THEORIA 2
(2020) : 63 : str. 135-164

https://doi.org/10.2298/THEO2002135M
Originalni naučni rad
Original Scientific Paper

Miljana Milojević
Vanja Subotić[1]

EKSPLORATIVNI STATUS (POST)KONEKCIONISTIČKIH MODELA[2]

APSTRAKT: *Cilj ovog rada je da pruži novo viđenje uloge konekcionističkih modela u istraživanju ljudske kognicije kroz konceptualizaciju istorije konekcionizma – od najjednostavnijih perceptrona do savremenih dubokih konvolucionih neuronskih mreža, kao i kritika poteklih iz domena rivalske simboličke kognitivne nauke. Naime, konekcionistički pristup u kognitivnoj nauci bio je meta oštrih kritika simbolista koje su u više navrata uzrokovale njegovu marginalizaciju i skoro potpuno napuštanje njegovih pretpostavki u izučavanju kognicije. Kritike su uglavnom ukazivale na njegovu eksplanatornu neadekvatnost kao teorije kognicije ili na njegovu biološku neplauzibilnost kao teorije implemetacije, a od konkretnih nedostataka nekih konekcionističkih modela napredovalo se do zaključaka o nedostacima konekcionizma uopšte. U ovom radu želimo da pokažemo da obe vrste kritike počivaju na pretpostavci da su jedina prava objašnjenja u kognitivnoj nauci instance homunkularnog funkcionalizma i da uklanjajući ovu pretpostavku i usvajajući alternativnu metodologiju – eksplorativno mehanicističku strategiju, možemo da uklonimo i većinu prigovora konekcionizmu kao irelevantne, da objasnimo napredak konekcionističkih modela uprkos njihovim nedostacima i da skiciramo putanju njihovog budućeg razvoja. Usvajanjem mehanicizma i kritikom funkcionalizma odbacićemo prigovore eksplanatorne neadekvatnosti, karakterisanjem konekcionističkih modela kao skica generičkih mehanizama odbacićemo prigovore biološke neplauzibilnosti, dok ćemo pripisivanjem eksplorativnog karaktera takvim modelima pokazati manjkavost prakse generalizovanja od trenutnih ka opštim neuspesima konekcionizma.*

KLJUČNE REČI: *duboko učenje*, *eksploracija*, *konekcionizam*, *mehanicistička objašnjenja*, *tradicionalna simbolička kognitivna nauka*.

---

1 Imena autora su navedena prema abecednom redu, a ne prema doprinosu.

2 Ovaj rad je nastao u okviru projekta „Dinamički sistemi u prirodi i društvu: filozofski i empirijski aspekti“, evidencioni broj 179041, koji finansira Ministarstvo prosvete, nauke i tehnološkog razvoja Republike Srbije.

Uvod

Veštačke neuronske mreže danas imaju izuzetno široku primenu, jer poseduju sposobnost učenja i modelovanja nelinearnih procesa – što ih čini izuzetno korisnim oruđem za rešavanje najrazličitijih zadataka poput prepoznavanja obrazaca, klasifikacije, klasterovanja, kompjuterske vizije i mnogih drugih. U tom smislu, njihova uloga kao veoma potentnih računskih sistema, bez kojih ne bismo mogli da zamislimo pametne telefone ili autonomne automobile, opšte je priznata i ne može se dovesti u pitanje. Ipak, iako su nastale prema ugledu na funkcionisanje neurona i njihovih sklopova u okviru mozga, status neuronskih mreža u oblasti kognitivne nauke, te pitanje njihove eksplanatorne moći povodom ljudske kognicije, ostaje predmet rasprave do danas.

Naime, pojava konekcionizma u kognitivnoj nauci najčešće je konceptualizovana kao pojava alternative tradicionalnom kompjutaciono-reprezentacionom pristupu kogniciji. Prema takvom stanovištu, konekcionistički modeli treba da postuliranjem odgovarajućeg tipa mrežnih mehanizama i kognitivne arhitekture objasne sve one fenomene na čije su objašnjenje pretendovali i tradicionalni modeli – poput korišćenja i razumevanja jezika, kategorizacije, zaključivanja, itd. – ali na biološki plauzibilniji način i na finijem nivou opisa koji je ujedno i eksplanatorno bogatiji. Međutim, istorija razvoja konekcionističkih modela prikazuje upravo niz osporavanja obe ove ideje – ideje da konekcionizam potencijalno može pružiti objašnjenje svih kognitivno i psihološki relevantnih fenomena, kao i ideje da se njime pruža biološki plauzibilno shvatanje kognitivnih mehanizama. Zapravo, istorija razvoja konekcionizma može se uprošćeno prikazati sledećom skicom etapa:

(1) predlaganje neke konkretne konekcionističke arhitekture i konkretnih modela;
(2) ukazivanje na činjenicu da takva *konkretna* konekcionistička arhitektura *ne može* da pruži objašnjenje za kognitivni fenomen *Φ*, odnosno da struktura postuliranih mehanizama ne objašnjava ili čak onemogućava proizvođenje fenomena *Φ*;
(3a) pokušaj konekcionista da argumentuju zašto konekcionizam *ne mora* da objasni fenomen *Φ*, pri čemu se obično pribegavalo odlasku na niži nivo opisa ili tvrđenjima da konekcionizam pruža objašnjenje *implementacionih* struktura, a ne samih kognitivnih fenomena, ili
(3b) izmena *strukturnih* obeležja neuronskih mreža tako da *Φ* sada može da se objasni.

Na ovakve odgovore konekcionista najčešće je potom odgovarano da:

(4a) konekcionizam ne pruža dobra objašnjenja implementacije, jer mehanizmi koje postulira nisu biološki plauzibilni, ili
(4b) postoji novi fenomen tipa *Φ'* koji ni novi modeli ne mogu da objasne.

Takvo apsolutističko shvatanje uloge konekcionizma u kognitivnoj nauci – da on mora pružiti ili dobru podlogu za objašnjenje svih kognitivnih, odnosno psiholoških, fenomena ili biološki plauzibilne mehanizme koji su odgovorni za pojavu takvih fenomena – dovelo je i do nekoliko čuvenih „zima" u razvoju neuronskih mreža, kao i do neodređenog statusa konekcionizma u okvirima kognitivne nauke. U ovom radu želimo da prikažemo jedan drugačiji pogled na ulogu konekcionističkih, odnosno postkonekcionističkih, modela koji pred njih ne stavlja ovako jake zahteve.

Tvrdićemo da se takvim modelima može dodeliti *eksplorativni status*. Model ima eksplorativni status onda kad se koristi u svrhe istraživanja postavki određene teorije, tako što se model modifikuje, parametrizuje ili precizira preispitivanjem toga da li postavke mogu biti drugačije, ili kako će se teorija promeniti kada se neke od postavki skroz izostave (Fisher 2006). Ukoliko konekcionističke i postkonekcionističke modele posmatramo na ovaj način, većinu prigovora koji su im do sada upućeni možemo tretirati kao nekonkluzivne. Dodeljivanje eksplorativnog statusa ovim modelima ostavlja dovoljno prostora za njihov budući *progres* u pogledu kompjutacione moći, koji svaka nova generacija kritičara iznova proglašava *u principu* nemogućim dajući prigovore tipa **(2)**, odnosno **(4b)**.

Eksplorativno shvaćeni modeli, kako će biti pokazano, nas približavaju kako-je-nešto-moguće (eng. *how-possibly*) objašnjenjima u pogledu kognitivnih mehanizama i fenomena, a uz inkorporiranje neurobioloških detalja može se govoriti i o približavanju kako-je-nešto-plauzibilno (eng. *how-plausibly*) objašnjenjima, pa se na taj način mogu razoružati i prigovori tipa **(4a)**. Drugim rečima, mi želimo da konekcionizam i postkonekcionizam tretiramo kao jedan moćan program koji ima potencijal da pruži kognitivno i neuralno plauzibilna objašnjenja, iako su se dosadašnje verzije suočavala sa eksplanatornim teškoćama u oba ova pogleda. Što bude intenzivnija saradnja između kognitivnih i neuronaučnika sa istraživačima na polju veštačke inteligencije, moći ćemo pre da se približimo kako-je-nešto-aktualno objašnjenima (eng. *how-actually*) (o napredovanju od kako-je-što-moguće do kako-je-nešto-aktualno objašnjenja v. Craver 2007: 114).

Prvi deo rada biće posvećen kratkom istorijatu nastanka koncepta neuronskih mreža, u drugom ćemo se baviti prvim sukobom standardnih i konekcionističkih modela, u trećem ćemo posvetiti pažnju trenutnom sukobu tradicionalista i postkonekcionista, dok ćemo u četvrtom dati naš predlog koji nastale sukobe, kao i argumente i protivargumente sa obe strane, ne vidi kao „sve ili ništa" stvar, već kao uvid u istoriju živog i progresivnog programa čije plodove tek možemo da očekujemo.

1. Sukob pre sukoba – neuronske mreže i problem izračunljivosti

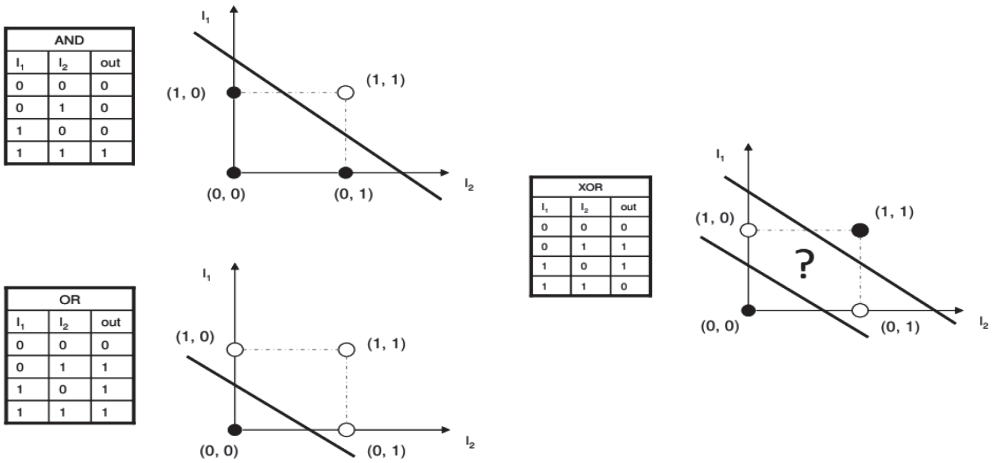
Napredak ka današnjim neuronskim mrežama počinje sa matematičkim modelom neuronske aktivnosti Mekaloha i Pitsa (McCullogh, Pitts) objavljenom u radu „A Logical Calculus of Ideas Immanent in Nervous Activity" (1943). Inspirisani neurofiziološkim nalazima Santjaga Ramona i Kahala i Ser Čarlsa Skota Šeringtona (Santiago Ramon y Cahal, Sir Charles Scott Sherrington) prema kojima biološki neuroni primaju stimuluse na svojim dendritima, a nakon procesiranja signala i postignutog aktivacionog praga transmituju izlazni signal preko aksona, Mekaloh i Pits konstruišu model prema kojem neuroni sa binarnim pragom aktivacije mogu da implementiraju funkcije koje odgovaraju onima logike prvog reda. Dalje, oni konstruišu matematičku metodu za dizajniranje neuronskih mreža, koje bi reprezentovale odgovarajuće inferencijalne relacije između iskaza, odnosno pružaju model prema kojem su određene kompjutacije matematički ekvivalentne odgovarajućim logičkim operacijama (Piccinini 2004: 203). Posebno interesantna posledica njihove tehnike bio je kompjutacioni dizajn Bulovih funkcija, a koji je Fon Nojman (1945) inkorporirao u svoju arhitekturu digitalnog računara. Tako, neuron može da reprezentuje Bulovu „i" funkciju za stimulus koji ima dve promenljive, *x* i *y*, ili dve input jedinice, a čije vrednosti mogu da budu 1 i 0, tako što će se prag aktivacije baždariti na 2. Ukoliko prag nije postignut autput signal će biti 0, a ukoliko je postignut autput će biti 1. Za sve moguće vrednosti promenljivih *x* i *y*: (0,0), (0,1), (1,0) i (1,1), vrednosti autputa *z* će redom biti 0, 0, 0, 1, što reprezentuje funkciju „i". Funkcija „ili" mogla je da se izračunava postavljanjem praga na 1, a funkcija „ne" postavljanjem samo jedne input jedinice.

