# Ali konekcionizem vodi v eliminativizem?

*OLGA MARKIČ*

*POVZETEK*

*Konekcionistični pristop h kognitivnemu modeliranju je poleg zanimivih empiričnih modelov sprožil tudi vrsto teoretskih, filozofskih diskusij. Tako se zastavlja vprašanje, ali konekcionistični modeli (ob predpostavki, da se izkažejo kot ustrezni kognitivni modeli) potrjujejo eliminativistični materializem. V nasprotju z Ramseyem, Stichem in Garonom, ki zagovarjajo tako stališče, skušam pokazati, da je tako sklepanje premalo utemeljeno in da so konekcionistični modeli s porazdeljenimi reprezentacijami združljivi z zdravorazumsko psihologijo.*

*ABSTRACT*
*DOES CONNECTIONISM LEAD TO ELIMINATIONISM?*

*The connectionistic approach to cognitive modelling, apart from producing some interesting empirical models, has triggered a series of theoretical and philosophical discussions. Thus, the question arises as to whether the connectionistic models - provided that these are adequate cognitive models - affirm the eliminationistic materialism. Unlike Ramsey, Stich and Garon, who defend this position, I have endeavoured to show that there are insufficient grounds for such a conclusion, and that connectionistic models with distributed representations are compatible with commonsensical psychology.*

## Uvod

Konekcionizem predstavlja nov pristop h kognitivnemu modeliranju. Poleg modelov, ki skušajo predstaviti delovanje posameznih delov možganov, znanstveniki načrtujejo modele spoznavnih nalog (npr. prepoznavanje slik, učenje angleških nepravilnih glagolov, prevajanje v drug jezik), kjer ne gre za dejanske modele možganov, ampak za modeliranje na višji ravni. Na ta način prihaja konekcionizem na področje, kjer je prevladoval kognitivizem in klasični simbolni modeli. Hipoteza kognitivizma je, da človekovo vedenje lahko razložimo s pomočjo simbolnih mentalnih reprezentacij in pravil nad njimi. Ker je tudi računalniški program množica pravil za manipuliranje s podatkovnimi strukturami, je možna analogija med delovanjem računalniškega programa in duha, in tako je iskanje ustreznih psiholoških razlag podobno iskanju ustreznih programov, ki jih duh/možgani izvajajo. To odpira možnost načrtovanja kognitivnih modelov, ki jih lahko tudi empirično preskusimo. Konekcionistični modeli (uporabljata

se tudi izraza nevronske mreže in vzporedno porazdeljeno procesiranje) se od klasičnih simbolnih modelov bistveno razlikujejo v tem, da ne poznajo eksplicitnih pravil nad reprezentacijami, temveč delujejo po pravilih, ki so določena na ravni posameznih enot.[1] Zato, in zaradi drugačnega pristopa k izdelavi modela, kjer je poudarek na učenju (prilagajanju okolju), predstavlja konekcionizem tudi nov pristop k razlagi kognitivnih fenomenov. Ob tem se zastavlja več vprašanj, povezanih z umestitvijo konekcionizma med kognitivne teorije: v kakšnem odnosu je konekcionizem do kognitivizma in klasičnih simbolnih modelov, kakšen je njegov odnos do nevroznanosti in ali je konekcionizem združljiv z razlagami in napovedmi vedenja, kot jih uporabljamo v naši vsakdanji praksi (zdravorazumska psihologija).

V tem prispevku se bom omejila na zadnje vprašanje. V kritični pretres bom vzela trditev Ramseya, Sticha in Garona (1991), da konekcionizem vodi v eliminativizem. Omenjeni avtorji iz ustreznosti konekcionizma kot kognitivnega modela sklepajo na eliminacijo zdravorazumske psihologije in s tem na nezdružljivost konekcionizma in zdravorazumske psihologije. V nadaljevanju bom skušala pokazati, da narava konekcionističnega procesiranja ne podpira tako radikalnega sklepa. Najprej bom opredelila pojem eliminativističnega materializma in navedla primere iz zgodovine znanosti za ontološko konzervativne in ontološko radikalne teoretske spremembe. Temu bo sledila predstavitev načela propozicionalne modularnosti, ki je po mnenju Ramseya, Sticha in Garona ključno za karakterizacijo zdravorazumske psihologije, in analiza njihovega eliminativističnega argumenta. V razdelku "Višje ravni analize v konekcionističnih modelih" bom pokazala, da njihov sklep temelji na preozkem gledanju na konekcionistične mreže, ki upošteva le raven posameznih procesnih enot. Če pa v razlagi upoštevamo tudi višje ravni analize, potem lahko najdemo stanja sistema, ki ustrezajo intencionalnim entitetam zdravorazumske psihologije.

**Eliminativistični materializem**

V filozofiji duha in v kognitivni znanosti je zdravorazumska psihologija (angl. folk psychology) že nekaj časa v središču pozornosti. Na eni strani so intencionalni realisti (kot npr. Fodor), za katere sta tako ontologija kot vzročna dinamika zdravorazumske psihologije v bistvu pravilni. Prepričanja, želje, namere itd. dejansko obstajajo in te intencionalne entitete so dejansko vzročno vključene v ustvarjanje in razlago vedenja. Tako stališče jemlje zdravorazumsko psihologijo za osnovo kognitivne znanosti. Kognitivizem in klasični simbolni modeli predstavljajo empirično teorijo, ki kaže, kako so lahko prepričanja in želje kot vzročno učinkujoča stanja uprimerjena v fizičnem mehanizmu.

