovni gradniki, t.i. idealizirani nevroni, preproste, neinteligentne enote, ki so lahko vklopljene ali izklopljene. Takšni umetni nevroni so med seboj povezani, vsak izmed njih pa ima določeno aktivacijsko vrednost, ki jo preko vezi posreduje drugim enotam in s tem pripomore k povečanju (ekscitiranje) ali zmanjšanju (inhibiranje) njihove aktivacijske vrednosti, kar vpliva na to, ali se sprožijo ali ne.¹² Konekcionistični modeli so lahko zgrajeni na različne načine, ki se med seboj, kot pravi Markič (2010), po kateri jih povzemam, razlikujejo glede na (i) število plasti mrež, (ii) smerjo povezave med njenimi enotami, (iii) vrsto aktivacijske in izhodne funkcije in (iv) izbrano učno pravilo.

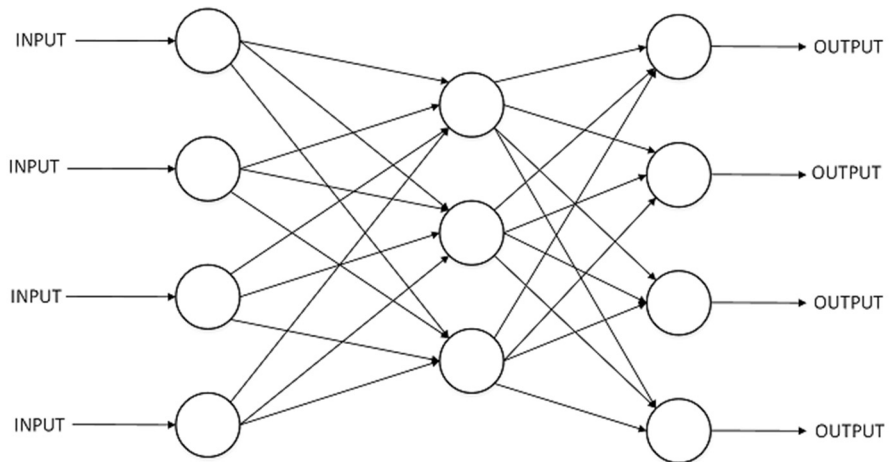
Z vidika števila plasti mrež ločimo *enoplastne*, kjer je vsaka enota hkrati vhodna in izhodna, *dvoplastne*,¹³ kjer prvi nivo enot predstavlja vhodne podatke, drugi pa izhodne, in *večplastne*, kjer je med vhomom in izhodom še več »skritih« ravni.

¹² Glej sliko 2.

¹³ Npr. Rosenblatt (1962) in njegov »Perceptron«; danes je to kar drugo ime za nevrnsko mrežo.

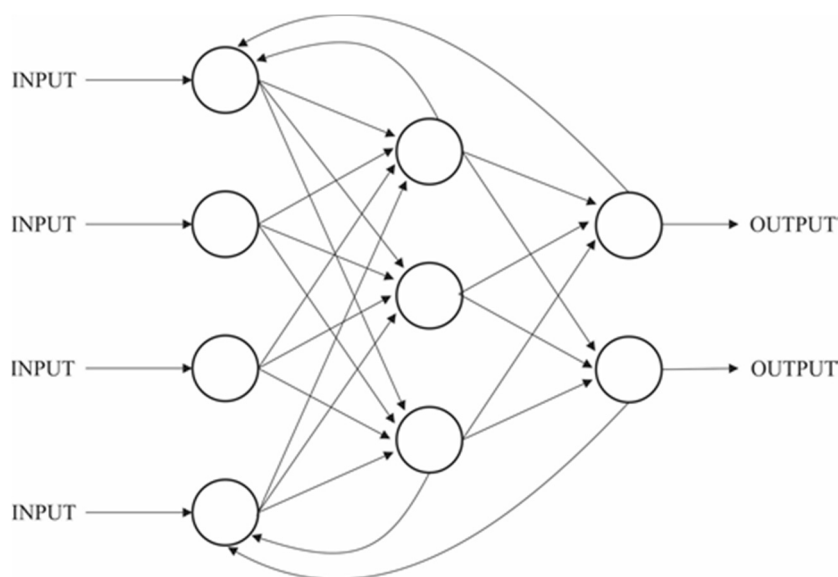


Slika 4: Dvoplastna mreža



Slika 5: Troplastna mreža

Z vidika smeri povezave so lahko slednje *enosmerne*, kjer so enote nižje ravni povezane samo z enotami višje ravni (ang. *feedforward*) in informacije potujejo zgolj v eni smeri (sliki 4 in 5), tj. naprej (od vhoda do izhoda), in *povratne*, kjer so tudi enote višje ravni povezane z enotami nižje ravni (ang. *recurrent*), kar omogoča pretok podatkov v obeh smereh, tj. naprej in nazaj.



Slika 6: Povratna mreža

Aktivacijska funkcija, ki določa način preslikave vhodnih podatkov v aktivacijsko stanje, je za a_i npr. lahko $\sum w_{ji} * a_j + \theta_i$ (w_{ji} je moč povezave med i in j , a_j stanje j enote, θ_i pa prag i enote), lahko pa so to tudi druge funkcije, npr. kvadratna $\sum w_{ji} * a_i^2 + \theta_i$ ali sferična $\sum (a_i - w_{ji})^2 - \theta_i$. Izhodna funkcija, ki določa način preslikave aktivacije enote v njen izhod, pa je lahko deterministična, v tem primeru vhod enolično določa izhod, ali stohastična, kjer vhod zgolj z večjo ali manjšo verjetnostjo določa izhod.

Učno pravilo, ki ga izberemo, da bi se sistem prilagajal okolju, pa je lahko *nenadzorovano* ali *nadzorovano*. V prvem primeru mreža sama postopoma okrepi vezi med dvema enotama, da bi v bodočnosti dosegala ustrezen izhod. Tukaj sta najbolj znani pravili *Hebbovo*, kjer se okrepijo povezave med istočasno aktivnimi enotami, ostale se oslabijo, in *tekmovalno*, kjer enote med seboj najprej tekmujejo, katera bo prva postala aktivna, tista, ki ji to uspe, pa potem okrepi svojo zvezo s tistimi, ki so ji do zmage pomagale, z ostalimi pa jo oslabi. V drugem primeru pa mrežo učimo tako, da izhodne podatke primerjamo s pravilnimi odgovori, zato ga imenujemo tudi učenje z učiteljem. Tukaj so najbolj znana pravila *delta*, kjer učenje poteka s pomočjo učnih primerov, tj. z znanimi pari vhodnih in izhodnih podatkov, pri čemer se pri danem vhodu najprej izračuna izhod, potem pa še razlika med dobljenim in želenim izhodom, skladno z njo pa se potem ustrezno zmanjšajo ali povečajo moči povezav med enotami.¹⁴

¹⁴ Za podrobnosti glej npr. Bechtel in Abrahamsen (2002), Peruš (2001), Markič (2010), Guid in Strnad (2007).