Prvo unapređenje modela neuronske aktivnosti i prve neuronske mreže sposobne da uče pojavljuju se krajem pedesetih godina prošlog veka u radu psihologa Frenka Rozenblata (Frank Rosenblatt), koji spaja ideju Mekaloha i Pitsa sa uvidima Dejvida Heba (David Hebb 1949). Naime, Heb je primetio da ukoliko jedna neuralna ćelija često ekscitira drugu dolazi do pojačanja veze između njih i efikasnosti buduće ekscitacije čime se ujedno objašnjava i sposobnost učenja i pamćenja. Rozenblat je matematički formulisao Hebov neurofiziološki postulat, prema kom se proces učenja bazira na simultanoj aktivaciji između jedinica (odnosno, neurona), koja zauzvrat ojačava sinaptičke veze između tih jedinica (odnosno, neurona), i primenio ga kao algoritam za obučavanje dvoslojnih neuronskih mreža, tzv. *perceptrona* (Rosenblatt 1958: 386).[3]

3 Među savremenim autorima postoji neslaganje u pogledu toga koliko slojeva Rozenblatovi perceptroni imaju: Bengio et al. (2015) tvrde da je u pitanju jednoslojna neuronska mreža, Buckner & Garson (2018) tvrde da je dvoslojna. Međutim, interesantno je da Rozenblat opisuje i troslojne, četvoroslojne ili pak višeslojne perceptrone (Rosenblatt 1962: chs. 5, 15 & 16), ali su Minski i Papert, u svojoj knjizi iz 1969. godine, doprineli zabludi formalnim dokazom o ograničenosti učenja dvoslojne proaktivne mreže, u kojoj se signal direktno prenosi od jedinica inputa do jedinica autputa, pa se danas za osnovnu karakteristiku perceptrona uzima odsustvo dodatnih skrivenih slojeva između input i autput sloja.

Dok Hebov postulat konstituiše nenadgledano učenje, perceptron po imenu Mark I je mogao da se obučava i putem nadgledanog učenja, to jest preko metoda pokušaja i pogreške, tako što se sračuna razlika između željenog autputa i aktualnog autputa (u slučaju da mreža na početku ne daje očekivane rezultate), pa se potom „manuelno" podesi stepen aktivacije jedinica, odnosno snaga inputa, i to se ponavlja sve dok mreža za određeni input ne počne da pruža odgovarajući autput (Rosenblatt 1958: 395). Na ovaj način perceptronima je podarena fleksibilnost i mogućnost učenja, koju neuronske mreže Mekaloha i Pitsa nisu posedovale.

Minski (Marvin Minsky) i Rozenblat su debatovali godinama o vrednosti perceptrona, svaki zauzimajući ekstremnu poziciju: dok je Minski tvrdio da perceptroni ne mogu da urade ništa i da je Rozenblatov rad bez naučne vrednosti (Minsky & Papert 1969: 4), Rozenblat je bio ubeđen da perceptroni mogu da se obuče da urade bilo šta, jer nisu morali da se programiraju pomoću eksplicitnih pravila koja bismo morali prethodno da otkrijemo, već su rešavali probleme zahvaljujući obrascima aktivacije do kojih se dolazilo treniranjem koje je operisalo samo nad zadatim inputima i željenim autputima. U koautorskoj knjizi sa Papertom, Minski je ponudio formalni dokaz da dvoslojni perceptron, koji ima samo sloj inputa i autputa, pa posledično i dvoslojne neuronske mreže Mekaloha i Pitsa, ne mogu da procesiraju funkciju ekskluzivne disjunkcije ili XOR. Dokaz je upućivao na činjenicu da dvoslojni perceptron može da računa samo linearne funkcije, to jest funkcije čije različite vrednosti mogu biti razdvojene pravom linijom, a XOR nije linearna funkcija, budući da nema prave koja bi prostor inputa podelila na odgovarajući način.



Slika 1.1.[4]

4 Grafički prikaz je preuzet sa internet adrese https://mc.ai/solving-xor-with-a-single-perceptron/.

Čuvena „I/ili teorema“ (eng. *The And/Or Theorem*), Minskog i Paperta, tumačena je potom tako da će i kompleksnije mreže nailaziti na slična ograničenja iako su Minski i Papert svoje rezultate ograničili na jednostavni dvoslojni perceptron. Njihova knjiga je bila do te mere uticajna da je došlo do marginalizacije istraživanja neuronskih mreža usled poteškoća sa dobijanjem finansija od naučnih fondova Sjedinjenih Američkih Država (Olazaran 1996). Međutim, perceptron-neizračunljivost XOR ne važi za višeslojne neuronske mreže, čega su Minski i Papert bili svesni iako to nisu potencirali. Već sa umetanjem jednog skrivenog sloja jedinica između inputa i autputa, čije bi jedinice ili neuroni bili konektovani sa oba spoljašnja sloja, moguće je postaviti težine jedinica na takav način da se prostor inputa transformiše u linearno deljiv. Ipak, treba takođe imati na umu da u to vreme nije postojao efikasan način za obučavanje mreža sa skrivenim slojevima – što je bio dodatni razlog za značajno slabljenje ovog programa. Teškoće u pogledu učenja višeslojnih neuronskih mreža rešene su tek sa primenom algoritma propagiranja greške unazad (eng. *backpropagation of error*) na neuronske mreže Rumelharta, Hintona i Vilijamsa (1986).

Stoga, nakon Rozenblatove prerane smrti 1971. godine, nastupio je period neprimećenog tehničkog rada na neuronskim mrežama, a u kognitivnoj nauci, koja se aktivno razvija tek od sredine sedamdesetih godina prošlog veka, preovladali su modeli u kojima se kompjutaciono procesiranje vrši sekvencijalno nad simboličkim repezentacijama nalik onome digitalnih računara. Ovo je bio sasvim razumljiv tok razvoja, budući da je Tjuring (Turing 1936) dao model idealne mašine koja može računati sve funkcije koje danas nazivamo Tjuring-izračunljivim, a Fon Nojman (von Neumann 1945/1993) i arhitekturu digitalnog računara, koji takve funkcije može efektivno računati. Ideja da um funkcioniše tako što vrši logičke operacije, napokon je mogla da se objasni bez pozivanja na nekog unutrašnjeg mislioca, pa je um postao biološki realizovan digitalni računar.

2. Tradicionalna kognitivna nauka i konekcionizam

2.1. Poreklo sukoba

Prvi pravi sukob između dva pristupa u kognitivnoj nauci nastaje tek sa jasno formulisanim programom konekcionizma i rešenjima problema sa kojima su se Rozenblatovi perceptroni suočavali. Naime, u narednoj deceniji, to jest osamdesetih godina prošlog veka, Makleland (McClelland) i Rumelhart (Rumelhart), u saradnji sa lingvistima, psiholozima i istraživačima veštačke inteligencije okupljenim u istraživačku grupu „PDP“ (paralelno distribuirano procesiranje) objavljuju takozvanu „PDP Bibliju“ – dva toma u kojima se detaljno opisuju svojstva modela višeslojnih neuronskih mreža i njihove primene u istraživanju različitih zadataka i kognitivnih procesa koji leže u osnovi obavljanja tih zadataka. Nedugo potom, Mekleland i Rumelhart

takođe objavljuju i udžbenik koji se prodavao uz dva flopi diska na kojima je bio softver sa primerima većine modela opisanih u „PDP Bibliji“, čime je konekcionistički istraživački poduhvat postao dostupan svakom studentu, ekspertu ili laiku (Berkeley 2019: 194). Glavna prednost konekcionizma, smatralo se, bila je biološka plauzibilnost modela, bazirana na prirodi neuronske aktivnosti, uz empiristički pristup poreklu kognitivnih procesa, koji je počivao na učenju kroz treniranje mreža.

Kako bismo jasno sagledali razlike između dva tada ponuđena programa – tradicionalnog i konekcionističkog, poslužićemo se sledećim uvidima Rodžersa (Rogers) i Meklelanda. Naime, oni u svom preglednom tekstu iz 2014. godine definišu kognitivnu nauku kao naučnu disciplinu kojom se traže odgovori na sledeća tri međusobno povezana pitanja: **(i)** koji procesi učestvuju u konstituisanju kompleksnog ponašanja inteligentnih sistema, kao što su ljudi, **(ii)** kakve reprezentacije su produkt pomenutih procesa, i **(iii)** kakva je osnova procesa i reprezentacija – urođena ili se formira zahvaljujući učenju? Odgovor na svako od tri pomenuta pitanja *teorijski* obavezuje kognitivnog naučnika na određene metodološke i ontološke pretpostavke – od toga za kakav tip modela će se odlučiti, kako će podesiti parametre, koji podaci će biti inkorporirani u model, pa do toga da li nam objašnjenja, koja dobijamo zahvaljujući modelima, govore nešto i o prirodi ljudskih kognitivnih procesa.

Teorijski okvir za proučavanje ljudske kognicije, karakterističan za rane dane kognitivne nauke sedamdesetih godina prošlog veka, kao što je već pomenuto, podrazumevao je analogiju između digitalnog računara i mozga. Ova analogija nesumnjivo je bila utemeljena u uspesima Tjuringa i Fon Nojmana, a neuspesima ranih modela neuronskih mreža kao računskih sistema, a koji su se ticali izračunljivosti. Klasični kompjutacionizam ili tradicionalna simbolička kognitivna nauka se može, stoga, definisati kroz tri odgovora na Rodžersova i Meklelandova pitanja:

> **(Ti)** kognitivni procesi su nalik programima digitalnog računara, to jest liče na uređene liste eksplicitnih ili implicitnih pravila, koja su *domenospecifična* i *sekvencijalna*;
> **(Tii)** reprezentacije su simboličke, i odlikuje ih *kombinatorijalna sintaksa i semantika*;
> **(Tiii)** ovako shvaćeni procesi i reprezentacije moraju biti *urođeni*, jer je broj mogućih lista pravila virtualno beskonačan, te je potrebno pretpostaviti nekakvo prvobitno ograničenje kako bi se specifikovala pravila.

U ovom periodu, vladao je entuzijazam kako u zajednici istraživača veštačke inteligencije u pogledu prospekata pravljenja inteligentnih mašina, tako i u zajednici kognitivnih naučnika u pogledu spremnosti simboličke kognitivne nauke da eksplanatorno obuhvati i niže i više kognitivne procese. Krajem sedamdesetih izgledalo je kao da naučnici više utvrđuju *status quo* nego što prave kartu do tada neistraženog

nepoznatog područja (Buckner & Garson 2018: 78). Vreme je, prema tome, bilo sazrelo za formulisanje rivalskog pristupa klasičnom kompjutacionizmu, a koje je bilo omogućeno uvođenjem algoritma propagiranja greške unazad i novih potentnih višeslojnih modela neuronskih mreža. Konekcionizam kao teorija o procesiranju informacija kognitivnih sistema koja se služi modelima neuronskih mreža može se definisati preko sledeća tri odgovora na pitanja Rodžersa i Meklelanda:

> **(Ki)** kognitivni procesi su nalik programu *analognog* računara, jer je neuronska mreža usmerena ka tome da pronađe najviše asocirani autput sa arbitrarnim inputom. Skrivene jedinice u trećem sloju, učestvuju u stvaranju šablona, koji predstavlja „ponašanje" neuronske mreže;
> **(Kii)** reprezentacije su *distribuirane* i sastoje se od niza jedinica nižeg ili subsimboličkog nivoa, čiji potpuni formalni opis može samo da aproksimira viši ili simbolički nivo;
> **(Kiii)** ovako shvaćeni procesi i reprezentacije se konstituišu „obučavanjem" neuronske mreže putem primera iz korpusa, to jest zahvaljujući njenom „učenju" iz *iskustva*.