Na drugi strani so intencionalni eliminativisti, za katere je zdravorazumska psihologija primer napačne teorije. Kot trdi zagovornik eliminativističnega materializma Paul Churchland, je zdravorazumska psihologija napačno in zavajajoče pojmovanje vzrokov vedenja in narave kognitivne aktivnosti. Ni le nepopolna predstava naše notranjosti, temveč je napačna predstava naših notranjih stanj in aktivnosti. Zato ne moremo pričakovati ustrezne nevroznanstvene teorije, ki bi se ujemala s kategorijami našega zdravorazumskega okvirja. Posledica tega je, da bo namesto redukcije na

[1] Najpreprosteje si konekcionistični model predstavljamo kot množico med seboj povezanih enot, ki prek različno močnih povezav (moč povezave predstavlja utež) prenašajo impulze. Sestavni del modela je še učni algoritem, s pomočjo katerega se mreža na osnovi primerov uči obvladovanja naloge (oziroma se prilagaja okolju), ne da bi ji vnaprej povedali pravila. Več o matematičnih osnovah glej. Dobnikar (1990), poljudnejši uvod pa v Markič (1992).

nevroznanost stari okvir zdravorazumske psihologije (prepričanja, želje, namere, bolečine, občutki itd.) preprosto odstranjen (eliminiran) in nadomeščen z razvito nevroznanstveno teorijo. Razlage vedenja se bodo sklicevale na nevrofarmakološka stanja in nevralne aktivnosti v specializiranih področjih možganov, oziroma na tisto, kar bo relevantno v novi teoriji (Churchland, 1988). V uspešnost redukcije zdravorazumske psihologije na kognitivno znanost dvomi tudi Stephen Stich (1983), ki prav tako meni, da za intencionalna psihološka stanja ni mesta v kognitivni teoriji.

Tak odnos med novo, uspešnejšo teorijo in staro teorijo ni posebnost psihologije. Eliminativistični materialisti navajajo flogiston, eter in čarovnice kot primere za eliminativizem v zgodovini znanosti.

Jasen in za večino nesporen je primer flogistona. Vse do Lavosiera (1743-1794) so bili znanstveniki prepričani, da se ob gorenju sprošča nekakšna duhu podobna substanca, ki so jo imenovali flogiston. Njegova odkritja o sestavi zraka in vlogi kisika pri dihanju in gorenju pa so prispevala k nastanku moderne kemije, ki je postavila zakon o ohranitvi materije. Ker se je izkazalo, da pri gorenju sodeluje kisik iz atmosfere, je flogiston postal ne le nepopolen, ampak radikalno neustrezen opis dogajanja. V novi teoriji flogiston ni imel več nobene vloge in ker ni bil primeren za redukcijo ali identifikacijo s kakšnim pojmom znotraj nove kemije, je bil preprosto odstranjen iz znanosti. Podobno se je zgodilo tudi z etrom, za katerega so fiziki v preteklosti menili, da omogoča prenašanje svetlobe, toplote in elektrike. Kasnejše teorije so tako razlago zavrgle in tako eter danes nima več mesta v sodobni fiziki.

Ob dveh primerih s področja naravoslovnih znanosti eliminativisti kot očiten primer eliminacije navajajo še čarovnice. V srednjem veku obstoj čarovnic ni bil pod vprašanjem. Duševne bolezni so razlagali kot primere, kjer se je Satanov duh naselil v žrtve, ki so zato kazale nenormalno vedenje. Dandanes smo mnenja[2], da je pojem čarovnice del pojmovnega okvirja, ki povsem napačno prikazuje in razlaga pojave, zato je dobesedno uporabo pojma čarovnice treba zavreči. Nove teorije o duševnih boleznih vodijo k eliminaciji čarovnic iz resne ontologije.

Eliminativist mora torej najprej pokazati, da je teorija, v kateri nastopajo določene entitete, neustrezna in jo je treba zamenjati z boljšo teorijo. (Kdaj je neka teorija boljša kot druga, je samo po sebi težko vprašanje. Običajno velja, da je nova teorija, s pomočjo katere dobimo natančnejše napovedi in boljšo razlago na večjem področju, boljša od stare. Pri tem se mora vsaj tako dobro kot stara teorija ujemati tudi s teorijami s sosednjih področij.) Vendar to še ni dovolj. Nova teorija lahko nadomesti staro teorijo, vendar hkrati ne eliminira njenih postavk. Prva možnost je, da te iste postavke še bolj natančno opredeli. Primer za to je Kopernikova heliocentrična teorija. Čeprav je povsem drugače opredelila naravo planetov in njihovo gibanje kot stara Ptolomejeva teorija, je kljub temu še vedno govorila o planetih. Druga možnost pa je, da postavke stare teorije reducira na postavke nove teorije. Največkrat omenjan primer za to možnost je redukcija temperature na povprečno kinetično energijo molekul.