Pomembna značilnost konekcioniistične mreže je ravno učenje, ki ji omogoča samostojno posploševanje iz vzorcev (brez pomoči programerja), kar sta pokazala že Rumelhart in McClelland (1987), ko sta jo naučila tvoriti preteklik angleških pravih in nepravih glagolov. Učenje je potekalo preko mehanskih prilagajanj moči povezav med enotami in je med drugim vključevalo, npr., dodajanje končnice »ed« vhodnim podatkom pri formulaciji izhodnih odgovorov v primeru pravih glagolov, konekcioniistična mreža pa je potem sama tvorila preteklik tudi za tiste glagole, ki jih v procesu učenja ni spoznala. Začela sta s tipičnimi glagoli, ki se jih naučimo najprej, npr. priti, imeti, dati, in jih kot nedoločnike predstavila na vhodni ravni, potem pa sta primerjala dobljene in želene izhode ter glede na razliko med njimi spreminjala moči povezav med enotami, kot to določajo pravila delta¹⁵ (Copeland, 1993: 215).

Prednosti konekcioniističnih sistemov

Privlačnost konekcioniistične obdelave podatkov tiči v njeni podobnosti človeškemu procesiranju informacij, konekcioniistične mreže pa imajo po Copelandu (1993) šest lastnosti, ki jih najdemo tudi pri človeku. Ni presenečenje, da so na tem seznamu ravno tiste, v katerih se možgani razlikujejo od klasičnih računalnikov, saj konekcionizem, kot smo omenili, govori prej v prid trditvi, da jim ljudje nismo podobni, kot pa, da smo jim. Prva je *biološka podobnost*, ki pa se odvija zgolj na abstraktni ravni, saj konekcioniistične mreže npr. zanemarijajo razlike, ki obstajajo med dejanskimi nevrotransmiterji, ne izkazujejo kompleksnosti vzorcev bioloških nevronske povezave ali ne simulirajo zapletenega procesa dejanskega vzbujaanja, vendar »[t]e razlike niso problematične, v bistvu so celo zaželene, če se osredotočamo na konekcioniistične modele kot na kognitivne in ne biološke modele« (Bechtel, Abrahamsen, 2002: 57). Druga je *prepričanje, da poteka učenje v možganih na podoben način kot v mrežah*, tj. s spreminjanjem uteži povezav med nevroni preko izkušenj, kar možganom in mrežam omogoča učenje ter prilagajanje novim in odprtim vsakdanjim situacijam. Pomembno je ločiti med aktivacijsko vrednostjo enote, ki predstavlja njeno trenutno spremembo stanja glede na vhodne podatke, in utežmi povezav med enotami, ki so posledica ponavljajočih se sprememb ter konekcioniistični mreži omogočajo obdelavo vhodnih podatkov tako, kot se je naučila, tj. da znanje, ki je shranjeno v povezavah, uporabi pri odgovoru na novo vprašanje (Markič, 2010: 55). Tretja je *vzporedno procesiranje*, saj so v mrežah izhodni podatki posledica motnje vhodne plasti, ki povzroči z vidika aktivacije enot masovno vzbujaanje in zaviranje, v katerem sodelujejo vse hkrati, kar konekcioniističnim mrežam zaradi velikega števila med seboj povezanih enot omogoča, da v istem času izvedejo (približno) isto število operacij kot možgani. Četrta je *razpršeno shranjevanje podatkov*, kar pomeni, da znanje v mreži ni shranjeno na enem mestu (čeprav obstajajo tudi take), kot npr. v računalniku, am-

¹⁵ Za praktične primere glej npr. Mcleod, Plunkett, Rolls (1998), Peruš (2001), Guid in Strnad (2007).

pak je porazdeljeno po povezavah tako, da vsaka izmed njih hrani različne drobce iste informacije. V omenjenem primeru učenja tvorbe preteklika, znanje, ki je konekcionistični mreži omogočalo pravilno tvorbo preteklika danega nepravilnega glagola, ni bilo shranjeno v eni povezavi, ampak je bilo razpršeno po celem vzorcu povezav, ki so bile aktivne v času učenja in so se tako okrepile (Copeland, 1993: 222). Peta je *prepričanje, da so mreže v nekem smislu odporne na okvare*: če pride do prekinitve nekaterih povezav med enotami, ali celo do uničenja enote, konekcionistična mreža še vedno deluje, vendar se poveča možnost izhodnih napak. Z drugimi besedami, njena učinkovitost opravljanja kognitivnih funkcij pada s številom uničenih enot (dokler uspešno nadomešča manjkajoče podatke s približki, še nekako gre), podobno kot sposobnost izvajanja kognitivnih operacij možganov pada s številom njihovih uničenih nevronov. »V sistemih, ki uporabljajo porazdeljene reprezentacije, lahko izločimo precejšnje število enot, pa bo obnašanje takšnih sistemov še vedno zgolj malenkostno izkrivljeno« (Bechtel, Abrahamsen, 2002: 62). Šesta pa je *vsebinsko naslavljanje spomina*, ki mreži, ko se ne spomni takoj zahtevanega podatka, omogoča dostop do njega preko shranjenih namigov ali ključev, kot to počnemo ljudje, npr. če vemo, da smo knjigo posodili študentu, ne spomnimo pa se kateremu, si pri njegovi identifikaciji pomagamo s tem, da se skušamo spomniti, ali je bil npr. visok, svetlolas, debel itd. »Mreža je lahko narejena tako, da deluje po načelu vsebinskega naslavljanja spomina v smislu, da če je vhodni podatek zgolj *del* zapomnjenega vzorca, mreža kljub temu kot izhodni podatek proizvede popoln zapomnjen vzorec« (Copeland, 1993: 222).

