Uloga Marrovih razina objašnjenja u kognitivnim znanostima

Marko Jurjako*
mjurjako@ffri.uniri.hr
https://orcid.org/0000-0002-7252-8627

https://doi.org/10.31192/np.21.2.13
UDK: 159.9Marr, D.
165.194
Izvorni znanstveni rad /
Original scientific paper
Primljeno: 20. ožujka 2023.
Prihvaćeno: 10. svibnja 2023.

Ovaj rad razmatra pitanje može li se utjecajno razlikovanje između razina objašnjenja koje uvodi David Marr koristiti kao opći okvir za razmišljanje o razinama objašnjenja u kognitivnim znanostima i psihologiji. Marr je razlikovao tri razine na kojima možemo objašnjavati kognitivne procese: računalna, algoritamska i implementacijska razina. Neki tvrde da se Marrove razine objašnjenja poglavito mogu primjenjivati na modularne kognitivne sustave. Budući da su mnogi psihološki procesi nemodularni, čini se da Marrove razine objašnjenja ne mogu objasniti takve psihološke procese. U ovom radu se evaluira takva vrsta razmišljanja. Da bi se pokazalo da ovaj način razmišljanja nije uvjerljiv, u radu se prikazuje utjecajna paradigma iz kognitivnih znanosti koja se temelji na principu slobodne energije. Na temelju te paradigme, u radu se tvrdi da se čak i nemodularni psihološki procesi mogu uspješno analizirati iz računalne i algoritamske perspektive. Zaključak je rada da se, pod pretpostavkom da je funkcija uma minimiziranje slobodne energije, Marrov pristup razinama objašnjenja može uspješno primijeniti kao opći okvir za razumijevanje psiholoških procesa.

Ključne riječi: *aktivno zaključivanje, kognitivne znanosti, Marrove razine objašnjenja, prediktivno kodiranje, princip minimizacije slobodne energije.*

* Izv. prof. dr. sc. Marko Jurjako, Sveučilište u Rijeci, Filozofski fakultet, Odsjek za filozofiju i Katedra za kognitivne znanosti; Sveučilišna avenija 4, HR-51000 Rijeka.

Uvod

David Marr je ponudio jedno od najutjecajnijih gledišta na to kako analizirati psihološke fenomene.[1] On analizira psihološke fenomene iz tri perspektive: računalne, algoritamske i implementacijske. Obično se pretpostavlja da Marrova trodijelna analiza dobro zahvaća modularne psihološke fenomene, poput ranog vizualnog procesiranja, dok nemodularni fenomeni, poput sposobnosti za svjesno donošenje odluka, ostaju izvan eksplanatornog dohvata njegove analize.[2] U tom kontekstu, stječe se dojam da je Marr ponudio ograničenu perspektivu koja ne može zahvatiti sve fenomene koji se odnose na prirodu uma.

U ovom radu tvrdit će se da nema načelnih razloga zašto se Marrove razine psihološkog objašnjenja ne bi mogle koristiti kao temeljni okvir za razumijevanje funkcioniranja i modularnih i nemodularnih kognitivnih sustava. Konkretnije, argumentirat će se u prilog kondicionalne tvrdnje: ako se prihvati funkcionalistička slika uma, onda nema razloga smatrati da Marrova analiza ne može zahvatiti sve razine na kojima se mogu objašnjavati psihološki procesi.

Rad je podijeljen ovako: u prvom odjeljku predstavit će se Marrova podjela na tri razine objašnjenja u psihologiji; u drugom se odjeljku razmatraju razlozi za tvrdnju da se Marrove razine objašnjenja ne mogu primijeniti na um u cjelini; u trećem se odjeljku uvodi paradigma iz kognitivnih znanosti koja se temelji na principu slobodne energije. U nastavku se argumentira da, ako se um shvati kao sustav koji nastoji minimizirati slobodnu energiju, tada se Marrove razine objašnjenja mogu uspješno primijeniti na um u cjelini. U četvrtom odjeljku se razmatra nekoliko prigovora dosadašnjoj raspravi te se navode odgovori na njih.