Tenzija između konekcionističkih i simboličkih modela je najizraženija u pogledu odgovora na pitanje **(iii)**, budući da ih metodološke odluke upućuju na to da zastupaju ontološki suprotne teze u pogledu prirode kognitivnih procesa. S jedne strane, simbolisti tvrde da su kognitivni procesi urođeni i domenospecifični, što znači da su mnoge naše sposobnosti plod adaptiranosti, ili urođene spremnosti, za obavljanje specifičnih zadataka s kojima ćemo se susresti tokom odraslog života (Cowie 1999: 30). S druge strane, konekcionisti smatraju da su ključne karakteristike kognicije domenogeneralnost, to jest konstantno učenje putem iskustva, koje je, recimo, u obliku induktivnog rasuđivanja (Cowie 1999: 29), dok se urođenost specifičnih modula zaduženih za različite zadatke nastoji izbeći kao pretpostavka. Ova tenzija će igrati ključnu ulogu u dugogodišnjoj kritici konekcionističkih modela od strane zastupnika tradicionalne simboličke kognitivne nauke.

Kako bismo bolje sagledali goreopisanu prirodu sukoba između tradicionalista i konekcionista poslužićemo se jednim primerom u kojem se daju dva različita modela iste sposobnosti – sposobnosti izgradnje prošlog vremena u engleskom jeziku. Simbolički tradicionalni pristup pruža objašnjenje mehanizama zaslužnih za proizvodnju ovog kognitivnog fenomena, kao što Abrahamsen i Behtel (Abrahamsen & Bechtel 2006) primećuju, veoma blizu samih fenomena. Postuliraju se *dva* mehanizma, prvi koji izvršava operaciju „dodaj nastavak -ed" ako je glagol pravilan i drugi koji zahteva pronalaženje odgovarajućeg oblika u mentalnom leksikonu ukoliko je glagol nepravilan. Konekcionističko rešenje istog zadatka (Rumelhart et al. 1986, *PDP* poglavlje 18) pretpostavljalo je *samo jedan* mehanizam koji je operisao nad binarnim subsimboličkim inputom: svaka morfema iz korena glagola bila je kodirana na šest jedinica

inputa (kodirani su, recimo, zvučnost, da li je u pitanju samoglasnik ili suglasnik, itd.), a u skrivenom sloju ulazni signali su bili transformisani u vektor realnih brojeva, dok je sloj autputa nalikovao sloju inputa, to jest za svaku morfemu bilo je izdvojeno po 6 jedinica. Osim što je konekcionistički model nakon treniranja bio uspešan u izgradnji prošlog vremena pravilnih i nepravilnih glagola sa samo jednim mehanizmom, on je, budući da je ujedno dinamički model koji učeći predstavlja i evoluciju procesa dodatno reprezentovao i krivu učenja koja je izuzetno podsećala na onu koja se javlja kada deca uče prošlo vreme. Naime, i model i deca su u jednom trenutku, čak i nakon naučenih ispravnih oblika prošlog vremena, počinjali da preterano generalizuju dodavanje nastavka „-ed" na obe vrste korena glagola, drugim rečima i mašinsko učenje je ispoljavalo dečiji „razvojni profil oblika U", prvi put opisan u Berko (1958).

Na osnovu ovakvih slučajeva, gde su oba programa imala pretenziju da objasne isti fenomen i da pruže mehanicistička objašnjenja, koja su zauzvrat postulirala sasvim različite mehanizme dubinski nespojivih karakteristika, možemo reći da je sukob između simboličkog i konekcionističkog pristupta bio stvaran, a ne samo prividan kao što su neki autori poput Brodbenta (Broadbent 1985) tvrdili ističući da je konekcionizam samo teorija *implementacije*, a ne i kognicije.

2.2. Kritika Fodora i Pilišina: bauk sistematičnosti

Fodor i Pilišin (1988) su smatrali, slično Brodbentu, da konekcionizam ima ozbiljnijih problema ukoliko zaista ima pretenzije na kognitivna objašnjenja. Ovi autori su inaugurisali čuveni *argument o sistematičnosti* (kasnije rafiniran u Fodor & McLaughlin 1990): budući da neuronske mreže nemaju kombinatorijalnu sintaksu i semantiku, one ne objašnjavaju svojstvo sistematičnosti na odgovarajući način, štaviše one prema njima ne mogu čak ni da ispolje svojstvo sistematičnosti. Značaj ovog argumenta ogleda se u činjenici da je sistematičnost, prema njima, *suštinska* karakteristika ljudske kognicije. Čuveni primer sistematičnosti jezika i misli tiče se toga da ukoliko neko razume rečenicu Džon voli Meri, on mora biti u mogućnosti da razume i reprezentuje rečenicu Meri voli Džona. Sistematičnost, i s njom povezana *produktivnost* – mogućnost razumevanja i produkovanja virtualno beskonačnog skupa rečenica, nisu odlike koje misao može imati, već ih ona mora imati, i svaka teorija o kogniciji mora imati spremno objašnjenje zašto i kako one nastaju.

Uprkos tome što je bilo više konekcionistički nastrojenih autora koji su se uhvatili u koštac sa izazovom sistematičnosti (Smolensky 1988, Chalmers 1990), tvrdeći da nije potrebno da neuronska mreža bude strukturirana da bi mogla da bude senzitivna na sistematičnost, zastupnici tradicionalne simboličke nauke su insistirali na razlici između reprezentacija koje su *aktualno* sistematične i reprezentacija koje *predstavljaju* sintaksičku strukturu, čime sistematičnost *emergira* iz ponašanja, ali nije svojstvo same mreže (McLaughlin 1993: 178). Drugim rečima, ukoliko neuronske mreže mogu

i ne moraju da simuliraju sistematičnost pozivanje na njih ne može objasniti ovu suštinsku odliku ljudske misli, što dovodi do zaključka da je konekcionizam eksplanatorno manjkav kada su kognitivni fenomeni u pitanju. Za razliku od konekcionizma, kompjutaciono-reprezentacioni modeli upareni sa idejama lingviste Noama Čomskog (Noam Chomsky) mogli su spremno da objasne sistematičnost i produktivnost jezika. Kognitivni procesi se odvijaju nad sintaksičkim svojstvima simboličkih reprezentacija uz pomoć pravila nalik onima „univerzalne gramatike", to jest internalizovane strukture hijerarhijski postavljenih kategorija koje su uređene putem urođenih pravila, koja omogućavaju da deca u vrlo kratkom roku i na osnovu nekvalitetnih podataka ovladaju maternjim jezikom (1965: 57-58).

Ovde treba primetiti da čitav argument počiva na pretpostavci da je sistematičnost suštinska odlika misli čiji su glavni konstituenti simboličke reprezentacije, a ne samo jedna od karakteristika nekih misli. Mnogi autori su osporavali ovu pretpostavku tako što su ili u potpunosti negirali reprezentacionu strukturu misli (poput Churchland 1986, Churchland & Sejnowski 1990), ili tako što su argumentovali da je kod Fodora i Pilišina uloga sistematičnosti prenaglašena (npr. Dennett 1991). Takođe, mnoga empirijska istraživanja ukazuju na to da misao ni na fenomenalnom nivou često ne ispoljava logičku strukturu koja se postulira simboličkim pristupima: tako, na primer, Vejsonov (Wason 1966) zadatak selekcije otkriva slabe performanse u deduktivnom rasuđivanju ljudi u određenim kontekstima, a drugi autori ukazuju na slabu sposobnost ljudi da aktivno barataju sintaksičkim drvetima (Gordon, Hendrick & Johnson 2001; v. takođe Marcus 2014). Ipak, ukoliko se složimo da je sistematičnost, makar i ne bila suštinska odlika misli, zasigurno važan kognitivni fenomen, onda se čini da bi konekcionizam trebalo da pruži neko objašnjenje zašto i kako do nje dolazi – ako uopšte ima pretenzije na davanje kognitivnih objašnjenja. Tim putem pošli su Prins i Smolenski (Prince & Smolensky 1993/2004) ponudivši „Teoriju optimalnosti" koja kombinuje prednosti oba pristupa. Oni su, zapravo, zaslužni za ekspliciranje suštinskog pitanja: „Zašto bi statistički mehanizmi poput neuronskih mreža uopšte morali da imaju tako rigidne i uređene autpute u vidu jezika?" Nažalost, ipak nisu uspeli da daju konkluzivan odgovor na to pitanje. Iako je sasvim jasno da moraju postojati određena ograničenja na inputu ili samom procesiranju, ni do dan danas nije otkriveno kakve su ona prirode, pa samim tim nije ni pružen konkretan model koji bi bio sistematičan u jakom smislu.

Poslednje utočište konekcioniste koji prihvata sistematičnost kao bitan kognitivni fenomen je da odustane od pružanja kognitivnih ili psiholoških objašnjenja koja bi bila na algoritamskom nivou Marove trosolojne analize[5] i da pristane na tvrdnju da je njegova teorija samo teorija implementacije. Prema takvom stanovištu, simbolički

5 Mar (Marr 1982) izdvaja tri nivoa analize sistema koji procesiraju informacije – kompjutacioni, algoritamski i implementacioni nivo, dok ih Pilišin (1984) naziva semantičkim, sintaksičkim i fizičkim. Prema Pilinišovom mišljenju, samo su prva dva nivoa kogitivno relevantna.

pristup bi nam pružao objašnjenja kognitivnih fenomena kao izvesnih zakonitosti pružajući deduktivno-nomološka objašnjenja, a konekcionizam bi nam pružao mehanicistička objašnjenja na implementacionom nivou bez pretenzija na objašnjenje kognitivnih zakonitosti (vidi Abrahamsen & Bechtel 2006). Međutim, ovde dolazimo do novog problema, a to je problem biološke plauzibilnosti. Ukoliko konekcionizam treba da objasni kako su kognitivni procesi implementirani, od njega se onda očekuje da objasni ne samo kako izvesni procesi mogu biti implementirani generalno, već i kako su oni implementirani u slučaju ljudske kognicije. Na kraju krajeva, kao što smo videli na primeru dva modela izgradnje prošlog vremena, ono što se može staviti u prilog konekcionističkom modelu je njegova biološka plauzibilnost (jedinstveni mrežni mehanizam koji uči i razvija se na sličan način kao i dete) u odnosu na simbolički (dva mehanizma, prelistavanje mentalnog inventara, posebno memorijsko skladište, itd.). Ipak, pretenzije na biološku plauzibilnost suočile su se sa više problema. Jedan od njih predstavlja algoritam nadgledanog učenja putem propagiranja greške unazad, koji ne izgleda kao da je implementiran u mozgu (Buckner & Garson 2018: 80), a koji je, setimo se, uveden kako bi se predstavile vrednosti skrivenih jedinica i kako bi se omogućilo učenje mreža sa više slojeva.[6] Takođe, osim propagiranja greške unazad veštačke neuronske mreže iz osamdesetih godina propagirale su signale skoro uvek unapred, što nije slučaj sa signalima neuronskih moždanih sklopova gde postoje višestruki rekurentni putevi signala (v. npr. Kravitz et al. 2013).

Takođe, konekcionistički modeli nisu se suočili samo sa problemima biološke plauzibilnosti, već i sa problemima efikasnosti. Iako su se pokazali odličnim za modelovanje percepcije, viši kognitivni procesi su iziskivali kodiranje unapred da bi neuronska mreža uopšte mogla da produkuje smislene rezultate. Ključni nedostatak tadašnjih neuronskih mreža ležao je u „katastrofičnom zaboravljanju“ – tendenciji da se zaborave prethodno formirane veze između jedinica kada je potrebno da se obavi novi zadatak, ili kada nastupi obučavanje na prethodno neobrađenom delu korpusa (Buckner & Garson 2018: 81). Za simuliranje kompleksnijih kognitivnih domena potrebno je, dakle, da mreža „pamti“ prethodne informacije da bi mogla ponovo da ih iskoristi.