Teoretske spremembe, pri katerih se entitete in procesi stare teorije ohranijo ali pa reducirajo na nove, Ramsey, Stich in Garon (1991, str. 202)[3] imenujejo ontološko konzervativne teoretske spremembe. Teoretske spremembe, ki niso ontološko konservativne, pa so ontološko radikalne. Ker ni splošnega načela, s katerim bi si lahko pomagali pri ugotavljanju, kdaj je teoretska sprememba ontološko radikalna, Ramsey, Stich in Garon predlagajo, da se pogleda postavke stare teorije. Če jih lahko iden-

2 Tu gre za mnenje znanosti zahodne družbe. V diskusijo, ki jo sproži Feyerabendov "anything goes", ki enakovredno vzpostavlja različne pojmovne okvirje, se tu ne bomo spuščali.

3 Za navedbo tega članka bom uporabljala kratico RSG.

tificiramo ali reduciramo na postavke nove teorije, potem smo priča teoretsko konzervativni spremembi, če pa so postavke nove teorije tako drugačne, da to ni mogoče, lahko z veliko verjetnostjo sklepamo, da je teoretska sprememba ontološko radikalna in ima za posledico eliminacijo postavk stare teorije.[4]

Klasični simbolni modeli so predstavljali močno podporo zdravorazumski psihologiji. Kako pa je s konekcionizmom? Ali je novi pristop h kognitivnemu modeliranju v skladu z zdravorazumskim okvirjem, ali pa so njegove postavke nezdružljive s prepričanji, željami in drugimi postavkami zdravorazumske psihologije? Ramsey, Stich in Garon v svojem članku "Konekcionizem, eliminativizem in prihodnost zdravorazumske psihologije" zagovarjajo drugo stališče. Njihov argument je izražen s pogojnikom: če se izkaže, da so konekcionistični modeli ustrezna karakterizacija naših kognitivnih procesov, potem v kognitivni znanosti v prihodnje ni prostora za zdravorazumsko psihologijo.

Za intencionalnega realista Fodorjevega tipa tako sklepanje ne predstavlja nevarnosti, ampak ga le še utrjuje v njegovem prepričanju. Iz $P \supset Q$ (če konekcionizem, potem intencionalni eliminativizem) in P (ustreznost konekcionizma) sklepamo po modus ponens na Q (intencionalni eliminativizem). Če pa smo intencionalni realisti in prepričani klasicisti, lahko uporabimo modus tollens in iz $P \supset Q$ in ne-Q sklepamo na ne-P in tako pokažemo, da konekcionistični modeli niso ustrezni kognitivni modeli.

Kaj pa intencionalni realist, ki se hkrati navdušuje nad konekcionizmom? Ali ob konekcionizmu zdravorazumsko psihologijo res čaka tako črna prihodnost, kot ji jo napovedujejo Ramsey, Stich in Garon? V nadaljevanju bom skušala pokazati, da konekcionistični modeli ne predstavljajo ontološko radikalne spremembe in da konekcionizem ni nezdružljiv z zdravorazumsko psihologijo.

## Zdravorazumska psihologija in propozicionalna modularnost

V karakterizaciji zdravorazumske psihologije Ramsey, Stich in Garon kot ključno navedejo načelo *propozicionalne modularnosti*, ki je sestavljeno iz trditev, "da so propozicionalne naravnanosti *funkcionalno diskretna, semantično interpretabilna* stanja, ki igrajo *vzročno vlogo* v ustvarjanju drugih propozicionalnih naravnanosti in nazadnje v ustvarjanju vedenja" (RSG, str. 204).

O tem, da so propozicionalne naravnanosti funkcionalno diskretna stanja, lahko govorimo, kadar je smiselna trditev, da je oseba pridobila ali izgubila posamezno prepričanje ali spomin. (Na primer, ko se Maja zjutraj zbudi, popolnoma pozabi, kam je spravila spričevalo, čeprav ni pozabila ničesar drugega.) Propozicionalne naravnanosti so semantično interpretabilne, če dopuščajo posplošitve zdravorazumske psihologije, ki se naslanjajo na semantične lastnosti prepričanj. Pri tem so predikati, ki izražajo te semantične lastnosti (npr. verjame, *da bo imel vlak zamudo*), taki, da jih lahko projeciramo v prihodnost in jih uporabljamo v nomoloških posplošitvah. (Na primer, vse ljudi, ki so prepričani, da bo imel vlak zamudo, uvrstimo v eno skupino in na podlagi tega pričakujemo, da bodo v njihovem vedenju neke zakonitosti pravilnosti.) In končno, ta funkcionalno diskretna, semantično interpretabilna stanja igrajo vlogo v ustvarjanju vedenja, saj posamezno prepričanje lahko navedemo kot vzrok določenega dejanja.

---

[4] Odnos med kvantno mehaniko in klasično mehaniko (odnos med novo, bolj splošno in bolj natančno teorijo ter staro teorijo) RSG uvrščajo med ontološko konzervativne teoretske spremembe. Analogija med odnosom kvantne mehanike in klasične mehanike ter odnosom konekcionizma do klasičnih kognitivnih modelov (Smolensky, 1988) zato po njihovem mnenju že vnaprej rešuje pomembno vprašanje, ki ga skušajo rešiti sami v tem članku.