1. Marrove razine objašnjenja

Svaki sustav za obrađivanje informacija može se analizirati na računalnoj, algoritamskoj i implementacijskoj razini.[3] Razliku između ove tri razine možemo razmotriti na sljedećem primjeru. Marr se bavio istraživanjem vizualnog sustava i načinom kako on omogućava percepciju predmeta u okolini. Na računalnoj razini se razmišlja o funkciji, tj. o kognitivnom zadatku koji vizualni sustav obavlja. Marr je smatrao da je funkcija *ranog* vizualnog procesiranja formiranje objektivne (tj. o promatraču neovisne) 3D predodžbe predmeta u okolini. Konkretnije, funkcija ranog vizualnog procesiranja jest, na temelju svjetlosti koja se odbija od vanjskih predmeta i pada na retinu oka, proizvesti mentalnu

[1] Usp. David MARR, *Vision. A computational investigation into the human representation and processing of visual information*, San Francisco, W. H. Freeman, 1982.

[2] Usp. José Luis BERMÚDEZ, *Philosophy of psychology. A contemporary introduction*, London, Routledge, 2005, poglavlje 2.

[3] Usp. Marr, *Vision...*

predodžbu tog predmeta. Na algoritamskoj razini nastoji se podrobnije odrediti, koje bi bile reprezentacije (ili mentalne predodžbe) i računalne procedure (tj. algoritmi) koji njima barataju, a kao ishod imaju stvaranje 3D predodžbe predmeta. Na primjer, na ovoj razini će se govoriti o algoritmima koji na temelju svjetlosnih zraka koje padaju na retinu oka mogu detektirati rubove predmeta te daljnjim obrađivanjem informacija mogu derivirati informacije o veličini, obliku i položaju predmeta. Konačno, na implementacijskoj razini istražuju se neurološki korelati vizualnog procesiranja. Kod ljudi značajnu ulogu u procesiranju vizualnih podražaja imaju procesi u okcipitalnom režnju mozga.[4]

Marr je predstavio te tri razine objašnjenja kao relativno autonomne. Na primjer, može se razmišljati o funkciji nekog kognitivnog sustava, čak i u nedostatku znanja koji algoritmi ga implementiraju. Slično tome, kognitivni znanstvenici mogu osmisliti apstraktni matematički model, tj. algoritam koji simulira ili predočava neki kognitivni proces, a da ne znaju nužno gdje se on točno u mozgu odvija.

Unatoč tome, postoji konsenzus da su Marrove tri razine objašnjenja međusobno povezane, barem u smislu da postavljaju ograničenja što će biti uvjerljivo objašnjenje na određenoj razini. Na primjer, ako je funkcija nekog kognitivnog sustava detekcija predatora u okolini, onda se može očekivati da će algoritam koji implementira tu funkciju biti takav da izvršava svoje operacije relativno brzo u vremenu.[5] Algoritam koji presporo omogućava detekciju predatora neće organizmu omogućiti da dugo preživi u svojoj okolini. Slično tome, poznavanje neurofiziologije i anatomije mozga može postaviti ograničenja ograničenja na koji način se smije razmišljati o kognitivnim sposobnostima na računalnoj i algoritamskoj razini. Na primjer, sam Marr je došao do zaključka da je funkcija ranog vizualnog procesiranja stvaranje 3D predodžbe predmeta na temelju neuropsiholoških istraživanja mozga. Naime, istraživanja su pokazala da pacijenti koji imaju oštećenja na dijelovima parijetalnog režnja mogu prepoznati predmete samo ako ih gledaju iz neke perspektive. Na temelju toga Marr je zaključio da funkcija ranog vizualnog procesiranja mora biti stvaranje objektivne 3D predodžbe predmeta (tj. predodžbe predmeta koja ne ovisi o položaju promatrača).[6]

Iz toga je jasno da Marrova trodijelna analiza može objasniti funkcioniranje kognitivnih sustava koji se mogu relativno lako identificirati i čija se funkcija može jasno odrediti. Međutim, ostaje pitanje može li se Marrovo gledište uspješno primijeniti na sve psihološke fenomene tako da ih se analizira iz računalne, algoritamske i implementacijske perspektive? U daljnjem tekstu raz-

[4] Za detaljnije objašnjenje Marrove trodijelne analize usp. Bermúdez, *Philosophy of psychology*…, odjeljak 2.1.