Med dejanskimi nevroni in konekcionističnimi mrežami sicer ni popolne podobnosti in ta z vidika kognicije, kot pravita Bechtel in Abrahamsen (2002), tudi ni potrebna, zaradi česar nekatere razlike med možgani in konekcionističnimi sistemi niso pomembne, npr. različne vrste oziroma oblike bioloških nevronov, ki glede na to opravljajo določene in različne funkcije, česar konekcionistični modeli ne upoštevajo, prispevek nevrotansmitorjev, ki s kemičnim prispevkom premestijo špranjo (sinapso) na stiku med dvema povezavama, o čemer v konekcionističnih mrežah ni govora, in urejena ter specificirana geometrija možganov, ki jo sestavljajo prefinjeno urejeni popolni vzorci nevronske povezave, ki temeljijo na naloženih plasteh vodoravno položenih živčnih vlaken in tako tvorijo različne geometrijske oblike, kar v geometrijsko nedoločenih konekcionističnih sistemih zaman iščemo. Vendar pa vse ni tako rožnato. Strinjam se namreč s Copelandom (1993), da se konekcionistični modeli in dejanski možgani kljub vsemu razlikujejo v eni točki, ki je z vidika nesimbolnega pojasnjevanja naše kognicije s pojmi proženja in neproženja formalnih enot pomembna, in to je proces učenja. Mreža se uči izključno iz izkušenj oziroma števila ponovitev iste naloge, zanjo ne obstaja noben drug način, da bi se dokopala do znanja, kar pa v primeru človeka ne drži. Nekdo z dobrim spominom lahko nekaj zna že po prvem branju in mu v procesu učenja ni treba izgubljati časa in energije z neskončnimi ponovitvami, da bi si pridobil zadostno mero izkušenj. Poleg tega se mreže, ki uporabljajo nadzorovano učenje,

nečesa naučijo izključno s pomočjo učitelja, kar pa za naše možgane ne velja; ti so v tem pogledu veliko bolj samozadostni. Ko se učimo obvladati neko večšino ali spretnost, naši možgani nimajo predhodnega znanja o tem, kaj je pravi izhod, in ne računajo na zunanjega vsevednega znanstvenika, ki bo s specifikacijo poskrbel za proženje ustreznih nevronov, da bi ga proizvedli.

Searle (1990) pa z miselnim eksperimentom, imenovanim »kitajska telovadnica«, izpostavlja nekaj drugega, zaradi česar nobena konekcioniistična mreža naj ne bi bila sposobna kognicije. To je njena nezmožnost razumevanja pomenov, s katerimi operira, »kitajska soba«, ¹⁶ zdaj preurejena v telovadnico, pa naj bi to nedvoumno pokazala. Zamislite si telovadnico namesto sobe, v kateri imamo veliko zgolj angleško govorečih ljudi, ki izvajajo iste operacije kot enote in povezave v konekcioniistični mreži. Način, kako ljudje v ekipi simulirajo enote v mreži, ponudi Copeland (1993): posredovanje zelenih žetonov sosedom predstavlja vzbujevalno vez, rdečih zaviralno, število posredovanih žetonov enega sosedu drugemu pa predstavlja moč te povezave; trener oskrbi sodelujoče z navodili, v katerih so določene količine žetonov, ki jih je treba predati, in igralci, ki jih morajo dobiti, da bi izvedli neko akcijo (aktivacijski prag), kar predstavlja specifikacijo povezav; da pa bi ekipa dosegla zelene rezultate, igralci v času treninga spreminjajo oziroma prilagajajo navodila skladno s trenerjevimi zahtevami, kar predstavlja način učenja mreže. Searle trdi, da je rezultat na koncu enak kot v sobi, kjer s kitajskimi simboli s pomočjo knjig s pravili manipulira zgolj en človek. V ekipi (mreži) nihče ne govori kitajsko, čeprav se zunanjim opazovalcem tako zdi, in ne obstaja noben način, da bi se celota (mreža) naučila pomena kitajskih besed (Searle, 1990: 22).

Kakorkoli, tudi konekcioniistični model – ki ga Beer definira kot »/.../ večplastno mrežo preprostih, nevronom podobnih elementov, ki so naučene številčno vhodno predstavo pretvoriti v številčno izhodno« (Beer, 2000: 96) – naj bi bil sposoben rešiti razlagalno nalogo, tj. odgovoriti na omenjene štiri lastnosti naše kognicije: produktivnost, sistematičnost, kompozicionalnost in sklepalno koherentnost, kot poudarjajo njegovi zagovorniki (npr. Pollack, 1990, Chalmers, 1990, Markič, 2010), kar pa pomeni, da simbolni modeli niso več »edina igra v mestu«. ¹⁷ Ni več potrebno, da obdelava predstav naše spoznave (podatkov) poteka izključno preko programske vodenih operacij s simboli, ki temelji na fizičnem simbolnem sistemu, imenovanem »mentalese«, ampak lahko procesiranje informacij naše kognicije poteka tudi brez tega, tj. kot ekscitiranje in inhibiranje medsebojno povezanih umetnih nevronov. Sicer je res, da močna simbolna hipoteza pušča vprašanje na-

¹⁶ Za miselni eksperiment, imenovan »kitajska soba«, glej Searle (1980/1990), Preston (2002), Bregant (2013).

¹⁷ Za ugovore temu, da lahko konekcioniistični modeli kognicije resnično pojasnijo vse štiri lastnosti, ki jim jih pripisujemo, glej npr. Fodor, Pylyshyn (1988), Ramsey, Stich, Garon (1991), Bechtel, Abrahamsen (2002).

čina manipulacije s simboli povsem odprto, pravi zgolj, da so samo univerzalni simbolni sistemi (računalniki) zmožni misliti, a pojava konekcionističnih modelov kljub temu ne preživi. Z njimi se namreč pojavi alternativna razlaga tega, kako možgani funkcionirajo, ta pa ne vključuje nobenih simbolov in programov, ki vodijo in nadzirajo operacije z njimi, ampak zgolj kompleksne mreže med seboj povezanih enot, ki ena drugo vklaplajo in izklaplajo.