Ramsey, Stich in Garon podkrepijo zgornje trditve, da zdravorazumska psihologija jemlje propozicionalne naravnanosti kot funkcionalno diskretna, vzročno aktivna stanja, z dvema primeroma.

Prvi primer se nanaša na Alice, ki bi rada poslala e-mail in je prepričana, da to lahko naredi iz svoje pisarne. Prav tako si želi govoriti s svojo asistentko, za katero je tudi prepričana, da bo v pisarni. V takem primeru zdravorazumska psihologija dopušča, da je vzrok Alicinemu odhodu v pisarno prva želja/prepričanje (poslati e-mail), druga želja/prepričanje (pogovor z asistentko), ali pa oboje. Tako je povsem mogoče, da v tem konkretnem primeru Alicina želja po pošiljanju e-maila ni igrala nobene vloge v njenem vedenju in da je odšla v pisarno le zaradi pogovora z asistentko. Vendarle bi se lahko zgodilo, da bi vseeno odšla v pisarno zaradi želje poslati e-mail, čeprav si ne bi želela govoriti z asistentko. Želja po pošiljanju e-maila, ki je bila sicer vzročno nedejavna pri dejanski odločitvi, bi lahko postala vzročno dejavna. Zdravorazumska psihologija je tako pripravljena prepoznati par različnih semantično karakteriziranih stanj, od katerih je eno vzročno dejavno, drugo pa ne (RSG, str. 205).

Drugi primer je podoben prvemu, le da se namesto na želje in dejanja nanaša na prepričanja in sklepanja. Predpostavimo, da inšpektor Clouseau verjame služabniku, da je preživel večer v vaškem hotelu in da se je vrnil domov z jutranjim vlakom. Predpostavimo, da Clouseau prav tako verjame, da je vaški hotel v tem letnem času zaprt in da jutranji vlak ne vozi več. Na podlagi teh prepričanj (skupaj s splošno sprejetim poznavanjem ozadja) bi inšpektor Clouseau lahko sklepal, da je služabnik lagal. Če je to dejansko storil, je svoj sklep lahko utemeljil ali na prepričanjih o hotelu, ali na prepričanjih o vlaku ali pa na obojem. Z gledišča zdravorazumske psihologije je povsem mogoče, da čeprav je inšpektor Clouseau vedel, da je hotel v tem letnem času zaprt, to prepričanje v tem konkretnem primeru ni igralo nobene vloge v njegovem sklepanju. "Zopet lahko vidimo, da se zdravorazumska psihologija sklicuje na par različnih propozicionalnih naravnanosti, od katerih je v dani situaciji ena vzročno dejavna, druga pa ne" (RSG, str. 206).

Ramsey, Stich in Garon menijo, da je v psihologiji mnogo modelov, ki sledijo zdravorazumski psihologiji in so v skladu z zahtevo po propozicionalni modularnosti (npr. modeli klasične umetne inteligence, kot so produkcijski sistemi ali semantične mreže). Drugače pa je s konekcionističnimi modeli s porazdeljenimi reprezentacijami, kjer je po njihovem mnenju velika neskladnost med konekcionističnimi reprezentacijami in predpostavko propozicionalne modularnosti.

**Od konekcionizma k eliminativizmu**

Da bi čim bolj nazorno predstavili svoj eliminativistični argument, Ramsey, Stich in Garon predstavijo konekcionistični model (Mreža A). Njena naloga je oceniti resničnost (neresničnost) 16 propozicij, ki so kot vhodni podatki predstavljene mreži (npr. psi imajo dlako). Mreža A je tronivojska mreža, ki je sestavljena iz 16 vhodnih, 4 skritih in 1 izhodne enote. Izhodni podatek, ki je blizu 1, pomeni, da je propozicija resnična, če pa je blizu 0, je propozicija napačna. Mreža se je naučila naloge s pomočjo učnega algoritma, ki temelji na posplošenem delta pravilu (back-propagation). Če naučeni mreži predstavimo propozicijo kot vhodni podatek (vektor vrednosti za vhodne enote), bo mreža prek uteži, ki si jih je pridobila med učenjem, posredovala ustrezen odgovor na izhodni enoti. Naučene uteži povezujejo vseh 16 vhodnih enot s štirimi enotami skritega nivoja in z eno izhodno enoto. Pri tem je pomembno, da so s pomočjo učnega algoritma vrednosti vseh uteži tako prilagojene, da mreža ne deluje samo za eno

propozicijo, ampak za vseh 16. Znanje sistema je tako shranjeno superpozicijsko v množici uteži.