[5] Usp. *isto*, 26.

[6] Usp. José Luis BERMÚDEZ, *Cognitive science: an introduction to the science of the mind*, Cambridge University Press, 2014, https://doi.org/10.1017/CBO9781107279889, odjeljak 2.3.

motrit će se neki razlozi koji ukazuju na to da bi odgovor na ovo pitanje mogao biti negativan.

2. Općenitost Marrovih razina objašnjenja

Marrove razine objašnjenja najbolje zahvaćaju funkcioniranje podosobnih procesa koji imaju jasno definirane funkcije. Za potrebe ovog rada, podosobne procese možemo shvatiti kao one kognitivne procese koji se odvijaju ispod razine svijesti.[7] Osobna razina bi bila ona koja se odnosi na svjesno funkcioniranje osobe te uključuje obrasce psihološkog objašnjenja koji se temelje na pripisivanju zdravorazumskih mentalnih stanja, poput vjerovanja, želja i namjera.[8] Marrova analiza ranog vizualnog procesiranja tipičan je podosobni sustav koji funkcionira ispod razine svijesti i ima jasno određene funkcije. Taj sustav u konačnici proizvodi predodžbe predmeta kojih smo svjesni, međutim, sam proces kojim se realiziraju predodžbe se odvija ispod razine svijesti. Štoviše, ono čega su ljudi najčešće svjesni je predmet koji se nalazi u njihovoj okolini, a ne predodžba tog predmeta.

Sustavi koji imaju jasno definirane funkcije su obično modularni. Modularni sustavi imaju više karakteristika, te se gledišta razlikuju ovisno o tome koje karakteristike se naglašavaju.[9] Međutim, obično se minimalno pretpostavlja da modularne sustave karakterizira specifičnost domene primjene. Modularni kognitivni sustavi su zaduženi za procesiranje nekih tipova informacija (specifičnih za neku domenu funkcioniranja) koje se transformiraju i šalju u druge sustave na daljnje obrađivanje ili uzrokuju motorne radnje. Marrov primjer ranog vizualnog procesiranja tipičan je primjer takvog sustava. Oko je specijalizirano da prima svjetlosne podražaje iz kojih vizualni sustav izvodi informacije o svojstvima predmeta izvan osobe te u konačnici dovodi do formiranja predodžbe predmeta. Ta predodžba se dalje šalje u druge kognitivne sustave koji je koriste za svoje svrhe. Na primjer, osoba može koristiti svoju maštu da manipulira predodžbom predmeta da bi izvela neki zaključak o njemu ili razmišljati o strategiji za rukovanje predmetom ako odluči nešto s njim poduzeti.

Marrove razine analize dobro objašnjavaju funkcioniranje modularnih sustava.[10] Budući da modularni sustavi imaju jasnu domenu primjene, onda se

[7] Za neke od razloga zašto to nije dobar općeniti princip za razlikovanje osobnih od podosobnih stanja, usp. Bermúdez, *Philosophy of Psychology*…, 30.

[8] Za raspravu usp. Mason WESTFALL, Constructing persons. On the personal–subpersonal distinction, *Philosophical Psychology*, (2022) 1-30, https://doi.org/10.1080/09515089.2022.2096431; Zoe DRAYSON, The personal/subpersonal distinction, *Philosophy Compass*, 9 (2014) 5, 338-346, https://doi.org/10.1111/phc3.12124.