Tako imamo naenkrat napravo, tj. možgane, ki so sposobni misliti, niso pa računalniki v klasičnem smislu, kar postavlja trditev zagovornikov simbolnih modelov, da je pojasnitev opisanih štirih lastnosti kognicije možna zgolj v primeru, ko možgane razumemo kot računalnik, na laž, s čimer se spet kot feniks iz pepela dvigne prepričanje, da smo drugačni od njih.

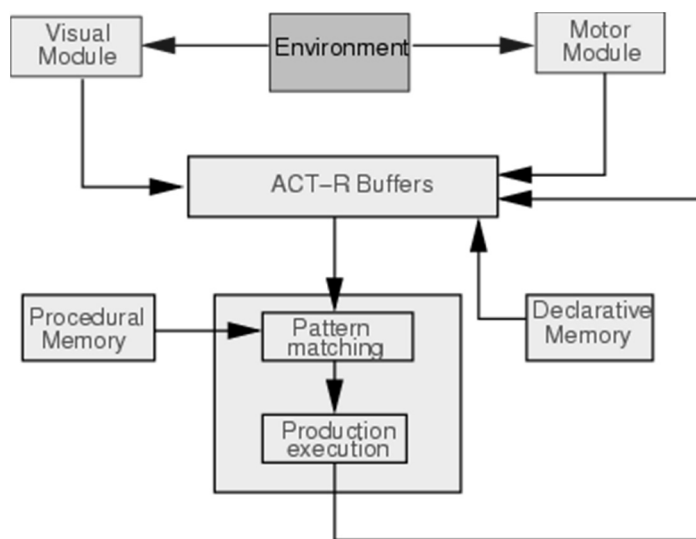
4 Drugi modeli kognicije

Hibridni model

Če na kratko povzamemo prednosti in slabosti simbolnih modelov in konekcionističnih mrež, vidimo, da je prednost simbolnih sistemov njihova uporabnost pri modeliranju višjih kognitivnih operacij, npr. izpeljevanja, njihova slabost pa posnemanje nižjih spoznavnih funkcij, npr. zaznavanja. Dodaten problem predstavlja njihova togost, na kar je opozoril že H. Dreyfus (1979), ki je posledica množice točno določenih navodil za manipuliranje s simboli, ki jim morajo slediti, ko modelirajo našo kognicijo, kar jim preprečuje posnemanje našega vsakodnevnega razmišljanja v nejasnih in odprtih situacijah. Na to, da je s simbolnimi modeli umetna inteligenca dosegla kritično točko in da pri izdelavi inteligentnega stroja ne napreduje več, je v obliki naslednje metafore opozoril S. Dreyfus: »/.../[napredek v tem], da nam bo uspelo narediti inteligentne računalnike, je podoben napredku moža, ki je za to, da bi bil bližje luni, splezal na drevo« (Dreyfus, Dreyfus, 1986: 10). Nasprotno je prednost konekcionističnih sistemov njihova primernost za modeliranje nižjih kognitivnih funkcij, npr. prepoznavanje vzorcev, slabost pa posnemanje nižjih spoznavnih operacij, npr. igranje šaha. Njihova dodatna težava je velika poenostavitev v primerjavi z biološkimi nevronske mrežami, kar postavlja pod vprašaj njihovo primernost za pojasnjevanje kognicije, ki črpa svojo verodostojnost iz dejanske zgradbe in delovanja možganov. Zato ne preseneča, da so se pojavile težnje po izgradnji kognitivnega modela, ki bi združeval prednosti simbolnega in konekcionističnega pristopa ter tako izničil njune posamezne slabosti.

Najbolj znan poskus povezovanja simbolnega modeliranja, ki v splošnem stavi na to, da ni obdelava podatkov nič drugega kot manipulacija z nizi simbolov in njihova transformacija, s konekcionističnim, ki v splošnem razume obdelavo podatkov kot vklapljanje in izklapljanje medsebojno prepletenih formalnih enot (umet-

nih nevronov), je t.i. ACT-R (ang. *Adaptive Control of Thought – Rational*), katerega avtor je Anderson (2007). Takšne modele imenujemo *hibridni*, zanje pa je značilno, da vsebujejo tako simbolno (zaporedno), kot tudi podsimbolno (vzporedno) obdelavo informacij, in ACT-R ni s tega vidika nobena izjema. Sestavljen je iz treh osnovnih elementov: (i) modulov, (ii) medpomnilnikov (ang. *buffers*) in (iii) usklajevalca vzorcev (ang. *pattern matching*). Spodnja slika ponazarja njegovo osnovno zgradbo:¹⁸



Slika 7: Osnovni deli ACT-R

Obstajata dve vrsti modulov: (a) *zaznavno-motorni*, ki je odgovoren za interakcijo z resničnim okoljem oziroma z njegovo simulacijo – tukaj sta najbolj razvita vizualni in manualni, obstajata pa še, npr., govorni in slušni –, in (b) *spominski*, ki je sestavljen iz *deklarativnega spomina* (ali znanja, da) – ta hrani podatke tipa »Ljubljana je glavno mesto Slovenije«, »Človek je sesalec« ali »5+5=10« –, in *proceduralnega spomina* (znanje, kako), ki proizvaja t.i. produkcijska pravila, ki niso nič drugega kot znanje o tem, kako delamo stvari, npr. znanje o tem, kako se vozi, kako se kuha, kako se seštevata itd. Medpomnilniki se uporabljajo za dostop do modulov (razen do proceduralnega spomina) in predstavljajo vmesnik med njimi in sistemom. Njihova vsebina v danem trenutku predstavlja trenutno stanje sistema. Usklajevalec vzorcev išče produkcijsko pravilo, ki se ujema s trenutnim stanjem oziroma zahtevami medpomnilnikov, vsebuje pa navodilo, kako naj se sistem v tem trenutku obnaša oziroma kaj naj naredi. Pomembno je, da je lahko v

¹⁸ Vir: <http://act-r.psy.cmu.edu/about/>.

vsakem koraku, ker deluje ta modul kot zaporedni procesor, aktivno le eno produkcijsko pravilo, ki mu sistem sledi, kar povzroči prilagoditev medpomnilnikov in spremembo stanja sistema. Vprašanje je, kako se ACT-R odloči, katero produkcijsko pravilo bo za izvedbo določenega koraka izbral oziroma uporabil. Pri tem se usklajevalec vzorcev ravna po načelu najvišje koristi za celoten sistem v danem trenutku, ki izhaja iz *teorije racionalne izbire*,¹⁹ imenu ACT-R pa prispeva črko »R«, ki predstavlja »racionalnost«. Skratka, usklajevalec vzorcev izvaja neke vrste analizo stroškov in koristi, t.i. *podsimbolno računanje*, ki vključuje uporabo ustreznih enačb (simbolov) in na osnovi tega brez kakršnegakoli stika z osrednjo nadzorno enoto (ki je sploh ni) sam od sebe izbira najbolj učinkovita produkcijska pravila.