Taka ugotovitev po mnenju Ramseya, Sticha in Garona že vodi v neskladnost z zdravorazumsko psihologijo. V zdravorazumski psihologiji velja, da je prepričanje, da imajo psi dlako, tisto, ki povzroči odgovor na vprašanje "Ali imajo psi dlako?". Če pa je spomin tako organiziran, kot ga kaže opisani konekcionistični model, potem ni jasno, ali je smiselno govoriti o določeni informaciji kot o povzročitelju izhodnih podatkov, ali pa je to kar celota informacij. "Informacija, kodirana v Mreži A, je shranjena holistično in porazdeljeno skozi mrežo. Kadarkoli dobimo informacijo iz Mreže A, ... *mnogo* vezi in *mnogo* skritih enot sodeluje v preračunavanju. Vsaka določena utež ali enota bo pripomogla k kodiranju informacije o *mnogih* različnih propozicijah. Preprosto se nima pomena spraševati, ali reprezentacija določene propozicije igra vzročno vlogo v komputaciji mreže ali ne" (RSG, str. 212). Lastnosti konekcionističnega modela s porazdeljenimi reprezentacijami, kjer je znanje shranjeno v množici uteži (superpozicija), so zato na prvi pogled v globokem neskladju z zahtevo propozicionalne modularnosti zdravorazumske psihologije. Ta je predpostavljala, da v splošnem obstaja odgovor na vprašanje, ali določeno prepričanje ali spomin igra vzročno vlogo. Po mnenju Ramseya, Sticha in Garona ne moremo več trditi, da vedenje mreže, da imajo psi dlako, povzroči odgovor "da". Prav toliko ga povzroči tudi na primer njeno vedenje, da imajo mačke dlako ali katerakoli od ostalih 16 propozicij, saj so vse shranjene v isti množici uteži.

Nekompatibilnost med propozicionalno modularnostjo in konekcionizmom so Ramsey, Stich in Garon ponazorili še z novim modelom, Mrežo B, ki je taka kot Mreža A, le da med učne primere dodamo še eno propozicijo (torej 17). Ko se bo Mreža B naučila naloge (odgovoriti z da/ne na 17 propozicij), bodo vrednosti uteži, gledano v celoti, različne od uteži v Mreži A. Zaradi načina shranjevanja v modelih s porazdeljenimi reprezentacijami (superpozicija) je namreč način kodiranja propozicije odvisen od ostalega znanja, ki je shranjeno v mreži. Posledica tega je, da mreža s 17 propozicijami shrani le-te nekoliko drugače (vrednosti uteži so različne) kot mreža s 16 propozicijami, čeprav je teh 16 propozicij podmnožica 17 propozicij. Informacija, ki se tiče katerekoli propozicije, namreč ni shranjena na enem mestu, ampak je razpršena po mreži. Za razliko od konekcionističnega modela pa v klasičnem simbolnem modelu ne bi bilo težko ugotoviti, katera stanja sistema kodirajo dodatno propozicijo in ali reprezentacija nove propozicije igra vlogo v konkretni kognitivni nalogi, ki jo sistem modelira.[5] Zato Ramsey, Stich in Garon sklepajo, da konekcionističnemu modelu manjka podstruktura, ki bi bila funkcionalno različna, ki bi jo lahko interpretirali kot reprezentacijo posamezne propozicije, in bi ustrezala psihološki naravni vrsti. "Nauk je, da, čeprav obstaja neomejeno število konekcionističnih mrež, ki tako kot Mreža A predstavljajo informacijo, da imajo psi dlako, te mreže nimajo skupnih projektivnih moči, opisljivih v jeziku konekcionistične teorije. ... razred mrež, ki bi modelirale kognitivnega predstavnika, ki je prepričan, da imajo psi dlako, ni naravna vrsta, ampak le kaotična disjunktivna množica. Zdravorazumska psihologija obravnava razred ljudi, ki so prepričani, da imajo psi dlako, kot psihološko naravno vrsto; konekcionistična psihologija pa ne" (RSG, str. 213).

---

5 V klasičnem simbolnem modelu, kjer je npr. seznam 2 enak seznam 1 + nov stavek, bi se v seznamu 2 ohranil seznam 1 kot podmnožica:

| seznam 1 | seznam 2 |
|---|---|
| psi imajo dlako | psi imajo dlako |
| mačke imajo tace | mačke imajo tace |
| | mačke imajo dlako |

Dva primera konekcionističnih mrež naj bi torej pokazala, da konekcionistični modeli, ki uporabljajo porazdeljene reprezentacije, nimajo funkcionalno diskretnih, semantično interpretabilnih podstruktur, ki bi ustrezale intencionalnim entitetam zdravorazumske psihologije. Menim, da se sklepanje Ramseya, Sticha in Garona opira samo ne eno raven analize konekcionističnega modeliranja (raven posameznih procesnih enot), zanemarja pa višjo raven aktivacijskih vzorcev.[6] Pogled s te višje ravni v "kaotični množici" odkriva več strukture, kot so jo konekcionističnim reprezentacijam pripisali Ramsey, Stich in Garon, in ne podpira njihovega sklepa o ontološko radikalni teoretski spremembi.

**Višje ravni analize v konekcionističnih modelih**

Poleg ravni posameznih procesnih enot in uteži povezav obstajajo tudi lastnosti, ki se tičejo celotne mreže. Ena izmed značilnosti konekcionističnega modela s porazdeljenimi reprezentacijami so vzorci aktivacij skritih enot, ki nastanejo v toku procesiranja in so posledica učenja mreže. Kadarkoli konekcionistični mreži podamo vhodne podatke, ki jih potem procesira do izhodnih enot, v mreži opazimo odgovarjajoči vzorec aktivacij procesnih enot na skritem nivoju. Ti vzorci niso nekakšna nedejavna lastnost mreže, ampak vzročno učinkujejo na procesno aktivnost v mreži. Predstavljajo rešitev sistema na zastavljen problem oziroma komputacijo, ki jo mreža kot celota izvaja kot odgovor na vhodne podatke in vodi do novih vzročnih aktivnosti, npr. kot vhodni podatki za novo mrežo ali kot navodila za mehanizme (motorne, optične, itd.), ki so priključeni na mrežo.