Ponovni progres došao je sa pojavom rekurentnih neuronskih mreža, unutar kojih postoji poseban sloj skrivenih jedinica između ulaznih i izlaznih jedinica, koje kodiraju ulazne i izlazne podatke, i predstavljaju „skladište“ informacija. Informacije iz ovog „skladišta“ bi bile poslate ulaznim jedinicama na ponovnu obradu. Uz primenu rekurentnih neuronskih mreža, tokom devedesetih godina neki metodološki problemi konekcionizma su počeli da se rešavaju putem inkorporiranja teorija dinamičkih sistema kako bi se omogućilo procesiranje u realnom vremenu, što se najviše koristilo

6 Propagiranje signala greške unazad menja snagu konekcije između nodova na takav način da snaga aktivacije nekog noda zavisi i od aktivnosti onih nodova koji nisu u neposrednoj komunikaciji sa njim što nije slučaj sa neuronskim sinapsama.

za modelovanje čitanja. Ipak, veliki deo problema koji se ticao eksplanatornosti, bilo kognitivne, bilo implementacione, i dalje je ostao na snazi – Kako jedan mehanički sistem učenjem može doći do apstraktnih simboličkih struktura? Kakvu eksplanatornu vrednost imaju modeli koji inkorporiraju biološki neplauzibilne elemente? Kako objasniti sticanje jezika?, itd., koji su doveli do još jedne „zime" u razvoju konekcionizma

3. Postkonekcionistički modeli: nove mreže i stare kritike

3.1. Duboko učenje i postkonekcionistički modeli

Inspirisan nalazima Hubela i Visela (Hubel & Wiesel 1968) koji su sugerisali da postoje dve vrste ćelija u vizualnom korteksu mačke i koji su pružili opis njihove organizacije i strukture receptivnih polja, Fukušima (Fukushima) 1980. godine dizajnira *neokognitron*, preteču modernih konvolucionih mreža – mreža koje u barem jednom svojem sloju koriste linearni operator konvolucije i koje predstavljaju višeslojne perceptrone, odnosno mreže čiji su slojevi potpuno povezani[7]. Međutim, ozbiljniji razvoj dubokih konvolucionih neuronskih mreža (DKNM)[8] nije bio moguć bez nove tehnologije koja se pojavila tek dvehiljaditih godina, pre svega grafičke jedinice za procesiranje (ili skraćeno GPU), dizajnirane za zahtevne video igre, i „*big data*"[9], koji će zajedno omogućiti procesiranje velikog broja parametara i postizanje velike dubine mreža koje danas imaju i do sto slojeva, za razliku od klasičnih konekcionističkih mreža koje su tipično imale tri do četiri sloja.

---

7 Svaka jedinica u sloju je povezana sa svakom drugom jedinicom sledećeg sloja, a funkcije aktivacije su nelinearne. Većina neuronskih mreža jeste potpuno povezana, mada to ne mora biti (za modelovanje parcijalno povezanih neuronskih mreža prema tipovima inputa v. npr. Kang & Isik 2005)

8 DKNM se razlikuju od ranijih konekcionističkih modela po konvolucijskim filterima i sistemom sjedinjavanja (eng. *pooling*). Konvolucijski filteri služe za uprošćavanje velikog broja parametara i izdvajanje pojedinih odlika ili, konkretno, za izoštravanje i zamućenje slika, i za razaznavanje ivica slika, ili dokumenata i rečenica u zavisnosti da li se DKNM koriste u domenu kompjuterske vizije ili za procesiranje prirodnog jezika (Gu et al. 2017). Primenom ovih filtera, u tzv. konvolucijskom sloju neuronske mreže, moguće je, dakle, izvući ciljane karakteristike slika ili objekata (LeCun et al. 2015: 438-439). Sistem sjedinjavanja potom redukuje dimenzije podataka tako što se pomoću njega kombinuju autputi grupe jedinica jednog sloja u novu jedinicu unutar narednog sloja (LeCun et al. 2015: 439). Ove neuronske mreže se obučavaju tako što se u ulazni sloj uvodi neka slika predstavljena preko svoje širine, visine i vrednosti piksela, i mreža pamti njene karakteristike.

9 Od posebnog značaja izdvaja se projekat *ImageNet* kojim je sakupljeno više od 14 miliona anotiranih slika za treniranje mreža; za poređenje, dosta korišćen skup podataka CIFAR-100 sadrži „samo" 60 hiljada anotiranih slika.

Takođe, dvehiljaditih je ostvaren veliki napredak i na polju rekurentnih neuronskih mreža – mreža koje zahvaljujući svojoj arhitekturi mogu da procesiraju sekvencijalne ili serijske inpute (inpute čiji članovi serije mogu imati uticaj na druge članove) promenljive veličine. S obzirom na ovu odliku, kao i jako ispoljavanje dinamičkog ponašanja, rekurentne mreže pokazale su se posebno potentnim za procesiranje govora, muzičkih i drugih vrsta sekvencijalnih inputa. Iako su prisutne od osamdesetih godina prošlog veka, devedesetih i dvehiljaditih se dizajniraju nove arhitekture, poput arhitekture dugog kratkoročnog pamćenja (otuda se skraćeno nazivaju LSTM) inspirisane biološkim modelom pamćenja prefrontalnog korteksa (Hochreiter & Schmidhuber 1997; Graves et al. 2009), čija efikasnost se uvećava zahvaljujući moćnijim jedinicama za procesiranje.[10]

Takvi tehnološki i arhitekturalni pomaci u dizajnu neuronskih mreža će posledično omogućiti i efikasno duboko učenje (eng. *deep learning*), statističku metodu za klasifikovanje šablona, koji nastaju kao rezultat obučavanja neuronske mreže na ogromnom broju podataka da sama pronađe intrinsičnu strukturu koja ih objedinjuje (videti LeCun et al. 2015: 436). Posebno zanimljiva osobina dubokog učenja je mogućnost da se ono vrši nenadgledano i na nestrukturiranim podacima (ranije se obučavanje mreža vršilo isključivo sa inputima koji su bili obrađeni u skladu sa preferiranim karakteristikama), koja je omogućila rešavanje novih zadataka u najrazličitijim oblastima.

S obzirom na ovakav značajan razvoj u oblasti neuronskih mreža i mašinskog učenja sasvim je opravdano postaviti sledeća pitanja: Da li novi modeli pružaju bolja kognitivna objašnjenja? Takođe, da li su ovi modeli i biološki plauzibilniji? Da li napredak u njihovoj arhitekturi i mogućnostima izračunavanja može da se uporedi sa napretkom od jednostavnog perceptrona do mreža sa skrivenim slojevima koje su se obučavale putem algoritma propagiranja greške unazad, i koje su rešile probleme izračunljivosti nelinearnih funkcija?

3.2. Nova stara kritika s obzirom na neadekvatnost objašnjenja sistematičnosti

Argument protiv sistematičnosti Fodora i Pilišina nastavio je da se koristi kao oruđe za napad, ali sada na postkonekcionističke pristupe koji sa klasičnim konekcionizmom dele glavne osobine arhitekture kao što su neposedovanje kombinatorijalne semantike i sintakse. Ovaj argument ponavlja se u debatama koje se tiču procesiranja jezika i govora (Marcus, Vijayan, Bandi Rao, and Vishton 1999; Peña et al. 2002), a autori poput Aizave (Aizawa 2003) iznova insistiraju da nije dovoljno da model ispolji sistematičnost kao emergirajuće svojstvo podešavanja parametara mreže, već da ona

10 Više o različitim vrstama neuronskih mreža v. Hassabis et al. (2017: 246-247, 254). Važno je napomenuti da je najčešća praksa kombinovanje DKNM i rekurentnih mreža ukoliko zadatak zahteva više od perceptivnih „sposobnosti“ mreže, recimo, kada je potrebno da se generiše i opis slike koju mreža prethodno prepoznaje.

mora biti nomološki nužna posledica same arhitekture. Markus (2018b) takođe preslikava kritiku Fodora i Pilišina, ali tako da se sada primenjuje na duboko učenje, a ne oslanja se na univerzalnu gramatiku Čomskog kao njegovi prethodnici, već pak na minimalistički program Čomskog, koji predstavlja rafiniranu verziju njegovog stanovišta o urođenosti jezičke sposobnosti.[11]

Međutim, kao što smo već sugerisale ranije, ova vrsta argumenta zavisi kako od pretpostavke da je sistematičnost suštinska odlika mišljenja, tako i od pretpostavke da je jedina validna vrsta objašnjenja u kognitivnoj nauci objašnjenje putem funkcionalne analize kakvu opisuju Fodor i Kamins (Fodor 1968; Cummins 1983). Funkcionalna analiza zahteva funkcionalnu dekompoziciju kognitivnih sposobnosti i postuliranje jednostavnijih „ne sasvim kognitivnih" sposobnosti (Figdor 2018), ili „glupljih" homunkula, od kojih je kompleksnija sposobnost koja se objašnjava sastavljena, a koje se moraju razlikovati od onih samog sistema kako objašnjenja koja se na njih pozivaju ne bi bila cirkularna. Proces se ponavlja sve dok se ne dođe do nekognitivnih procesa i sposobnosti. Klasični simbolički pristup se tako u objašnjenju sistematičnosti pozivao na svojstva i sposobnosti delova njegove arhitekture koja je uključivala kompozicionalno struktuirane reprezentacije i kodirana pravila koja operišu nad strukturalnim svojstvima takvih reprezentacija, a smatralo se da konekcionizam nema resurse da pruži takvu vrstu objašnjenja. Međutim, značaj i relevantnost funkcionalne analize kao pravog modela za kognitivna objašnjenja sve više opada, jer se u oblasti kognitivne nauke javljaju alternativni pogledi na eksplanatornost.

Metjuz (Matthews 1997) podrobno kritikuje napad simbolista na konekcionizam ukazujući **(a)** da je sama sistematičnost neplauzibilno predstavljena u simboličkoj kognitivnoj nauci; **(b)** da simbolička kognitivna nauka zapravo ne pruža objašnjenje sistematičnosti; i **(c)** da ne mora svako kognitivno objašnjenje da se poziva na „gluplje" homunkule, već može biti i na nivou implementacije čak i u okvirima koje zadaje funkcionalne analiza. Povodom **(a)** Metjuz navodi da prema tradicionalnoj, simboličkoj kognitivnoj nauci sledi da, ukoliko neko može da zamisli *a*R*b*, on onda mora biti sposoban da zamisli, odnosno reprezentuje, *b*R*a*. Međutim, „ja mogu misliti misao da je *x* jedini član singlton skupa {*x*}, ali sam sasvim siguran da ne mogu da mislim misao da je singlton skup {*x*} jedini član *x*-a" – tvrdi Metjuz (1997: 162). Kada je **(b)** u pitanju, Metjuz ističe da je i sam Meklahlin (McLaughlin 1993) bio svestan toga da

11 Naime, u periodu od 1995. godine do početka dvehiljaditih, Čomski počinje da zastupa minimalistički program, prema kom razlikujemo jezičku sposobnost u širem smislu, koja podrazumeva senzorno-motorni sistem i pojmovno-intencionalni sistem kao i jezičku sposobnost u užem smislu, koju predstavlja apstraktni kompjutacioni sistem uređen prema urođenom mehanizmu rekurzije (Hauser et al. 2002). Specifičnost ljudske jezičke sposobnosti u odnosu na ostale ne-ljudske životinje (naročito primate) jeste mogućnost generisanja beskonačnog skupa iskaza na osnovu konačnog skupa elemenata putem mehanizma rekurzije. Budući da je ovaj mehanizam urođen, i dalje omogućava definisanje sintaksičkih kategorija, koje se tako postavljaju u hijerarhijski sistem.

simbolisti poput njega samo *obećavaju* da će pružiti odgovarajuće objašnjenje sistematičnosti, ali da nam ostaju *dužni* u pogledu formulisanja i ekspliciranja kompozicionalne sintakse jezika misli, odgovarajuće psihosemantike, kompjutacionih modela intencionalnih modusa, itd. (1997: 160). I na kraju, povodom **(c)**, on primećuje da i sam Kamins (1983) dopušta da neke kognitivne sposobnosti mogu biti objašnjene putem instancijacije (npr. izračunavanje funkcije „i“, kao što smo videli, može da se objasni mehaničkim nekognitivnim procesom). Ipak, Kamins je bio ubeđenja da interesantnije kognitivne sposobnosti ne mogu biti tako objašnjene, jer bi nam bez umetanja homunkula one ostale *misteriozne*. Odbacujući argument misterioznosti, Metjuz tvrdi da konekcionizam može da pruži jednu vrstu objašnjenja putem instancijacije, ali bez specifikovanja konkretnog mehanizma, već *indirektno* i *induktivno* pružajući objašnjenje relevantne funkcije uz pomoć *jednostavnog mehanizma* čiju strukturu bi delio i neki kompleksniji mehanizam (1997: 174-176).