Vidimo, da vsebuje ACT-R različne module, ki izvajajo različne kognitivne naloge, tip procesiranja v njih pa je odvisen predvsem od tega, kakšne vrste so. Glede na to izvaja *simbolno* in *podsimbolno*²⁰ obdelavo podatkov, pri čemer ga prvo dela za simbolni, drugo pa za konekcionistični sistem. Na simbolni značaj ACT-R tako kaže način zgradbe proceduralnega ali implicitnega spomina, ki se nanaša na obvladovanje pravil in postopkov, kjer je znanje predstavljeno kot procesna pravila v obliki parov *če-potem* (ang. *if-then*), ter način delovanja usklajevalca vzorcev, ki z obdelavo informacij korak za korakom funkcionira kot zaporedni procesor. Na njegov konekcionistični značaj pa kaže način zgradbe deklarativnega ali eksplicitnega spomina, ki vsebuje podatke o dejstvih in dogodkih, kjer je znanje predstavljeno v obliki semantičnih mrež, način delovanja njegovega priklica, tj. ali bo neko dejstvo priklicano in kako hitro, kar je odvisno od njegove relevantnosti v danem kontekstu in njegove uporabnosti v preteklosti, ter način izbire najbolj učinkovitega produkcijskega pravila, ki se odvija na osnovi izračuna stroškov in koristi: produkcijska pravila so res predstavljena simbolno in potek njihovega delovanja je vnaprej določen, ni pa vnaprej določeno, katera se bodo izvajala in v katerem primeru (večino časa namreč obstaja množica pogosto med seboj podobnih produkcijskih pravil, ki bi lahko bila aktivna, če bi bila izbrana).

Hibridni model kognicije je tako poskus združiti najboljše obeh svetov, simbolnega in konekcionističnega, v želji dokopati se do bolj celovitega razumevanja našega duha, ki nosilce informacij ali *gradnike*, obravnava kot simbolne predstave, katerih delovanje je določeno s procesnimi pravili tipa *če-potem* in ki so skrite znotraj deklarativnega spomina. Gradniki morajo biti za to, da bi bili učinkoviti, iz njega priklicani in nameščeni v medpomnilnik, kar pa urejajo od simbolne zgradbe gradnikov povsem neodvisne podsimbolne (porazdeljene) enačbe, ki do-

¹⁹ *Teorija racionalne izbire* ali *racionalna teorija ukrepanja* je stališče, ki kot kriterij pri izbiri poti za doseg nekega cilja poudarja in upošteva njeno učinkovitost oziroma koristnost, ki se odraža na delovanju celotnega sistema.

²⁰ Ali *porazdeljeno*.

ločajo aktivacijski nivo vsakega gradnika kot funkcijo njegove uporabnosti v preteklosti in koristnosti za trenutno situacijo.

Dinamični model

Dinamični pristop k modeliranju duha ali dinamična teorija, ki jo Horgan in Tien-son opredeljujeta kot »[s]kupek matematičnih pojmov, tehnik in rezultatov /.../« (Horgan, Tienson, 1992: 27), se pomembno razlikuje od simbolnega in konekcionističnega pogleda. Naše kognicije ne skuša pojasniti s pojmom mehanizma, ampak se pri odgovarjanju na vprašanja, ki se zastavljajo v zvezi z njo, po pomoč spet zateče k t.i. deduktivno-nomološkemu modelu razlage. Tukaj je eksplanandum izpeljan iz eksplanansov, ki vsebujejo vsaj en naravni zakon, zaradi česar sledi nujno, hkrati pa je njegova razlaga tudi znanstvena,²¹ z drugimi besedami, s tem modelom nek pojav razložimo tako, da pokažemo, kako ga je mogoče izpeljati iz začetnih pogojev in naravnih zakonov.

Dinamični pogled ni nekaj resnično novega, saj je z nami že kar lep čas – in sicer od začetkov kibernetike, ki segajo v sredino 20. stoletja – kot alternativa računalniško-algoritemskemu (simbolnemu) in informacijsko-procesualnemu (konekcionističnemu) pristopu, na svoji priljubljenosti pa hitro pridobiva predvsem v zadnjem času (van Gelder, 1998, Spivey, 2007). Dinamični sistemi razumejo psihične procese kot kompleksne in nelinearne procese, ki vzniknejo iz preprostejših stanj ter so sposobni samoorganizacije. Razvijajo se skozi čas in predstavljajo zgolj opis ene izmed možnih alternativ, ki se, če se pojavi v stvarnosti, uresniči oziroma udejanji. Običajen razlog, zakaj naj bi bila dinamična teorija najboljši način za pojasnitev naše kognicije, se skriva v tem, da je zelo uspešna pri razlagi drugih naravnih pojavov, in ker so tudi naši možgani in naša mentalnost del narave, se zdi razumno pričakovati enak uspeh tudi pri pojasnitvi psihičnih procesov, mogoče celo večjega, kot smo ga dosegli s simbolnimi in konekcionističnimi modeli.

Bermudez (2010) opredeljuje dinamične sisteme na splošno kot katerekoli sisteme, ki se razvijajo po določenih zakonih skozi čas, ali, kot pravi Beer (2000), gre za matematične objekte, ki nedvoumno opisujejo, kako se stanje sistema spreminja s časom. Formalno jih lahko zapišemo kot trojko, ki jo sestavljajo urejen časovni niz T , prostor stanja S in razvojni operator ϕ_t , ki spremeni začetno stanje sistema, X_0 , po preteku določenega časa v neko drugo stanje sistema, npr. X_t (Beer, 2000: 92). Horgan in Tienson pravita, da je njihova tipična značilnost to, da vsebujejo specifikacijo dejanskih in hipotetičnih stanj, v katerih se lahko skozi čas znajdejo, tako opisana možna stanja sistema pa oblikujejo njegov abstraktni prostor stanj, v katerem slednje predstavljajo točke. Dinamični model je zato zbirka

²¹ Iz eksplanansa kot začetnih pogojev $C_1, C_2 \dots C_n$ in univerzalnih zakonov $L_1, L_2 \dots L_n$ sledi eksplanandum, npr. pojav P , kot znanstveno razložena deduktivna posledica.

smeri gibanja v času, ki jih ustvarijo različne točke prostora in ki bi jim sistem lahko sledil, zaradi česar ga lahko v nekem smislu identificiramo z nizom trajektorij, ki izhajajo iz stanj prostora, podobno kot lahko simbolni model enačimo z nizom njegovih teoremov (Horgan, Tienson, 1992: 30).