Taki vzorci aktivacij predstavljajo funkcionalno diskretne podstrukture, ki igrajo vzročno vlogo v procesiranju konekcionističnega modela in so lahko deležne semantične interpretacije. (V konekcionističnih kognitivnih modelih s porazdeljenimi reprezentacijami je to običajna praksa). Toda, ali so to podstrukture, ki bi lahko uprimerile intencionalne entitete zdravorazumske psihologije?

Ramsey, Stich in Garon menijo, da pozitiven odgovor ni mogoč, saj zdravorazumska psihologija pojmuje prepričanja in spomine kot trajna stanja (vsaj v veliki meri). Poleg tega ima človek lahko v splošnem veliko prepričanj in spominov, tudi kadar jih ne uporablja. Vzorec aktivacij procesnih enot v mreži pa ni trajno stanje in je prisoten le, kadar mreža sprejme ustrezne vhodne podatke. Prav tako v mreži ni možnih več vzorcev istočasno, ampak je lahko prisoten le eden (RSG, str. 216).

Negativni odgovor na zastavljeno vprašanje je prepričljiv, vendar konekcionist še ni izčrpal vseh možnosti. Ko smo govorili o konekcionističnih modelih, smo poudarili, da je znanje mreže shranjeno v utežeh povezav med enotami. Mreža je s pomočjo učnega algoritma skozi proces učenja popravljala vrednosti uteži tako, da je po končanem učenju zmožna obvladovati nalogo. Matrika povezav, ki je sestavljena iz aktivacijskih vrednosti enot in uteži povezav, določa, kako se bo mreža odzivala na dane vhodne podatke oziroma kakšen vzorec aktivacijskih vrednosti skritih enot bo v mreži nastal. Ta matrika povezav predstavlja dispozicijo mreže, da ob določenih vhodnih podatkih vodi k določenemu aktivacijskemu vzorcu, ki potem povzroči ustrezen izhod iz mreže.

Kot kandidat za uprimerjanje prepričanj in drugih intencionalnih entitet zdravorazumske psihologije se tako javlja matrika povezav, dispozicijsko stanje mreže, da v

6 Podobno stališče zastopajo tudi O'Brien (1991), Clark (1993) in Foster & Saidel (1994), na katerih dela se v svojem izvajanju tudi opiram.

danih okoliščinah proizvede določen vzorec aktivacij. Na ta način je rešena težava s trajnostjo, saj so dispozicijska stanja trajna in neodvisna od trenutnega vzorca aktivacij v enotah, prav tako pa zaradi superpozicijskega načina shranjevanja matrika povezav lahko kodira več informacij naenkrat (tj. predstavlja istočasno več prepričanj). Vendar tudi ta karakterizacija ni brez težav, saj je treba pokazati, da dispozicije, ki ustrezajo različnim prepričanjem, lahko štejemo za diskretna, neodvisno vzročno aktivna stanja, kot jih zahteva zdravorazumska psihologija in kot sta pokazala primera z Alice in inšpektorjem Clauseaujem.

Če sprejmemo, da morajo imeti trenutno aktivna prepričanja funkcionalno diskretno realizacijo in da morajo biti prepričanja dalj časa trajajoča stanja, potem je rešitev v kombinaciji matrike povezav (nekakšna implicitna reprezentacija), ki predstavlja prepričanje kot dolgo trajajoče stanje, in vzorca aktivacij (eksplicitna reprezentacija) kot trenutno aktivnega prepričanja. Na tak način konekcionist sicer ne more zadovoljiti intencionalnega realista, ki ozko povezuje zdravorazumsko psihologijo z računalniško metaforo in klasičnimi simbolnimi modeli[7], je pa dovolj, da zavrne prvi eliminativistični ugovor, ki so ga Ramsey, Stich in Garon podali s primerom Mreže A.

Ostane še drugi ugovor, primer Mreže B, ki meri na nezmožnost konekcionistične psihologije, da identificira psihološke naravne vrste. V tem primeru imamo težave, ker sta tako matrika povezav kot vzorec aktivacij, ki naj bi ustrezala prepričanju, da imajo psi dlako, vezana na konkretno mrežo, v kateri sta nastala. V drugih mrežah, ki naj bi prav tako kodirale prepričanje, da imajo psi dlako, to ustreza drugačni matriki povezav in drugačnemu vzorcu aktivacij. Tako se zdi upravičen ugovor, da te mreže nimajo nekih skupnih atributov, ki bi omogočali, da bi vse mreže, ki modelirajo prepričanje, da imajo psi dlako, uvrščali v naravno vrsto.