[9] Usp. Daniel WEISKOPF, Frederick ADAMS, *An introduction to the philosophy of psychology*, Cambridge, Cambridge University Press, 2015, poglavlje 3.

[10] Usp. Bermúdez, *Philosophy of psychology*…, 24-25.

očekuje da će imati funkciju koja se može jasno specificirati. Jednom kada je funkcija kognitivnog sustava od interesa specificirana, može se razmišljati o tome koji algoritam najbolje opisuje način na koji taj sustav funkcionira kod ljudi. Nakon toga se može provjeriti kako je taj sustav implementiran u mozgu i drugim fizičkim supstratima. No, postavlja se pitanje, može li se Marrova analiza primijeniti i na kognitivne procese za koje se smatra da nisu modularni?

Postoje razlozi za mišljenje da se nemodularni sustavi ne mogu uspješno objašnjavati Marrovom trodijelnom analizom. U tom pogledu, José Bermúdez objašnjava da su nemodularni kognitivni sustavi oni koji nemaju jasno specificiranu domenu primjene.[11] Tipični primjeri takvih sustava uključuju kognitivne procese koji se odvijaju na osobnoj ili svjesnoj razini funkcioniranja. Na primjer, ljudske racionalne sposobnosti nisu modularne jer nemaju neku posebnu domenu primjene. Štoviše, one su globalne jer se mogu primijeniti na bilo koji sadržaj misli, odluka i – u konačnici – omogućavaju različite tipove djelovanja. Omogućavaju razmišljanje o svakodnevnim stvarima, od toga što ćemo doručkovati ujutro do rješavanja logičkih teorema i razmišljanja o tome koji je smisao života. S obzirom na nepregledan raspon domena na koje se racionalne sposobnosti mogu primijeniti, nije jasno mogu li se one funkcionalno kodificirati, a kamoli da se može osmisliti algoritam koji bi mogao implementirati njihovu funkciju.[12] Štoviše, Bermúdez naglašava da je jedan od osnovnih ciljeva psihologije razumjeti kako um kao cjelina funkcionira.[13] Budući da um kao cjelina posjeduje sposobnosti koje nisu modularne, onda se očekuje da se Marrova analiza neće moći koristiti za objašnjavanje i razumijevanje velikog broja činjenica koje se odnose na um i mentalne pojave.

To gledište na neki način seže barem do Renéa Descartesa koji je smatrao da strojevi i životinje nemaju umove. On je argumentirao da, koliko god strojevi bili sofisticirani, uvijek ćemo ih moći razlikovati od ljudi prema ograničenosti njihovih kognitivnih sposobnosti:

> »(…) iako bi oni [strojevi ili životinje] činili mnoge stvari isto tako dobro ili možda bolje nego itko od nas, oni bi zacijelo otkazali u nekim drugima, uslijed čega bi se otkrilo, da ne postupaju svjesno, već samo zbog takvog rasporeda svojih organa. Dok je naime um opće oruđe, koje može služiti u svim mogućim prilikama, ovi organi moraju biti za svaku pojedinačnu radnju i na poseban način udešeni. Stoga je moralno (praktično) nemoguće, da bi bilo dovoljno različitih organa u jednom stroju, da postupa u svim slučajevima u životu na isti način kao što postupamo mi, jer imamo um.«[14]

[11] Usp. *isto*, 27.

[12] Usp. Hilary PUTNAM, Računarska psihologija i teorija interpretacije, u: Nenad Miščević, Nenad Smokrović (ur.), *Računala, mozak i ljudski um*, prev. Nenad Smokrović, Rijeka, Izdavački centar Rijeka, 2001, 154-169.

[13] Usp. Bermúdez, *Philosophy of Psychology*…, 27.

[14] René DESCARTES, *Rasprava o metodi*, prev. Niko Berus, Zagreb, Matica hrvatska, 1951, 47 prema Marko JURJAKO, Luca MALATESTI, *Filozofija uma. Suvremene rasprave o odnosu uma i tijela*, Rijeka, Filozofski fakultet Sveučilišta u Rijeci, 2022, 22.