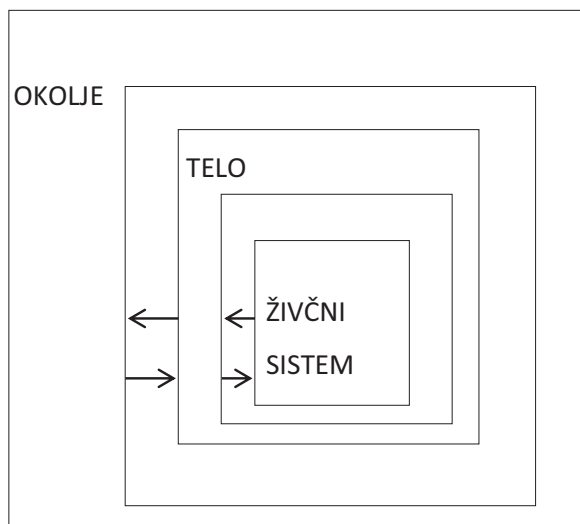
Beer (2000) navaja kot primer dinamičnega sistema *jezik*, za katerega se sicer misli, da je paradigmatični primer simbolnega sistema in tako neprimeren za konekcionistično in kakršnokoli drugo nesimbolno modeliranje, kar pa po njegovem ne drži. Pri tem se sklicuje na Elmana (1990), ki naj bi s posplošenim delta pravilom (ang. *backpropagation*) povratno naprej usmerjeno večslojno mrežo naučil predvidevati naslednjo besedo v stavku, pri čemer je uporabljal nabor 10000 kratkih stavkov, oblikovanih iz 29 samostalnikov in 21 glagolov. Vprašanje je, zakaj njegov poskus pooseblja dinamični in ne konekcionistični model. Beerov odgovor je, da zato, ker sta razumevanje in proizvodnja jezika v svojem bistvu časovna dogodka, kjer predhodne besede močno vplivajo na tolmačenje in izbiro tistih, ki jim sledijo, ta pomembnost časa pa je v Elmanovem modelu zaobjeta v kontekstualnih enotah, ki hranijo stanje skritih enot v prejšnjem časovnem koraku in jim ga v obliki signalov posredujejo v naslednjem ter tako vplivajo na oblikovanje izhodnih odgovorov sistema. »Takšne mreže so neavtonomni časovno-diskretni dinamični sistemi« (Beer, 2000: 91).

Kljub temu, da lahko npr. dinamične sisteme opišemo v simbolnem jeziku, da so, kot smo videli v opisanem Elmanovem primeru, nekateri konekcionistični modeli dinamični in da lahko nekatere dinamične sisteme predstavimo kot konekcionistične mreže, razlike med njimi obstajajo. Beer (2000) trdi, da se nanašajo na teoretični besednjak in slog pojasnitve, ki sta v njih uporabljena pri reševanju problemov, pred katere nas postavlja kognicija. V simbolnem modelu se pojasnitev osredotoča na vsebino uporabljenih reprezentacij in naravo ter učinkovitost uporabljenih algoritmov, v konekcionističnem na arhitekturo mreže, učni algoritem in razvijanje vmesnih porazdeljenih predstav, v dinamičnem pa na zgradbo prostora možnih poti/stanj in notranje in zunanje sile, ki z vzajemnim delovanjem oblikujejo tisto, ki se udejanji skozi čas.

Van Gelder (1998) je bil že pred več kot desetletjem prepričan, da bo kognitivna znanost prejkoslej ugotovila, da so kognitivni akterji dinamični sistemi, ki jih lahko izrazimo kot niz diferencialnih in/ali diferenčnih enačb,²² ki opisujejo, kako se njihova stanja spreminjajo skozi čas. Do sedaj se to sicer še ni zgodilo, saj se zdi, da sta glavna tekmovalca pri pojasnjevanju naše kognicije še vedno simbolni model in konekcionistična mreža, morda še Andersonov ACT-R kot kombinacija (hibrid) obeh, dinamični sistemi, čeprav so združljivi z nečim, kar imenujemo *ute-*

²² Diferencialna enačba je enačba, ki omogoča spremljanje zveznih sprememb sistema, npr. pospeševanje padajočega predmeta, diferenčna enačba pa enačba, ki omogoča spremljanje diskretnih sprememb sistema, npr. spreminjanje biološke populacije v času enega leta.

lešena kognicija (Varela et al., 1991), s čimer mislimo na to, da je kognicija odvisna od izkustev, ki jih priskrbi naše, v okolje umeščeno, telo,²³ kar izgleda razumno in privlačno, pa še vedno stopicajo v ozadju. Spodnja slika prikazuje dinamični pogled na utelešeno kognicijo:²⁴



Slika 8: Živčni sistem, telo in okolje so pojmovani kot dinamični sistemi, ki so v stalni interakciji.

Fussella (2013) meni, da dokler ne bo prišlo do spremembe temeljne paradigme, od tega, da smo digitalni računalniki, ki operirajo kot zaporedni procesorji simbolov, v to, da smo večdimenzionalni računalniki, ki se spreminjajo skozi čas, sestavljeni iz v naravo umeščenega in od nje neločljivega telesa, kognitivna znanost ne bo pripravljena na napredek.

5 Sklep: kaj smo?

Vrnimo se po tem izletu na področje alternativnih kognitivnih modelov k našemu začetnemu vprašanju, tj. ali resnično nismo nič drugega kot računalnik, pri odgovoru na katerega igrata glavno vlogo simbolni in konekcionistični model, ter si oglejmo, do kod nas je naša razprava pripeljala. Videli smo, da močna simbolna hipoteza možgane razume kot manipulatorje s simboli, zaradi česar naj ljudje ne bi bili nič drugega kot računalnik v klasičnem smislu, tj. naprava za obdelovanje predstav, prikazanih s simboli; konekcionistična hipoteza pa kot večplastne pre-

²³ Kognicija je tako posledica interakcije med telesom in okoljem.