Keri Fidžor (2018) ide dalje od Metjuza i argumentuje da je sasvim izlišan zahtev da se kognitivne sposobnosti objašnjavaju putem jednostavnijih sposobnosti, koje su u nekim pogledima različite od onih sa višeg personalnog nivoa, ali u drugim pak slične (one su „ne sasvim kognitivne“, ali ipak kognitivne u nekom smislu), pa posledično neeksplanatorne. Prema Fidžor (2018) funkcionalna analiza ili homunkularni funkcionalizam je samo zaostatak iz prošlih vremena kada nismo imali neantropocentričku perspektivu na um. Njome su se nepotrebno postulirali sve gluplji homunkuli, kako bi se perspektiva prvog lica postepeno gubila i kako bi se „misterija“ psihološkog sve duže odgađala, sve dok ne dođemo do procesa i mehanizama koji nemaju kognitivna i psihološka svojstva.

Konačno, kada su kognitivna objašnjenja u pitanju Pićinini i Krejver (Piccinini & Craver 2011) pokazuju da su objašnjenja funkcionalne analize i mehanicistička objašnjenja zapravo objašnjenja iste vrste, gde funkcionalna analiza pruža samo nezadovoljavajuće skice mehanizama koje treba upotpuniti. Ukoliko su ovi autori u pravu, (post)konekcionisti nisu u obavezi da pruže objašnjenja koja se pozivaju na jednostavnije sposobnosti od kojih je kompleksna sposobnost sastavljena, već mogu pružiti i neku vrstu mehanicističkog objašnjenja, koje je bliže nivou implementacije. Međutim, ovo ponovo otvara stare probleme zaodenute u novo ruho.

3.3. Nova stara kritika implementacije

Ukoliko postkonekcionistički modeli treba da nas približe mehanizmima koji proizvode kognitivne fenomene, onda se tvrdi da bi modeli trebalo da budu efikasni, to jest da verno simuliraju proizvodnju relevatnih fenomena, da budu biološki plauzibilni i strukturalno transparentni. Iako su i DKNM i rekurentne mreže poput LSTM, kao i njihove preteče – neuroni Mekaloha i Pitsa i Rozenblatov perceptron, modelirane po

uzoru na određene biološke karakteristike neuroanatomije[12], prelaz od ranih konekcionističkih na postkonekcionističke modele može da se okarakteriše kao sve veće zanemarivanje biološke plauzibilnosti. Naime, kognitivna nauka i neuronauka su tokom dobrog dela druge polovine dvadesetog veka bile tesno povezane sa istraživanjem veštačke inteligencije, što se očitovalo u vođenju računa o fiziološkim i neuralnim ograničenjima uključenim u rane konekcionističke modele, ali je saradnja u dvadeset prvom veku značajno manjeg intenziteta (Hassabis et al. 2017: 245). Razlog za to je pre svega specijalizacija težišta istraživanja – inženjeri veštačke inteligencije se bave veštačkim ekspertskim sistemima koji treba da obave određeni zadatak, i usled toga i kompjutacione detalje podređuju programiranju optimalnog obavljanja zadatka umesto biološki i kognitivno realističnom obavljanju zadatka. S druge strane kognitivni naučnici su se našli u živom blatu rasprave o prednostima i manama simbolizma odnosno konekcionizma, i primena postkonekcionističkih modela je stoga više fokusirana na izvlačenje kognitivnih posledica iz, recimo, modela koji je pobedio ljudskog eksperta u korejskoj igri go (Schubbach 2019), nego na brigu u pogledu toga koliko je sam model biološki plauzibilan.

Umanjenje biološke plauzibilnosti i kognitivne efikasnosti navelo je autore poput Gerija Markusa (2018a; 2018b) da se fokusiraju na trenutne neuspehe neuronskih mreža u modeliranju kognitivnih procesa. Markus (2018a) kritikuje neuronske mreže koje se služe dubokim učenjem tvrdeći da takve mreže ne mogu imati mehanizme za učenje apstrahovanja osim ako se unapred ne kodiraju eksplicitne verbalne definicije. Dalje, budući da su potrebni milioni primera za obučavanje, češći je slučaj da duboke mreže preterano generalizuju, nego što su u stanju da pariraju ljudskom kognitivnom procesiranju. Markus se, takođe, slaže sa psihologom Brendenom Lejkom (Brenden Lake) koji u preglednom radu sa Baronijem (Lake & Baroni 2018), uz obavezno citiranje Fodora i Pilišina, tvrdi kako neuronske mreže „i dalje nisu sistematične nakon toliko godina". Dalje se navodi da su postkonekcionistički modeli još uvek značajno pogođeni razlikama između rečenica korpusa na kome se obučavaju i rečenica korpusa na kome se testiraju prilikom obavljanja zadatka procesiranja prirodnog jezika. Za ljude bi takav zadatak bio trivijalan, ali neuronskim mrežama su potrebni preveliki skupovi rečenica i previše vremena da bi se na ograničenom skupu rečenica približile ljudskom procesiranju. Iako se u svojoj kritici Lejk fokusira na rekurentne neuronske mreže, Markus smatra da se ovakva argumentacija može podjednako odnositi na *bilo koju* vrstu neuronskih mreža kojima se pokušava simulirati *bilo koji* viši kognitivni proces.

12 Istraživanja u komparativnoj neuroanatomiji sugerišu jednako distribuiran skup vizuelnih šablona u korteksu mačke pri stimulaciji, umesto decenijskog razlikovanja jednostavnih i kompleksnijih ćelija koje stvaraju različite vizuelne šablone (Priebe et al. 2004). A, podsetimo se, biološka inspiracija za DKNM se sastojala u pokušaju da se napravi analogija između toga kako su ćelije u vizuelnom korteksu mačke osetljive na izdiferencirano vizuelno polje i kako su artificijelni neuroni raspoređeni u konvolucijskom polju osetljivi na određene oblasti slike.

Pored nemogućnosti da se na prirodan način nose sa hijerarhijskim strukturama poput jezika, sklonosti ka preteranoj generalizaciji i potrebi za ogromnim brojem primera pri obučavanju, kritičari iznova ističu i problem „crne kutije" kao i biološku neplauzibilnost postkonekcionističkih modela. Prema prigovoru „crne kutije" (post) konekcionistički modeli su opisani pomoću mapiranja inputa na odgovarajući autput. Šta se dešava između, kako se input transformiše, ostaje sasvim netransparentno – poput sadržaja crne kutije. Mozer i Smolenski su prihvatili ovu osobinu neuronskih mreža kao njihovu vrlinu jer je „ono što konekcionističke mreže imaju zajedničko sa mozgom to što kada ih otvoriš i proviriš unutra, sve što vidiš jeste samo gomila kaše" (Mozer & Smolensky 1989: 3). Ipak, jedan broj autora se uhvatio u koštac sa ovim problemom i ponudio veliki broj metoda za analizu neuronskih mreža, kao što su klaster analiza, analiza glavne komponente, analiza uvezivanja, itd. (za pregled v. Browne 1997, a za najnovije pokušaje v. Zednik 2019).

S druge strane, kada je biološka neplauzibilnost u pitanju ponovo se ističu algoritmi učenja koji se oslanjaju na propagiranje greške unazad, a posebno neplauzibilna osobina DKNM koju kritičari vole da izdvoje jeste deljenje težina u konekcijama neurona (eng. *weight sharing*) (v. Bartunov et al. 2018).

3.4. Da li je sve tako crno u crnim kutijama neuronskih mreža?

Ipak, nije sve tako crno kada su neuronske mreže u pitanju. Pored nekih neplauzibilnih strukturnih detalja, DKNM inkorporiraju i biološki plauzibilne mehanizme kao što je sama operacija konvolucije, a za koje se može ispostaviti da imaju značajne posledice za kognitivno procesiranje, kao što se, na primer, ispostavilo da mreže sa više slojeva mogu da računaju XOR funkciju za razliku od jednostavnog perceptrona. Takođe, oblast koja se bavi algoritmima za učenje neuronskih mreža donela je niz novih plauzibilnijih algoritama. Pa su tako neplauzibilnim algoritmima propagiranja greške unazad ponuđene alternative poput metode recirkulacije opisane još kod Hintona i Meklelanda (Hinton & McClelland 1988), metode inspirisane biološkim pojačanim učenjem koje se fokusiraju na očekivanje nagrade umesto na minimiziranje greške (Williams 1992), pa čak i metode koje modeliraju uticaj konkretnih neurotransmitera i neuromodulatora na parametre mreže (v. za pregled Buckner & Garson 2019: 84).

S druge strane, kada su u pitanju kritike koje se tiču neefikasnosti postkonekcionističkih modela, odnosno njihove nemogućnosti da verno simuliraju određene kognitivne zadatke, jedan broj autora im suprotstavlja mnogo pozitivnije viđenje trenutnog stanja u oblasti neuronskih mreža. Posebno interesantan pregled, pogotovo s obzirom na problem sistematičnosti i tvrdnje Markusa i Lejka da neuronske mreže „i dalje nisu sistematične nakon toliko godina", pruža zbornik Kalva i Sajmonsa (Calvo & Symons) iz 2014. pod nazivom *The Architecture of Cognition: Rethinking Fodor and Pylyshyn's Systematicity Challenge*. U njemu možemo naći rad Stefana Franka (Stefan L. Frank) u kojem se poredi efikasnost rekurentne neuronske mreže sa probabilističkim frazno

struktuiranim gramatičkim modelom rečeničkog procesiranja i koji zaključuje da su njihove performase zapanjujuće slične kada se testiraju pod uslovima koji uzimaju u obzir različita ograničenja stvarnog sveta. Eksplanatorni potencijal uspeha RNN se, s druge strane, otkriva u dinamičkim bazenima atrakcije, pa se pokazuje na koji način se mogu kombinovati postkonekcionistički modeli sa teorijom dinamičkih sistema kako bi se objasnili viši kognitivni fenomeni. Ališa Koram (Alicia Coram) u zborniku nudi jedan drugačiji pogled, koji je sasvim kompatibilan sa unutrašnjom postkonekcionističkom arhitekturom, a koji nas upućuje van bioloških granica subjekta. U skladu sa proširenim pristupom kogniciji, Koram predlaže da poreklo sistematičnosti ne treba tražiti u urođenim moždanim strukturama, već u okolini, odnosno jeziku i drugim javnim reprezentacionim shemama shvaćenim kao artefaktima koji povratno utiču na kognitivne sposobnosti (za pregled v. Symons & Calvo 2014).

Pored ovih pokušaja rešenja starih problema, koji ukazuju suprotno kritičarima da novi modeli mogu da simuliraju i na određeni način objasne fenomene poput sistematičnosti, ali da ih je potrebno posmatrati kao uronjene u stvarni svet, u recentnoj literaturi možemo naći i jedan primer objašnjenja moći apstrakcije pomoću strukturnih osobina DKNM. U pitanju je argument Kamerona Baknera (2018) da odgovarajuće aktivacione funkcije implementiraju jednu vrstu kognitivne apstrakcije čime na empiristički način možemo da objasnimo apstraktnu kategorizaciju baziranu na pojedinačnim percepcijama. On u svom tekstu iz 2018. iznosi dve teze: **(i)** da se korišćenjem DKNM može modelovati ključna karakteristika ljudske inteligencije: kategorijalno apstrahovanje, što bi predstavljalo korak ka pokazivanju koje karakteristike treba modelovati da bi se došlo do verne simulacije opšte inteligencije; **(ii)** da upotreba DKNM pruža potvrdu za staru empirističku ideju da informacije apstrahovane iz iskustva omogućavaju više kognitivne sposobnosti, poput zaključivanja, donošenja odluka, i sl. Bakner, zapravo, u liniji sa ostalim konekcionistima, izražava ista uverenja – da (post)konekcionistički modeli *generalno* mogu da obuhvate i niže i više kognitivne procese, kao i da (post)konekcionistički modeli *generalno* pružaju odbranu empirizma i ključnih pretpostavki da su učenje i domenogeneralne sposobnosti osnova za razumevanje ljudske kognicije.