Suvremenim rječnikom, prema Descartesu, um je nemodularno opće oruđe koje se može primijeniti na neprebrojivo mnogo situacija. Takve sposobnosti uma nije moguće rekreirati kroz modularne sustave (tj. *organe*) od kojih bi svaki bio zadužen za obavljanje nekog kognitivnog zadatka. Naime, s obzirom na gotovo beskrajne mogućnosti uma da kontemplira različite sadržaje, smišlja pojmove, donosi zaključke i upravlja voljom, nije moguće napraviti ogroman broj modularnih sustava koji bi zadovoljili sve te funkcije uma.

Iako prethodna razmatranja djeluju uvjerljivo, u daljnjem tekstu će se tvrditi da Marrova analiza također može doprinijeti boljem razumijevanju nemodularnih kognitivnih procesa. Prvo, bit će prikazano da se neke nemodularne sposobnosti mogu uspješno objasniti putem Marrovih razina objašnjenja. Zatim, u sljedećem odjeljku, tvrdit će se da se u načelu Marrove razine objašnjena mogu uspješno primijeniti ne samo na pojedine nemodularne kognitivne sposobnosti, već i na um kao cjelinu.

Što se tiče prve tvrdnje, treba primijetiti da suvremeni dosezi u računalnoj neuroznanosti sugeriraju da se neke nemodularne sposobnosti ipak mogu funkcionalno i algoritmički specificirati. Dobar primjer za to su neuropsihološka istraživanja sposobnosti za donošenje odluka.[15] Na računalnoj razini, donošenje odluke se može definirati kao sposobnost za biranje radnji koje će uspješno doprinijeti ostvarenju unaprijed zadanih ciljeva. Upravo zbog svojeg nemodularnog karaktera, sposobnost za donošenje odluka je kompleksan psihološki konstrukt koji se može dalje specificirati i analizirati na različite faktore, ovisno o tome kako se ta sposobnost testira kroz različite bihevioralne paradigme. Unatoč toj kompleksnosti, istraživači u računalnoj neuroznanosti su razvili općenite algoritme kojima se ona može uspješno modelirati.

Jedna vrsta takvih algoritama se temelji na učenju putem potkrepljivanja (engl. *reinforcement learning models*). To je vrsta strojnog učenja koja modelira djelatnika čiji je cilj naučiti koji niz radnji (tj. usvajanje kojeg plana) će maksimizirati kumulativnu nagradu koju može steći tijekom nekog vremenskog perioda. Takvi algoritmi simuliraju ponašanje djelatnika koji istražuje svoju okolinu te na temelju primljene nagrade ili kazne ažurira plan djelovanja da bi poboljšao svoju izvedbu u budućim interakcijama s okolinom.

Algoritmi učenja putem potkrepljivanja imaju zanimljivu povijest. Njihov je razvoj sinergija matematičkih istraživanja o optimalnom planiranju i empirijski utemeljenih psiholoških modela učenja kojima su se nastojali kvantitativno objasniti različiti aspekti klasičnog i operantnog uvjetovanja. U novije vrijeme, osim za potrebe modeliranja ljudskog donošenja odluka, takvi algoritmi

[15] Usp. Peter DAYAN, Nathaniel D. DAW, Decision theory, reinforcement learning, and the brain, *Cognitive, Affective, & Behavioral Neuroscience* 8, (2008) 4, 429-453.

se koriste za rješenja širokog raspona zadataka, igranja video igara, te razvoja robotike i autonomnih vozila.[16]