Naloga, ki se postavlja pred konekcionista, je torej poiskati višje lastnosti mreže, ki bi omogočale identifikacijo takih skupnih atributov v različnih mrežah. Znanstveniki, ki raziskujejo konekcionistične modele, se pogosto poslužujejo statistične analize vzorcev aktivacij skritih enot pri interpretaciji svojega modela. S tem dobijo sliko načina, kako se je sistem naučil razdeliti kognitivni prostor pri dani nalogi. Najpogosteje omenjan primer take analize, hierarhične analize skupkov (cluster analysis), je analiza mreže NETtalk Sejnowskega in Rosenberga (Rosenberg and Seynovski, 1987). Zanimiva je tudi uporaba metode glavnih osi (principal component analysis), ki jo je pri analizi svoje mreže uporabil Elman (Elman, 1991). Gre za podobno statistično metodo, ki pa skuša poleg statičnih lastnosti (podobnost med reprezentacijama) zajeti še časovno dinamiko mreže.

Na kratko si poglejmo analizo mreže NETtalk. Gre za konekcionistični model, ki naj bi se naučil spreminjati pisno besedilo v foneme (kar bi ob priključitvi na sintetizator zvoka omogočilo slušni izhod). Mreža je precej velika, saj je sestavljena iz 203 vhodnih enot, 80 skritih enot in 26 izhodnih enot in uporablja posplošeno delta pravilo kot učni algoritem. Na izbranih učnih primerih (nekaj različnih besedil, za katera so podani pravilni izhodni podatki - fonemi) se mreža ob velikem številu ponavljanj precej dobro nauči zadane naloge, ne da bi jo seznanili s pravili spreminjanja teksta v foneme. Toda če bi ostalo samo pri tem, to ne bi bilo posebej zanimivo za kognitivno znanost. Imeli bi sistem, ki uspešno izvaja nalogo, hkrati pa nam ne bi bilo jasno, kako jo opravlja, manjkala bi razlaga. (Pri klasičnem pristopu s tem ne bi imeli težav, saj znanstvenik najprej opravi analizo naloge in potem na podlagi komputacijske teorije

[7] Tako je Fodorjeva reprezentacijska teorija duha (Fodor, 1987) zavezana eksplicitnim reprezentacijam vsebine misli in intencionalne entitete ne morejo biti vzročno dejavne, če jih ne moremo identificirati z eksplicitnimi in s tem funkcionalno diskretnimi reprezentacijami v jeziku misli.

konstruira algoritem, ki pripelje do delujočega programa.) Zato je treba konekcionistične modele, ki do znanja pridejo na podlagi učenja iz množice učnih primerov, podvreči analizi, ki sledi obdobju učenja. Analizirati mora že delujoč model in v njem poiskati elemente za razlago.[8] V primeru analize skupkov pri NETtalku raziskovalec za različne vhodne podatke posname aktivacije skritih enot in izhodne podatke, ki so jih skrite enote povzročile. Nato vzame vse tiste vhodne podatke, ki so vodili do enakega fonema, in poišče povprečni vektor aktivacij skritih enot. To naredi za vse foneme (79 skupkov). S hierarhično analizo skupkov nato primerja najbolj podobne vektorje (podobnost se meri z razdaljo med vektorji), poišče povprečje za tak par in nato ponavlja postopek, dokler ne dobi zadnjega para. Rezultat takega razvrščanja je dendogram, drevesna struktura, ki prikazuje hierarhično razdelitev vektorskega prostora notranjih enot. Iz njega lahko razberemo, da je največja razlika med soglasniki in samoglasniki, najmanjša pa med b in p.

Za izpodbijanje drugega eliminativističnega argumenta Ramseya, Sticha in Garona so relevantne predvsem ugotovitve, vezane na mreže, ki so imele različne začetne vrednosti uteži. Ker je v matriki uteži shranjeno znanje mreže (četudi nikakršno, kot na začetku učenja nove naloge), začetne vrednosti uteži vplivajo na to, kakšne bodo vrednosti uteži po končanem učenju. Mreže, ki so jih učili z istimi učnimi primeri, a z različnimi začetnimi vrednostmi uteži, imajo zato različne opise na ravni enot in uteži. Toda če pri vseh mrežah naredimo še hierarhično analizo skupkov, nam ta pokaže, da imajo vse te mreže enako drevesno strukturo skupkov. Z višjo ravnijo opisa je tako mogoče zajeti skupne lastnosti mrež, ki jih na nižji ravni opisa ne bi mogli odkriti. Pomembno pri tem je, da se v skupke uvrščajo vzorci aktivacij skritih enot glede na to, katere izhodne podatke povzročajo.

Podobno bi lahko trdili v primeru Mreže A in Mreže B (oziroma bolj kompleksnih mrež z več izhodnimi enotami), če bi lahko naredili tako post hoc analizo, kjer bi imeli semantične entitete za oznake skupkov. Ko bi mreži posredovali določene vhodne podatke, bi ti potovali do skritih enot, kjer bi oblikovali vzorec aktivacij. Ta vzorec aktivacij bi padel v skupek, ki bi ga označili s "psi imajo dlako". Potem bi lahko rekli, da smo dobili določene izhodne podatke, ker je bila mreža v tistem trenutku prepričana, da imajo psi dlako.