²⁴ Slika je iz Beer, 2000: 97.

pletene mreže, zaradi česar naj ljudje ne bi bili računalnik, vsaj ne v tradicionalnem smislu. Vse lepo in prav, problem se pojavi, ko ugotovimo, da lahko računalnik simulira ali reproducira tudi obnašanje oziroma delovanje umetne nevronske mreže, saj se naenkrat pojavi sum, da sta oba, baje tako različna, pristopa k pojasnjevanju naše kognicije, simbolni in konekcionistični, v resnici združljiva. Ali to drži?

Zagovorniki močne simbolne hipoteze odgovarjajo, da vzporedno procesiranje informacij z možnostjo vsebinskega naslavljanja spomina, kjer je medij izmenjave podatkov vzbujanje in zaviranje aktivnosti umetnih nevronov, ne pomeni zavrnitve simbolno-programskega pristopa k umetni inteligenci, saj bi bilo načeloma možno narediti (neklasičen) računalnik, nek hibrid, ki bi bil vzporedni procesor z informacijami, predstavljenimi s simboli, in bi deloval po načelu manipuliranja z njimi. »Vzporedni stroji, ki jih imamo v mislih, bi podatke še vedno hranili v simbolnem zapisu, njihove operacije pa bi še vedno temeljile na strukturno občutljivih manipulacijah z gradniki tega zapisa« (Copeland, 1993: 243). (Nekaj takega danes vsaj deloma pooseblja predstavljeni hibridni model naše kognicije ACT-R s svojo simbolno zgradbo proceduralnega spomina in zaporednim delovanjem usklajevalca vzorcev na eni strani, ter konekcionistično zgradbo deklarativnega spomina in vzporednim načinom izbire najbolj učinkovitega produkcijskega pravila na drugi.) Oglaševalci konekcionistične hipoteze se s tem ne strinjajo in trdijo, da takšen program, ki bi lahko do popolnosti simuliral obnašanje možganov, človeku, ki bi mu ga npr. vsadili v glavo, sam po sebi še ne bi omogočal mišljenja, saj bi slednjega zgolj simuliral, simulacija mišljenja pa še ni dejansko mišljenje, tako kot simulacija mleka še ni pravo mleko. Z drugimi besedami, simboli programa se ne nanašajo na mišljenje, ampak zgolj na enote v mreži, njihov aktivacijski prag in prepletenost, zato pove program zgolj nekaj o mrežah, nič pa o mišljenju. Sutherland (1986) to podkrepi s tem, da četudi bi bilo s programom možno simulirati mišljenje, takšen program ne bi bil misleč nič bolj, kot bi bil vetroven program, ki bi simuliral orkan.

Predpostavimo zaradi argumenta, da simbolni modeli in konekcionistične mreže resnično niso združljivi in da nekaj takega, kot je vzporedni procesor z informacijami, predstavljenimi s simboli, ki bi deloval po načelu manipuliranja z njimi, npr. hibridni sistem, ne premosti ključnih razlik med njima, ki bi obema pristopoma dovoljeval identifikacijo možganov z računalnikom. Kljub temu bi naši možgani in s tem tudi mi sami še vedno lahko bili računalnik, ne sicer v von Neumannovem smislu, ki ga opredeljuje kot zaporednega procesorja informacij, zapisanih kot niz simbolov, pa tudi ne v hibridnem smislu, ki ga definira kot vzporednega obdelovalca podatkov, predstavljenih v simbolnem jeziku. Računalnik lahko namreč razumemo tudi širše, in sicer kot *napravo, ki obdeluje kakršnokoli vrsto predstav (ne zgolj simbolne) na kakršenkoli način (ne zgolj zaporedno)* (Copeland, 1993: 205). Če posvojimo to definicijo, se tudi v primeru, ko mislimo, da kognitivizem in konekcionizem nikakor nista združljiva in da je pravi pristop k

pojasnjevanju naše kognicije drugi, ne pa tudi prvi, ne moremo izogniti temu, da naši možgani, s tem pa tudi mi sami, ne bi bili v nekem tretjem, tj. splošnem, smislu računalnik. »Posledica tega, kar sem do sedaj povedal, je, da četudi bi možgani bili sestavljeni iz vzporedno procesirajočih mrež [v katerih znanje ni predstavljeno posredno, tj. s simboli, ampak neposredno], bi še vedno lahko bili računalnik« (Copeland: 1993: 246). Brezizhodnost situacije, v kateri se znajdejo zgovorniki tega, da možgani s tem pa tudi ljudje nismo računalniki, ponazarja naslednja enostavna dilema:

1. Ali je resnična močna simbolna hipoteza ali konekcionizem.
 2. Če je resnična močna simbolna hipoteza, potem smo računalniki (v von Neumannovem smislu).
 3. Če je resničen konekcionizem, potem smo računalniki (v splošnem smislu).
- ∴ 4. *Ljudje smo računalniki.*

Sklepno misel prepustimo La Mettriju, ki sicer v mislih ni imel računalnikov, je pa že v 18. stoletju materialistično zapisal, da nismo nič drugega kot stroji: »Zaključimo, da je človek v grobem stroj in da je celotno veselje zgrajeno zgolj iz ene substance /.../. Ali bi bila potem organizacija takšnega stroja dovolj, da bi lahko pojasnili vse? Na to bom še enkrat odgovoril z da.« (La Mettrie, 1748/1996: 39, 26).

Brains in the hands of computers

The article defends with the help of the simple dilemma the belief that people, i.e. our brains, are in some particular sense a computer. Copeland (1993) claims that empirical differences between a computer and a brain do not reject the classic approach to the identification of a brain with a computer, i.e. the strong symbol hypothesis grounded in symbolic representation of our knowledge of the world. He argues that these differences merely show that our brains are not the traditional von Neumann computer. The connectionist approach to the explanation of our cognition fails in this as well. Despite applying the non-symbolic way of representing information about the world, which is supposed to be alien to computers, it does not succeed in preventing their identification with brains. We may still be computers, but of some other sort, a sort which is not possible only in principle as it is demonstrated by the most developed hybrid model of our cognition ACT-R, which combines both approaches. And even if we, for the sake of argument, accept the belief that a computer as a symbol manipulator is not compatible with neural networks, we are still not safe to claim that we are not computers in some general way.