Važno za kontekst ovoga rada jest to što korištenje takvih modela jasno ilustrira kako se Marrove razine objašnjenja mogu uspješno koristiti za modeliranje nemodularnih kognitivnih sposobnosti. Već je spomenuto da se sposobnost za donošenje odluka može na računalnoj razini definirati kao sposobnost za biranje radnji koje će zadovoljiti zadane ciljeve. Tako definirana funkcija odlučivanja se može uspješno modelirati korištenjem različitih algoritama koji se temelje na učenju potkrepljivanjem. Nadalje, korištenje takvih modela omogućava da se biheviorални podaci na kvantitativno precizan način povežu s aktivacijama dijelova mozga koji se nalaze u podlozi donošenja odluka. U tom pogledu vrijedi istaknuti da je učenje putem potkrepljivanja korelirano s karakterističnim aktivacijama dijela mozga koji se naziva *striatum* te da je važan aspekt tog procesa kodiran u reakcijama dopaminskih neurona.[17]

Međutim, čak i ako se dopusti da se neke nemodularne kognitivne sposobnosti mogu uspješno analizirati putem Marrovih razina objašnjenja, i dalje ostaje otvoreno pitanje može li se um kao cjelina analizirati na sličan način. U nastavku će se braniti pozitivan odgovor na ovo pitanje. Da bi se to postiglo bit će razmotrena jedna utjecajna paradigma iz kognitivnih (neuro)znanosti koja funkcioniranje svih kognitivnih procesa objašnjava kao minimiziranje slobodne energije.

3. Princip slobodne energije i Marrove razine objašnjenja

Modeli uma ili mozga koji se temelje na principu slobodne energije (skraćeno PSE) pripadaju skupini modela koji se nazivaju bayesijanska teorija mozga.[18] Karakteristika koja je zajednička bayesijanskim teorijama u kognitivnim znanostima je tumačenje kognitivnih procesa kroz prizmu probabilističkih modela. Unutar te skupine probabilističkih modela, posebnost PSE-a je u tvrdnji da je osnovna funkcija svih kognitivnih procesa, uključujući uma shvaćenog kao ukupnost svih mentalnih događaja koji karakteriziraju neku osobu, minimizacija veličine koja se naziva slobodna energija.[19] Pojam slobodna energija dolazi iz fizike, međutim, u ovom kontekstu označava gornju granicu informacijsko-teorijske veličine koja se naziva »iznenađenje« (engl. *surprisal*). U ovom

[16] Usp. Richard S. SUTTON, Andrew BARTO, *Reinforcement learning. An introduction*, Cambridge, MA London, The MIT Press, ²2018.

[17] Usp. Tom SCHÖNBERG i dr., Reinforcement learning signals in the human Striatum distinguish learners from nonlearners during reward-based decision making, *The Journal of Neuroscience*, 27 (2007) 47, 12860-12867.

[18] Usp. Thomas PARR, Giovanni PEZZULO, Karl J. FRISTON, *Active inference. The free energy principle in mind, brain, and behavior*, Cambridge, Mass., The MIT Press, 2022.

[19] Za detaljni pregled ove paradigme, usp. *isto*.

kontekstu, »iznenađenje« označava stupanj neočekivanosti nekog događaja te se definira kao negativan logaritam vjerojatnosti ostvarivanja tog događaja. Što je vjerojatnost nekog događaja manja to je veće iznenađenje kada se on dogodi i obrnuto. Budući da je slobodna energija gornja granica iznenađenja nekog događaja, slijedi da se minimiziranjem slobodne energije ujedno minimizira iznenađenje nekog događaja.

Osnovna je ideja PSE-a da svi organizmi imaju tendenciju minimizirati vjerojatnost događaje koji su iz perspektive njihovog fenotipa iznenađujući (tj. imaju visoku slobodnu energiju) te je cilj ostati u poznatim stanjima (tj. stanjima koja imaju nisku slobodnu energiju). Tipičan primjer koji se koristi za ilustraciju ove opće ideje je riba izvan vode.[20] Normalan, odnosno poznat i očekivan okoliš za ribu jest život u vodi. To znači da bi bilo vrlo iznenađujuće za ribu da se nađe izvan vode. Da bi riba minimizirala iznenađenje, a time ujedno maksimizirala šanse za preživljavanje, mora poduzeti one radnje koje će joj omogućiti da ostane u okolišima koji su očekivani, odnosno imaju nisku slobodnu energiju s obzirom na njezin fenotip.