Dosadašnje modelovanje prepoznavanja objekata na osnovu percepcije, a koja potpadaju pod svakodnevne kategorije poput: „koala“, „krevetac“ ili „pelena“, nailazilo je na problem perspektive, jer se u odnosu na perspektivu menjaju i svojstva percipiranja. Prema tome, kako bi se omogućilo ispravno kategorisanje potrebno je ujediniti različite perspektive radi prepoznavanja jednog predmeta. U terminima modelovanja DKNM ovo znači da je u zadacima vizuelnog prepoznavanja potrebno kontrolisati skup „ometajućih“ promenljivih (eng. *nuisance variables*) – koje se tiču veličine, pozicije i ugaone rotacije objekata. Napokon, kategorijalno apstrahovanje koju obavljaju DKNM bi prema Bakneru moglo da se definiše na sledeći način, imajući u vidu postojanje „ometajućih“ promenljivih (2018: 5348): jedan reprezentacioni

format je apstraktniji od drugog u odnosu na način obavljanja zadatka klasifikacije; specifičnije, ukoliko je jedan format tolerantniji na ometajuće varijacije – koje moraju da se prevaziđu radi uspešnog obavljanja zadatka – on je ujedno apstraktniji. Ova hipoteza nam daje objašnjenje teze **(i)**. Međutim, kako bi se zasnovala teza **(ii)** potrebno je pokazati da je ovakav način apstrahovanja DKNM biološki plauzibilan, odnosno da se i ljudsko apstrahovanje vrši na analogan način.

Predlog za zasnivanje teze **(ii)** Bakner razvija primećujući najpre da je način na koji ljudi klasifikuju više predmeta kao pripadajućih istoj kategoriji u vezi sa procesiranjem perceptualnih sličnosti, koje se događa u mozgu sisara. Za potrebe modelovanja ovog procesa, prostor perceptualnih sličnosti se razumeva kao višedimenzionalni vektorski prostor u kojoj svaka dimenzija predstavlja jedno od perceptualno razlučivih svojstava. „Otisak" (eng. *manifold*) je deo ovog vektorskog prostora koji se uzima za označavanje granica reprezentacije kategorije, a „ometajuće" promenljive utiču na ovaj deo prostora tako što onemogućavaju razlučivanje granice različitih apstraktnih kategorija. Svaki sisar tokom vizuelnog procesiranja uči da sprovede niz transformacija unutar prostora perceptualne sličnosti kako bi se smanjio uticaj „ometajućih" promenljivih, kao što uče i DKNM. Važna pretpostavka je da je ovaj proces subpersonalan i nedostupan introspekciji. Stoga, Lok, čije shvatanje apstrakcije Bakner smatra bliskim načinu operisanja DKNM u zadacima apstrahovanja, ne bi morao da se interpretira kao da govori o apstraktnoj ideji trougla u vidu mentalne slike nekonzistentnih svojstava dostupne introspekciji, već takva ideja može biti nešto subpersonalno kao što je transformisan „otisak" kategorije. Tako, Lokov trougao bi u isti mah uključivao međusobno nekonzistentna svojstva (odnosno „ometajuće" promenljive koje utiču na „otisak"), ali bi se putem apstrakcije transformisale idiosinkrazije pojedinačnih trouglova tako da bi se primerci iste kategorije trougla grupisali u isti „otisak" (Buckner 2018: 5349).

4. Eksplorativni (post)konekcionistički modeli: od biološke plauzibilnosti do biološke aktualnosti?

Do sada smo nastojale da opišemo dugogodišnju kritiku konekcionizma kao neadekvatnog programa u oblasti kognitivne nauke i razloge za njegovo napuštanje u različitim vremenskim periodima. Ono što se kristalizuje u ovom pregledu jeste forma njegove kritike. Kritika konekcionizma ima dva aspekta: **(a)** glavni *opšti* aspekat, gde se tvrdi da za konekcionističke mreže simuliranje nekog kognitivnog procesa *nije uopšte moguće*, **(b)** *uži* aspekat, gde se tvrdi da za konekcionističke mreže *nije moguće trenutno* da obave zadatak. Ono što je zajedničko svim kritičarima (Minsky & Papert 1969, Fodor & Pylyshyn 1988, Marcus 2018a, 2018b, Lake & Baroni 2018, i dr.) jeste da oni polaze od **(b)**, odnosno pokazuju da neki model nešto ne može da učini, a odatle

generalizuju tvrdnju **(a)** da konekcionistički modeli uopšte to nešto ne mogu da učine. Sasvim je jasno da se generalizacija od **(b)** do **(a)** može opravdati samo ako se ukaže na neku suštinsku odliku konekcionističke arhitekture koja onemogućava proizvodnju odgovarajućeg efekta.

Međutim, kao što smo videli u slučaju Minskog generalizacija je bila sasvim neopravdana. Fenomen koji nije mogao da bude simuliran bilo je rešavanje nelinearnih funkcija poput ekskluzivne disjunkcije. Međutim, svojstvo jednostavnog perceptrona da rešenja funkcije deli isključivo na linearan način, nije ujedno i svojstvo mreža koje su sastavljene od takvih perceptrona. Na osnovu jednostavne mereološke greške konekcionizam je morao da sačeka osamdesete godine da se ponovo vrati na scenu.

U slučaju kritika Fodora, Pilišina, Meklahlina i Markusa postoji posredno opravdanje generalizacije koje ide preko metodoloških pretpostavki. Naime, usvajanjem funkcionalne analize zahteva se postuliranje svojstava i sposobnosti, poput reprezentacionih struktura koje su kompozicionalne, a koje neuronske mreže naprosto ne mogu imati. Onda kada se naiđe na neku mrežu koja ne može da simulira neki kognitivni fenomen – argumentuje se „pa očekivano, one to i ne mogu činiti, jer nemaju kombinatorijalnu semantiku i sintaksu". Međutim, ovde su jake ontološke pretpostavke kompozicionalnosti, urođenosti i domenospecifičnosti viših kognitivnih procesa na osnovu kojih se opravdava generalizacija, bazirane na relativno slabim metodološkim pretpostavkama. Dok trenutnu limitiranost neuronskih mreža ovi autori pejorativno opisuju preko klasičnog asocijativizma u psihologiji (Fodor & Pylyshyn 1988: 64), a svaku inovaciju dočekuju kao nešto što smanjuje biološku plauzibilnost koja bi donela prednost konekcionizmu makar kao teoriji implementacije (Marcus 2018b), oni kao da ne primećuju da njihova sopstvena pozicija zavisi od relativno zastarelih metodoloških principa funkcionalne analize koja nastoji da izbegne „misterioznost" psihološkog i kognitivnog postuliranjem homukula čije osobine mogu imati samo simboličke arhitekture. Ukoliko se takva metodološka pretpostavka ukloni, tradicionalna kognitivna nauka gubi na svojoj eksplanatornosti, a njegove ontološke pretpostavke ostaju jednako opravdane kao i konekcionističke, ako ne i manje s obzirom na biološku neplauzibilnost simboličke arhitekture.

Inspirisane radovima Metjuza, Fidžor, Pićininija, Krejvera, Stinson, Fišera i dr., koji ukazuju na drugačiju metodologiju kognitivne nauke želimo da ukažemo na neuspešnost i irelevanciju dosadašnje kritike konekcionizma usvajajući eksplorativnu mehanicističku metodologiju koja modele vidi kao nepotpune skice mehanizama. Takođe, usvajajući takvu metodologiju moguće je objasniti napredak (post)konekcionističkih modela u simuliranju različitih kognitivnih fenomena koji bi ostao sasvim misteriozan ukoliko bismo zadržali metodološke pretpostavke funkcionalne analize.

4.1. Kako izgleda proces eksploracije u postkonekcionističkim modelima?

Važno je odmah napomenuti da konekcionistički modeli mogu da opisuju kognitivne procese na bilo kom nivou detalja, iako predstavljaju idealizovane modele fiziologije mozga. Naime, ovi modeli rekreiraju fiziološke detalje aktualnih kognitivnih procesa tako da se preispita da li kvalitativna razlika u određenim detaljima doprinosi ili odmaže funkcionisanju procesa (Stinson 2018: 127). Međutim, konekcionistički modeli nisu realistične simulacije mozga, niti je potrebno da budu takvi kako bi bili eksplanatorno vredni (Buckner 2018: 5367). Zaključke koje donosimo na osnovu takvih modela nisu ništa manje pouzdani nego što bi to bili zaključci koje usvajamo iz eksperimentalnih naučnih disciplina (v. Stinson *forthcoming*).

Ukoliko (post)konekcionističke modele shvatimo na takav način, odnosno da se njima ne obavezujemo na jaku biološku plauzibilnost i detaljna mehanicistička objašnjenja, a uz odbacivanje pretpostavke funkcionalne analize i njenih ontoloških posledica na kognitivnu arhitekturu, te prihvatanja onoga što je Kamins (1983) nazvao „misterioznošću" psihološkog, (post)konekcionističkim modelima možemo dodeliti eksplorativni status. Stinson, slično Metjuzu (1997) predlaže da svrha implementacije neuronskih mreža nije dedukovanje preciznih aktivnosti moždanih struktura ili opisivanje konkretnih mehanizama, već u istraživanju i otkrivanju *generičkih mehanizama* koji vode funkcionisanje mozga (2008: 121). Generički mehanizam, prema Stinsonovoj (2018: 129), nastaje kada se otkrije regularnost između dva specifično ustrojena skupa činjenica, i operiše u okviru određenog raspona vrednosti parametara.[13] Proces eksploracije bi tekao na sledeći način: matematičkim formulisanjem i empirijskim posmatranjem postavila bi se hipoteza da generički mehanizam M_1 ima tendenciju da pod uslovima U_1, U_2, U_3...U_n produkuje određeni tip ponašanja P_1. Sledeći korak je da otkrijemo, na osnovu prethodnog znanja o odnosu M_1 i P_1, odnos moždanih struktura prema određenom kognitivnom procesu (Stinson 2018: 130).

Pogledajmo na primeru postkonekcionističkih modela koji implementiraju DKNM kako izgleda proces eksploracije. Zanimljivo je da su aktualne DKNM konstruisane tako da oponašaju vizuelni korteks mačke: putem matematičke formulacije i primene konvolucionih filtera i sistema sjedinjavanja, i empirijskog ispitivanja procesa opažanja sisara, preciznije mačaka, postavlja se hipoteza da generički mehanizam M_1 (raspoređivanje neurona, odnosno jedinica, u konvolucijskom polju tako da se povežu sa neuronima, odnosno jedinicama, u drugim slojevima) ima tendenciju da pod uslovima U_1, U_2, U_3...U_n (kada se, recimo, ćelije učine senzitivnim na receptivna polja, odnosno kada jedinice prime signal iz sloja inputa tokom obučavanja) produkuje ponašanje P_1

13 Stinsonova je inspirisana shvatanjem generičkih mehanizama Džona Stjuarta Mila (John Stuart Mill). Milu je ovaj termin bio potreban da ispita odnos uzroka i posledice: prvo je potrebno razložiti događaj na činjenice do one mere do koje nam je to potrebno, da bismo onda posmatrali koje činjenice slede iz kojih, to jest koje su uzroci a koje posledice u odnosu na okolnosti koje se eksperimentalno variraju (Mill 1843: ch. 7).