Keywords: strong symbol hypothesis, connectionist networks, ACT-R, dynamical systems, language of thought

Literatura:

- Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe*. Oxford: Oxford University Press.
- Bechtel, W., Abrahamsen, A. (2002). *Connectionism and the Mind*. Oxford: Blackwell.
- Beer, R. D. (2000). »Dynamical approaches to cognitive science«. *Trends in Cognitive Science*, 3, str. 91–99.
- Bermúdez, J. J. (2010). *Cognitive Science*. Cambridge: Cambridge University Press.
- Block, N. (1980/2007). »Uvod: Kaj je funkcionalizem?«. V Markič, O in Bregant, J. (ur.), *Narava mentalnih pojavov*, Maribor: Aristej, str. 147–166.
- Bregant, J. (2010). »Ali lahko stroj misli?«. *Analiza*, 4, str. 55–72.
- Bregant, J. (2013). »Funkcionalizem, računalniki in 'kitajska soba'«. *Analiza*, 3, str. 67–95.
- Chalmers, D. J. (1990). »Syntactic Transformations on Distributed Representations«. *Connection Science*, 2, str. 53–62.
- Churchland, P. S. (1986). *Neurophilosophy: Toward a Unified Science of the Mind/Brain*. Cambridge: The MIT Press.
- Copeland, J. (1993). *Artificial Intelligence: A Philosophical Introduction*. Oxford: Blackwell.
- Darwin, C. (1859/2009). *O nastanku vrst*. Ljubljana: ZRC SAZU.
- Descartes, R. (1637/2007). *Razprava o metodi*. Ljubljana: ZRC SAZU.
- Dreyfus, H. L. (1979). *What Computers Can't Do: The Limits of Artificial Intelligence*. New York: Harper and Row.
- Dreyfus, H. L., Dreyfus, S. E. (1986). *Mind Over Machine*. New York: Macmillan.
- Elman, J.L. (1990). »Finding Structure in Time«. *Cognitive Science*, 14, str. 179–211.
- Fodor, J. (1975). *The Language of Thought*. Cambridge: Harvard University Press.
- Fodor, J. A., Phylyshyn, Z. W. (1988). »Connectionism and Cognitive Architecture: A Critical Analysis«. *Cognition*, 28, str. 3–71.

- Fusella, P. V. (2013). »Dynamic Systems Theory in Cognitive Science: Major Elements, Applications, and Debates Surrounding a Revolutionary Meta-Theory«. *Dynamical Psychology*, <http://dynapsyc.org/2013/Fusella.pdf>.
- Guid, N., Strnad, D. (2007). *Umetna inteligenca*. Maribor: FERi.
- Haugeland, J. (1986). *Artificial Intelligence: The Very Idea*. Cambridge: The MIT Press.
- Hobbes, T. (1651/2006). *Leviathan (Revised Student Edition)*. Cambridge: Cambridge University Press.
- Horgan, T., Tienson, J. (1992). »Cognitive Systems as Dynamical Systems«. *Topoi*, 1, str. 27–43.
- Kim, J. (1996). *Philosophy of Mind*. Boulder: Westview Press.
- La Mettrie, Julien Offray de (1748/1996). *Machine Man and Other Writings*. Cambridge: Cambridge University Press.
- Lashley, K. S. (1950/1960). »In Search of the Engram«. V Beach, F. A., Hebb, D. O., Morgan, C. T., Nissen, H. W. (ur.), *The Neuropsychology of Lashley*, New York: McGraw Hill, str. 478–505.
- Markič, O. (2010). *Kognitivna znanost*. Maribor: Aristej.
- Marr, D. (1982). *Vision. A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W. H. Freeman.
- Marx, K. (1885/1997). »Capital II: The Process of Circulation of Capital«. V *Collected Works of Karl Marx and Frederick Engels*, New York: International Publishers.
- McCulloch, W., Pitts, W. (1943/1990). »A Logical Calculus of the Ideas Immanent in Nervous Activity«. V Boden, M. A. (ur.), *The Philosophy of Artificial Intelligence*, Oxford: Oxford University Press, str. 22–39.
- McLeod, P., Plunkett, K., Rolls, E. (1998). *Introduction to Connectionist Modelling of Cognitive Processes*. Oxford: Oxford University Press.
- Peruš, M. (2001). *Biomreže, mišljenje in zavest*. Maribor: Satjam.
- Pollack, J. (1990). »Recursive Distributed Representations«. *Artificial Intelligence*, 46, 77–105.
- Preston, J. (2002). »Introduction«. V Preston, J. in Bishop, M. (ur.), *Views into the Chinese Room*, str. 1–50.
- Pylyshyn, Z. W. (1984). *Computation and Cognition: Towards a Foundation of Cognitive Science*. Cambridge: The MIT Press.

- Ramsey, W., Stich, S. P., Garon, J. (1991). »Connectionism, Eliminativism, and the Future of Folk Psychology«. V Ramsey, W., Stich, S. P., Rumelhart, D. (ur.), *Philosophy and Connectionist Theory*, London: Erlbaum, str. 199–228.
- Rosenblatt, F. (1962). *Principles of Neurodynamics*. Washington: Spartan.
- Rumelhart, D. E., McClelland, J. L. (1987). »Learning the Past Tenses of English Verbs: Implicit Rules or Parallel Distributed Processing«. V MacWhinney, B. (ur.), *Mechanisms of Language Acquisition*, Mahwah: Erlbaum, str. 194–248.
- 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, 20–25.
- Spivey, M. J. (2007). *The Continuity of Mind*. Oxford: Oxford University Press.
- Sutherland, N. S. (1986). »Removing the Mystery from Intelligence«. *The Times Higher Education Supplement*, Feb. 28, str. 6.
- Turing, A. (1950). »Computing Machinery and Intelligence«. *Mind*, 59, str. 434–460.
- van Gelder, T. (1998). »The Dynamical Hypothesis in Cognitive Science«. *Behavioral and Brain Sciences*, 21, str. 615–628.
- Varela, F., Thompson, E., Rosch, E. (1991). *The Embodied Mind*. Cambridge: The MIT Press.
- von Neumann, J. (1958). *The Computer and the Brain*. New Haven: Yale University Press.