**Konekcionizem in zdravorazumska psihologija nista nezdružljiva**

Ramsey, Stich in Garon so zdravorazumsko psihologijo tesno povezali z jezikom misli. Ker v konekcionističnih modelih niso našli reprezentacij (stavkov jezika misli) kot objektov preračunavanja, so sklepali na eliminacijo zdravorazumskih entitet. Vendar samo dejstvo, da oblika kodiranja ni taka kot pri jeziku misli, še ni razlog, da ne bi mogli reprezentacijskih stanj sistema ustrezno zajeti s stavki. Zdravorazumska psihologija propozicionalnih naravnanosti ni vezana na komputacijsko zgodbo notranjega kodiranja ampak na vsebino notranjih reprezentacij. Z analizo konekcionističnega modeliranja smo prišli do sklepa, da pri konekcionističnih modelih s porazdeljenimi reprezentacijami, ki jih lahko opišemo na višji ravni in ki imajo tako arhitekturo, da omogočajo povratne zanke[9], trditev o ontološko radikalni teoretski spremembi ne velja. V

[8] Clark tak konekcionistični pristop k razlagi imenuje kopernikansko revolucijo v kognitivni znanosti. "Konekcionist učinkovito obrne časovni in metodološki potek razlage, tako kot je Kopernik obrnil običajni astronomski model svojega časa" (Clark, 1993, str. 50).

[9] To je potrebno zato, ker prepričanja lahko povzročajo tako dejanja, kot tudi druga prepričanja. Model sistema prepričanj mora omogočati, da prepričanja igrajo obe vlogi.

takih modelih prepričanju ustreza dispozicija mreže, da ob določenih vhodnih podatkih "pade" v določen skupek, ki predstavlja trenutno prepričanje. Čeprav je pri tem natančna dinamika mreže podana na ravni enot in uteži, to ne vodi v eliminacijo intencionalnih entitet zdravorazumske psihologije. Moč zdravorazumske psihologije je v posplošitvah in napovedih vedenja v običajnih situacijah, naloga konekcionističnega modeliranja pa je v odkrivanju mehanizmov, ki take napovedi omogočajo in obenem tudi razlagajo pojave, za katere zdravorazumska psihologija ni našla (ali pa niti ni iskala) razlag (npr. nižji spoznavni procesi, nenormalno vedenje ob poškodbah možgan, itd.)

Konekcionist torej kljub eliminaciji notranjega jezika misli lahko ostaja intencionalni realist. Če pa želi zagovarjati intencionalni eliminativizem in nevrološki konekcionizem, bo moral za to poseči po drugih argumentih, tako kot sta to storila Feyerabend, ki je zagovarjal povsem fiziološki pristop (Feyerabend, 1963), in Quine, ki je trdil, da moramo znanost očistiti intencionalnih izrazov, in je zagovarjal behavioristično analizo vedenja (Quine, 1960).[10]

## LITERATURA


Churchland, P.M. (1988), *Matter and Consciouness*, MIT Press, Cambridge, MA.

Clark, A. (1993), *Associative Engines*, MIT Press, Cambridge, MA.

Dobnikar, A. (1990), *Nevronske mreže*, Didakta, Radovljica.

Forster, M. Saidel, E. (1994), "Connectionism and the fate of folk psychology: a reply to Ramsey, Stich and Garon", Philosophical Psychology, *vol. 7, no. 4, 437-461.*

Elman, J.L. (1991), "Distributed representation, simple recurrent networks and grammatical structure", *Machine learning 7, 195-225.*

Fodor, J. A. (1987), *Psychosemantics*, MIT Press, Cambridge, MA.

Fodor, J.A. and Pylyshyn, Z.W. (1988), "Connectionism and Cognitive Architecture: A Critical Analysis", *Cognition* 28, 3-71.

Feyerabend, P. (1963/1991), "Mental Events and the Brain", The Journal of Philosophy LX, 11, 295-296, ponatisnjeno v D. Rosenthal (ured.), *The Nature of Mind*, Oxford University Press, Oxford, 266-267.

Markič, O. (1992), "Konekcionizem in zdravorazumska psihologija", *Slovenski filozofski zvezki*, DAF, Ljubljana.

O'Brien, G.J., (1991), "Is Connectionism commonsense?", *Philosophical Psychology, zv. 4, št. 2, 165-178.*

Potrč, M. (1992), "Modeli duha", *Slovenski filozofski zvezki*, DAF, Ljubljana.

Quine, W.V. (1960), *Word and Object*, MIT Press, Ma.

Ramsey, W., Stich, S.P. and Garon, J. (1991), "Connectionism, Eliminativism, and the Future of Folk Psychology", v W. Ramsey, S. Stich and D. Rumelhart (ured.) *Philosophy and Connectionist Theory*, Erlbaum, London.

Rosenberg, C., and Seynovski, J. (1987), "Parallel networks that learn to pronounce English text", *Complex Systems 1, 145-168.*

Smolensky, P. (1988), "On the Proper Treatment of Connectionism", *Behavioral and Brain Sciences* 11, 1-74.

Stich, S. (1983), *From Folk Psychology to Cognitive Science*, MIT Press, Cambridge MA.


[10] Članek je dopolnjeno in nekoliko predelano poglavje magistrskega dela z naslovom "Kognitivni modeli" (mentor prof. dr. Matjaž Potrč), ki je nastalo v okviru projekta usposabljanja mladih raziskovalcev Ministrstva za znanost in tehnologijo na Oddelku za filozofijo Filozofske fakultete.