(diferenciranje vizuelnog polja tako da se prepozna specifičan objekat, odnosno detektovanje „otiska“). Na osnovu ovog eksplorativnog procesa u kome učestvuju DKNM možemo reći nešto o odnosu vizuelnog korteksa prema procesu opažanja, ili uz dodatne detalje, o procesu apstrahovanja, kako je Bakner pokazao svojom interpretacijom transformacione apstrakcije koju obavljaju DKNM.

Odbacivanjem funkcionalne analize kao jedine eksplanatorno vredne strategije u kognitivnoj nauci i usvajanjem eksplorativne strategije odnos između simbolicizma i konekcionizma se značajno menja, a većina kritika upućenih konkecionizmu na osnovu jakih ontoloških pretpostavki koje su sledile iz funkcionalne metodologije pada u vodu. Ukoliko se postkonekcionistički modeli koji sadrže duboke neuronske mreže shvate kao eksploratorni umesto teorijski, moguće je da se „igramo“ komponentama neuronske mreže i vrstama podataka kojim će se obučavati tako da istražimo kako se u svakoj od – milovski rečeno – varirajućih okolnosti ponaša. Na ovaj način bi se mogle testirati i nativistička i empiristička početna pretpostavka, tako što bi postale deo generičkog mehanizma, umesto da se unapred pretpostavljaju, a validacijom i interpretacijom P_1 se može ispitati korisnost svake od pretpostavki za potrebe modelovanja procesa ili obavljanja određenog zadatka.

Dalje, pripisujući eksplorativnu ulogu konekcionističkim modelima može se objasniti njihov dosadašnji napredak, a otvara se i prostor za njihov dalji razvoj. To što u svakom periodu, od Rozenblatovih perceptrona, preko pojave rekurentnih mreža devedesetih godina, pa do DKNM dvehiljaditih, dolazi do novih karakteristika i metodoloških prednosti ovih modela svedoči o tome da ne treba suditi o njihovim nedostacima *u principu*, već o njihovim *trenutnim* nedostacima, to jest da je inferencijalni skok sa trenutnih nedostataka na nedostatke *u principu* neopravdan. Kao što smo videli generalizacije u kritici konekcionizma poticale su ili iz nerazlikovanja kontingentih postavki koje se mogu menjati (recimo, broj slojeva, vrsta algoritma za učenje) od nužnih postavki bez kojih model ne bi mogao biti okarakterisan kao konekcionistički (recimo, relevantne strukturne sličnosti sa centralnim nervnim sistemom) ili su bile bazirane na ontološkim pretpostavkama o urođenosti i domenospecifičnosti, koje su zauzvrat branjene na osnovu metodoloških pretpostavki funkcionalne analize.

4.2. Pluralizam mehanicističkih objašnjenja

Metodologija koju branimo, a kojom treba da se udahne novi život (post)konekcionističkim modelima je eksplorativno mehanicistička. Stoga, kratko ćemo razmotriti da li se mehanicistička objašnjenja uopšte uklapaju u interpretaciju konekcionističkog teorijskog okvira. Behtel (William Bechtel) i Ričardson (Robert Richardson) (2000) navode sledeća svojstva mehanicističkih objašnjenja: **(i)** ova objašnjenja se odnose na ponašanje sistema referiranjem na funkcije koje obavljaju delovi sistema, kao i na interakcije između tih delova; **(ii)** heuristika za razvijanje ovih objašnjenja se sastoji u dekompoziciji sistema na delove i aktivnosti delova, što zauzvrat zahteva uvide iz

više naučnih disciplina. Može se videti da je svojstvo **(i)** u vezi sa tim kako se generalno interpretira ponašanje neuronske mreže – preko funkcije i interakcija jedinica u mnogostrukim slojevima. Svojstvo **(ii)** podseća dosta na obrazlaganje Stinsonove kako konekcionističke modele treba shvatiti kao da imaju eksplorativni status kojim se otkriva generički mehanizam.

Sledeći interpretaciju mehanizama uvedenih u (Machamer, Darden & Craver 2000: 3), a to je da postoji epistemički prelaz između nepotpune skice preko apstraktne sheme, pa do potpuno karakterizovanog mehanizma, možemo objasniti metodološki napredak konekcionističkih modela i predložiti dopunsko razmatranje njihovog eksplorativnog statusa. Naime, ako se konekcionistički modeli shvate kao skice, ili apstrakcije za koje se ne mogu u potpunosti eksplicirati entiteti i aktivnosti kojih se mehanizam tiče, i sadrže epistemičke „rupe" u pogledu hijerarhijskog odnosa između entiteta (Machamer, Darden & Craver 2000: 18), moguće je ostaviti prostora da se takve skice postepeno popunjavaju detaljima kognitivne i neuronauke (o odnosu između objašnjenja pomoću modela i mehanicističkih objašnjanja v. npr. Barberis 2013). Vremenom, konekcionistički modeli mogu napredovati do shema i potpuno karakterizovanih mehanizama, čime bi i objašnjenja koja pružaju bila bolje zasnovana. Ukoliko dalje napravimo razliku između kako-je-nešto-moguće, kako-je-nešto-plauzibilno, i kako-je-nešto-aktualno mehanicističkih objašnjenja (Craver 2007), možemo primetiti da konekcionistički modeli pružaju bar prve dve vrste objašnjenja. Ako se modelima dodeli eksplorativni status, konekcionista može ostati samo na nivou „igranja" komponentama modela, odnosno fokusiran na eksplorativnu ulogu *per se*, bez pretenzija na biološku plauzibilnost. U tom slučaju, konekcionistički modeli nam mogu pružiti kako-je-to-moguće objašnjenja, koja ne moraju biti u vezi sa detaljisanjem radi što vernije simulacije. Međutim, u nekim slučajevima, konekcionista može imati potrebu za inkorporiranjem bioloških ograničenja i stalo mu je do biološke plauzibilnosti, te samim tim eksplorativni modeli bi pružali kako-je-to-plauzibilno objašnjenje. Može se reći da je to slučaj sa modelima koji implementiraju DKNM – ovi modeli jesu biološki inspirisani vizuelnim korteksom mačke, efikasno simuliraju prepoznavanje slika do te mere da nekada prevazilaze i ljude u određenim zadacima. Međutim, budući da nam nije u potpunosti poznato zašto su DKNM uspešne, nije realno očekivati da mogu da pruže kako-je-nešto-aktualno objašnjenje.

Može se argumentovati da kognitivnim i neuronaučnicima nije ni potrebno da konekcionistički modeli pružaju kako-je-nešto-aktualno objašnjenja, budući da reprodukovanje i simuliranje nije isto što i objašnjavanje fenomena. Modeli treba da odraze strukturne aspekte sistema koji su relevantni za produkovanje fenomena koji treba objasniti. To dalje znači da je potrebno naći dovoljno precizan nivo apstrakcije na kom će biti moguće odraziti upravo samo relevantne strukturne aspekte, a reprodukcija se može ostaviti istraživačima na polju veštačke inteligencije, ukoliko imaju pretenzije na potpuno reprodukovanje ljudske inteligencije.

Zaključak

U ovom radu smo predstavile istorijat kritike konekcionizma, predložile jednu moguću konceptualizuju te kritike, kao i jedan način kako ona može da se odbaci i prevaziđe. Naime, iako se takva kritika najčešće predstavlja kao ontološka, ili kao sukob između nativističkih i empirističkih pristupa kogniciji, za nju se ispostavlja da počiva na određenim metodološkim pretpostavkama koje današnja filozofija dovodi u pitanje. Početna kritika Minskog i Paperta, može se reći, *predkonekcionističkih* modela, to jest perceptrona, zasnivala se na nedozvoljenom skoku od ukazivanja na nedostatke jednostavnih perceptrona na generalizovanje jednake neadekvatnosti kompleksnijih neuronskih mreža koje su od njih sastavljene. Kasnija kritika simbolista je sofisticiranija, ali primenjena u istom obliku i na *konekcionističke* i *postkonekcionističke* modele. Ona ukazuje na eksplanatorne nedostatke konekcionizma, to jest pokazuje da konekcionisti ne mogu da objasne suštinske kognitivne fenomene (poput sistematičnosti jezika) na zadovoljavajući način, i primorava konekcioniste koji prihvataju pretpostavku važnosti takvih fenomena da pribegnu stanovištu da konekcionizam pruža samo nekognitivnu teoriju implementacije simboličke arhitekture. Potom se ukazuje na nedostatke konekcionizma i kao teorije implementacije isticanjem biološki neplauzibilnih karakteristika različitih modela neuronskih mreža. Kako bi se teza simbolista ojačala navode se primeri neuspešnosti konkretnih modela u simuliranju relevantnih kognitivnih fenomena, koji se tumače kao da pokazuju da nijedan mehanizam, čija arhitektura ne inkorporira simboličke ontološke pretpostavke (postojanje kompozicionalno strukturiranih reprezentacija i urođenih sintaktičkih pravila) ne može uspešno da proizvede relevantne fenomene.

U radu smo nastojale da pokažemo da takva neuspešnost konekcionizma sledi samo pod pretpostavkama tradicionalne metodologije kognitivne psihologije koja insistira na funkcionalnoj analizi. Zahtevajući najpre da kognitivne sposobnosti budu objašnjene jednostavnijim sposobnostima od kojih je eksplanandum sačinjen, konekcionista je na kraju bio primoran da prizna da on takvo objašnjenje ipak ne može da pruži, jer se u njegovim objašnjenjima ne pojavljuju „ne sasvim kognitivni" homunkuli, već samo mehaničke operacije koje se vode matematičkim pravilnostima. Simbolisti su, nasuprot tome, na osnovu ove metodološke pretpostavke došli do ontoloških tvrdnji putem kojih su nastojali da dokažu da svaki empiristički pokušaj kognitivnih objašnjenja mora propasti. Napokon, iskrivljavanje pozicije konekcionizma u kognitivnoj nauci kao implementacione teorije i njegova nemogućnost da pruži sasvim efikasne mehanizme koji proizvode kognitivne fenomene doveli su do potpune marginalizacije ovog programa.

Odbacivanjem funkcionalne analize kao jedine ispravne metodologije u ispitivanju kognitivnog i psihološkog pokazale smo kako kritika konekcionizma postaje neopravdana i neuspešna. Usvajanjem eksplorativne strategije koja prihvata

mehanicistička kognitivna objašnjenja i viđenjem konekcionističkih modela kao nepotpunih skica mehanicističkih objašnjenja želele smo da razoružamo kritiku simbolista u sledećim pogledima:

(1) simbolista više ne može tvrditi da konekcionista ne može pružiti kognitivna objašnjenja, jer se dopuštaju objašnjenja koja nisu plod funkcionalne analize;
(2) simbolista više ne može insistirati na svojim nativističkim ontološkim pretpostavkama kao zasnovanim na neprikosnovenoj metodologiji, već ih mora opravdati na neki drugi način;
(3) simbolista više ne može kritikovati konekcionizam na osnovu biološke neplauzibilnosti, jer novousvojena strategija insistira samo na generičkim, ali ne i na konkretnim mehanizmima;
(4) simbolista više ne može samo ukazati na neadekvatnost nekog konkretnog (post)konekcionističkog modela kako bi kritikovao (post)konekcionizam uopšte, jer je status takvih modela eksplorativan i podložan izmeni relevantnih parametara.

Naš konačni zaključak je da je konekcionistički program bio nepravedno zapostavljen na osnovu nekritičkog prihvatanja metodologije čiji je glavni motiv izbegavanje „misterioznosti" psihološkog. Potreba da psihološke fenomene sagledavamo iz perspektive prvog lica dovela je do zapostavljanja jednog naučnog programa, koji prema našem mišljenju ima veliki potencijal za napredak, pogotovo ukoliko se upari sa teorijom dinamičkih sistema, kao i novim utelovljenim i proširenim pristupima kogniciji. Usvajanjem eksplorativne strategije i mehanicističkih objašnjenja u kognitivnoj nauci verujemo da će njegova uloga biti sve veća i značajnija, a njegova rezilijentnost tokom svih ovih godina, i tihi napredak uprkos oštrim kritikama povratno opravdavaju naše zauzimanje eksplorativne perspektive.

Miljana Milojević
Filozofski fakultet, Beograd
Vanja Subotić
Institut za filozofiju, Beograd

Literatura:

Abrahamsen, A. & Bechtel, W. (2006). Phenomena and Mechanisms: Putting the Symbolic, Connectionist, and Dynamical Systems Debate in Broader Perspective. In: R. Stainton (Ed.), *Contemporary debates in cognitive science*. Oxford: Basil Blackwell, 159-185.
Aizawa, K. (2003). *The Systematicity Arguments*. Kluwer Academic Publishers.

Barberis, S. D. (2013). Functional Analyses, Mechanistic Explanations, and Explanatory Tradeoffs. *Journal of Cognitive Science,* 14: 229-251.
Bartunov, S., Santoro, A., Richards, B. A., Hinton, G. E., & Lillicrap, T. (2018). Assessing the scalability of biologically-motivated deep learning algorithms and architectures. Preprint arXiv:1807.04587v2
Bechtel, W. & Richardson, R. (2000). *Discovering Complexity: Decomposition and Localization in Scientific Research*. Princeton University Press.
Berkeley, I. S. N. (2019). The Curious Case of Connectionism. *Open Philosophy* 2: 190-205.
Berko, J. (1958). The Child's Learning of English Morphology. *WORD*, 14 (2-3): 150-177.
Broadbent, D. (1985). A question of levels: Comment on McClelland and Rumelhart. *Journal of Experimental Psychology: General* 114 (2): 189-192.
Browne, A. (Ed.) (1997). *Neural Network Analysis, Architectures and Applications*. Philadelphia, PA: Institute of Physics,.
Buckner, C. (2018). Empiricism without Magic: Transformational Abstraction in Deep Convolutional Neural Networks. *Synthese* 195 (12): 5339-5372.
Buckner, C. & Garson, J. (2018). Connectionism and Post-connectionist Models. In: M. Sprevak, & M. Columbo (Eds.), *The Routledge Handbook of the Computational Mind.* Routledge University Press, 76-91.
Chalmers, D. (1990). Syntactic Transformations on Distributed Representations. *Connection Science* 2 (1&2): 53-62.
Chomsky, N. (1965). *Aspects of the Theory of Syntax*. The MIT Press.
Churchland, P. S. (1986). *Neurophilosophy: Toward A Unified Science of the Mind-Brain*. The MIT Press.
Churchland, P. S. & Sejnowski, T. J. (1990). Neural representation and neural computation. *Philosophical Perspectives* 4: 343-382.
Coram, A. (2014). Systematicity Laws and Explanatory Structures in the Extended Mind. In: P. Calvo & J. Symons, (Eds.). *The Architecture of Cognition: Rethinking Fodor and Pylyshyn's Systematicity Challenge*. The MIT Press, 277-304.
Cowie, F. (1999). *What's Within? Nativism Reconsidered*. Oxford University Press.
Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Clarendon Press.
Cummins, R. C. (1983). *The Nature of Psychological Explanation*. The MIT Press.
Dennett, D. (1991). *Consciousness Explained*. Penguin Books.
Figdor, C. (2018). The Fallacy of the Homuncular Fallacy. *Belgrade Philosophical Annual* 31: 41-56.
Fisher, G. (2006). The Autonomy of Models and Explanation: Anomalous Molecular Rearrangements in Early Twentieth-Century Physical Organic Chemistry. *Studies in History and Philosophy of Science Part A* 37 (4): 562-584.
Fodor, J. A. (1968). *Psychological Explanation: An Introduction To The Philosophy Of Psychology*. Random House.
Fodor, J. & Pylyshyn, Z. (1988). Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition* 28: 3-71.
Fodor, J. & McLaughlin, B. P. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition* 35 (2):183-205.

Frank, S. (2014). Getting Real about Systematicity", In: P. Calvo & J. Symons, (Eds.). *The Architecture of Cognition: Rethinking Fodor and Pylyshyn's Systematicity Challenge*. The MIT Press, 147-164.
Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* 36: 193-202
Gordon P. C., Hendrick R., & Johnson M. (2001). Memory Interference during Language Processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27 (6): 1411-1423.
Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., et al. (2009). A Novel Connectionist System for Improved Unconstrained Handwriting Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31 (5): 855-868
Gu, J., Wang, Z., Kuen, J. et al. (2017). Recent Advances in Convolutional Neural Networks. Preprint arXiv:1512.07108v6.
Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired Artificial Intelligence. *Neuron* 95 (2): 245-258.
Hauser, M. D., Chomsky, N., & Fitch, W. D. (2002). The Faculty of Language: What is it, Who has it, and How did it Evolve? *Nature* 98 (5598): 1569-1579.
Hebb, D. (1949). *The Organization of Behavior: A Neurophysiological Perspective*. Wiley.
Hinton, G. E. & McClelland, J. L. (1988) Learning Representations by Recirculation. In: D. Z. Anderson (Ed.), *Neural Information Processing Systems*. American Institute of Physics, 358-366.
Hochreiter, S. & Schmidhuber, J. (1997). LongShort-Term Memory. *Neural Computation* 9 (8): 1735-1780.
Hubel, D. H. & Wiesel, T. N. (1968). Receptive Fields and Functional Architecture of Monkey Striate Cortex. *The Journal of Physiology 195* (1): 215-243.
Kang, S. & Isik, C. (2005). Partially Connected Feedforward Neural Networks Structured by Input Types. *IEEE Transactions on Neural Networks* 16 (1): 175-184.
Kravitz D. J., Saleem K. S., Baker C. I., Ungerleider L. G., Mishkin M. (2013). The Ventral Visual Pathway: an Expanded Neural Framework for the Processing of Object Quality. *Trends in Cognitive Science* 17 (1): 26-49.
Lake, B. & Baroni, M. (2018). Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks. *OpenReview for the ICLR Conference*, https://openreview.net/pdf?id=H18WqugAb
LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep Learning. *Nature* 521: 436-444.
Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about Mechanisms. *Philosophy of Science* 67: 1-25.
Marcus, G., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven month-old infants. *Science* 283: 77-80.
Marcus, G. (2014). PDP and Symbol Manipulation: What's Been Learned Since 1986? In: P. Calvo & J. Symons, (Eds.). *The Architecture of Cognition: Rethinking Fodor and Pylyshyn's Systematicity Challenge*. The MIT Press, 103-114.
Marcus, G. (2018a). Deep Learning: A Critical Appraisal. Preprint arXiv:1801.00631
Marcus, G. (2018b). Innateness, AlphaZero, and AI. Preprint arXiv:1801.05667

Marr, D. (1982). *Vison*. W. H. Freeman.
Matthews, Robert J. (1997). Can Connectionists Explain Systematicity? *Mind and Language* 12 (2): 154-77.
McCulloch, W.S., Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics* 5: 115-133.
McLaughlin, B. (1993). The Connectionism/Classicism Battle to Win Souls. *Philosophical Studies* 71: 163-190.
Mill, J. S. (1843). *A System of Logic*. http://www.earlymoderntexts.com/authors/mill
Minsky, M., & Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. The MIT Press.
Mozer, M. C. & Smolensky, P. (1989). Using Relevance to Reduce Network Size Automatically. *Connection Science* 1: 3-16.
Olazaran, M. (1996). A Sociological Study of the Official History of the Perceptrons Controversy. *Social Studies of Science* 26 (3): 611-659.
Peña, M., Bonatti, L., Nespor, M., & Mehler, J. (2002). Signal-driven Computations in Speech Processing. *Science* 298: 604-607.
Piccinini, G. (2004). The First Computational Theory of Mind and Brain: A Close Look at McCulloch and Pitts' *Logical Calculus of Ideas Immanent in Nervous Activity*. *Synthese* 141 (2): 175-215.
Piccinini, G. & Craver, C. (2011). Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches. *Synthese* 183 (3): 283-311.
Priebe, N. J., Mechler, F., Carandini, M. et al. (2004). The Contribution of Spike Threshold to the Dichotomy of Cortical Simple and Complex Cells. *Nature Neuroscience* 7: 1113.
Prince, A. & Smolensky, P. (1993/2004). *Optimality Theory: Constraint interaction in Generative Grammar.* Blackwell Publishing.
Pylyshyn, Z. W. (1984). *Computation and Cognition*. The MIT Press.
Rogers, T. T., & McClelland, J. L. (2014). Parallel Distributed Processing at 25: Further Explorations in the Microstructure of Cognition. *Cognitive Science* 38 (6): 1024-1077.
Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review* 65: 386-408.
Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Cornell University Press.
Rumelhart, D., Hinton, G. & Williams, R. (1986). Learning Representations by Back- propagating Errors. *Nature* 323: 533-536.
Rumelhart, D. McClelland, J. L. & PDP Research Group (Eds.). (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations.* The MIT Press.
Schubbach, A. (2019). Judging Machines: Philosophical Aspects of Deep Learning. *Synthese* https://doi.org/10.1007/s11229-019-02167-z
Siegelmann, H. T. & Sontag, E. D. (1991). Turing Computability with Neural Nets. *Applied Mathematics Letters* 4 (6): 77-80
Smolensky, P. (1988). The Constituent Structure of Connectionist Mental States: A Reply to Fodor & Pylyshyn. *Southern Journal of Philosophy* XXVI: 137-162.

Stinson, C. (2018). Explanation and Connectionist Models. In M. Sprevak, & M. Colombo (Eds.), *The Routledge Handbook of the Computational Mind* Routledge University Press, 120-134.
Stinson, C. *forthcoming*. From Implausible Artificial Neurons to Idealized Cognitive Models:
Rebooting Philosophy of Artificial Intelligence.
Symons, J. & Calvo, P. (2014). Systematicity: An Overview. In: P. Calvo & J. Symons, (Eds.). *The Architecture of Cognition: Rethinking Fodor and Pylyshyn's Systematicity Challenge*. The MIT Press, 3-31.
Turing, A. M. (1936). On Computable Numbers, with an Application to the *Entscheidungsproblem*. *Proceedings of London Mathematical Society* 42 (2): 230-265. Corrections in (1937) *Proceedings of London Mathematical Society* 43 (2): 544-546.
Von Neumann, J. (1945/1993). *First Draft of a Report on the EDVAC*. *IEEE Annals of the History of Computing* 15 (4): 27-75.
Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New Horizons in Psychology, vol. 1*. Harmondsworth, UK Penguin, 135-151.
Williams, R. J. (1992). Simple Statistical Gradient-following Algorithms for Connectionist Reinforcement Learning. *Machine Learning* 8: 229-256.
Zednik, C. (2019). Solving the Black Box Problem: A General-Purpose Recipe for Explainable Artificial Intelligence. Preprint arXiv:1903.04361v2

Miljana Milojević
Vanja Subotić

The Exploratory Status of Postconnectionist Models
(*Summary*)

This paper aims to offer a new view of the role of connectionist models in the study of human cognition through the conceptualization of the history of connectionism – from the simplest perceptrons to convolutional neural nets based on deep learning techniques, as well as through the interpretation of criticism coming from symbolic cognitive science. Namely, the connectionist approach in cognitive science was the target of sharp criticism from the symbolists, which on several occasions caused its marginalization and almost complete abandonment of its assumptions in the study of cognition. Criticisms have mostly pointed to its explanatory inadequacy as a theory of cognition or to its biological implausibility as a theory of implementation, and critics often focused on specific shortcomings of some connectionist models and argued that they apply on connectionism in general. In this paper we want to show that both types of critique are based on the assumption that the only valid explanations in cognitive science are instances of homuncular functionalism and that by removing this assumption and by adopting an alternative methodology – exploratory mechanistic

strategy, we can reject most objections to connectionism as irrelevant, explain the progress of connectionist models despite their shortcomings and sketch the trajectory of their future development. By adopting mechanistic explanations and by criticizing functionalism, we will reject the objections of explanatory inadequacy, by characterizing connectionist models as generic rather than concrete mechanisms, we will reject the objections of biological implausibility, and by attributing the exploratory character to connectionist models we will show that practice of generalizing current to general failures of connectionism is unjustified.

KEYWORDS: connectionism, deep learning, exploration, mechanistic explanation, traditional symbolic cognitive science.