Prema principu slobodne energije, svi živi organizmi imaju sklonost odupirati se neredu (tj. entropiji) i minimizirati atipične ili iznenađujuće događaje u svom okolišu. S obzirom na općenitost PSE-a, u sklopu njegova pojmovnog aparata se jednostavno može objasniti adaptivno djelovanje i uloga pojedinih kognitivnih procesa u svemu tome. Polazi se od ideje da organizmi nastanjuju životno nesigurne okoliše.[21] Da bi se održali na životu moraju imati sposobnosti za adaptivno djelovanje. U temeljima adaptivnog djelovanja su sposobnosti percepcije stanja okoline i djelovanje na temelju tih percepcija. Prema PSE-u, djelovanje i percepcija su dvije strane istog novčića. Naime, jedna i druga sposobnost minimiziraju slobodnu energiju da bi organizam ostao u poznatim stanjima i okolišima. Organizmi to postižu poznavanjem modela okoline kojeg mogu ažurirati s novim percepcijama. Ideja modela u ovom kontekstu se različito shvaća. Neki imaju više kognitivističko shvaćanje modela kao mentalnih apstrakcija koje omogućuju probabilističko zaključivanje o procesima u okolini i unutrašnjosti organizma,[22] dok drugi shvaćaju modele kao doslovno utjelovljene u tjelesnim i neurofiziološkim fenotipovima koji karakteriziraju različite vrste organizama.[23] U svakom slučaju, zajednička je ideja da, kada se slobodna energija minimizira ažuriranjem modela koje mozgovi ili organizmi imaju o okolini, taj se proces naziva *perceptivno* zaključivanje. Alternativno, umjesto ažuriranja modela, slobodna energija se može minimizirati tako da se okolina uskladi s očekivanjima koja su utjelovljena u modelu organizma. U tom slučaju

[20] Npr. usp. Christopher L. BUCKLEY i dr., The free energy principle for action and perception. A mathematical review, *Journal of Mathematical Psychology*, 81 (2017) 55-79, https://doi.org/10.1016/j.jmp.2017.09.004.
[21] Npr., usp. Jakob HOHWY, *The predictive mind*, Oxford, Oxford University Press, 2013.
[22] Npr. *isto.*
[23] Usp. Parr, Pezzulo, Friston, *Active inference*…

govori se o *aktivnom* zaključivanju jer, umjesto ažuriranja modela na temelju percepcije, organizam zaključuje u kakvom stanju bi se trebao nalaziti s obzirom na svoj model okoline te aktivno djeluje da promjeni i uskladi okolinu sa svojim modelom.

Ako se prihvati ovakvo općenito gledište na kognitivne procese, onda postaje jasnije da možemo govoriti o umu kao cjelini u sklopu Marrove trodijelne analize. Na računalnoj razini funkcija uma se određuje kao minimizacija slobodne energije. Nadalje, PSE je kompatibilan s više različitih algoritama pomoću kojih se može modelirati perceptivne i akcijske procese koji omogućavaju minimizaciju slobodne energije.[24] Na primjer, jedna široko korištena skupina algoritama se temelji na ideji da se slobodna energija može minimizirati kroz proces minimizacije prediktivne pogreške.[25] Ta skupina algoritama se naziva prediktivno kodiranje. Prema ovoj ideji mozak ima hijerarhijski generativni model kojega čine prior ili prethodna vjerovanja (engl. *prior beliefs*) i vjerojatnost (engl. *likelihood*) – čija uloga je reprezentirati okolinu i uzroke koji proizvode podražaje kojima je osoba izložena.[26] Prethodna vjerovanja su u osnovi predviđanja ili hipoteze mozga o tome što bi moglo uzrokovati podražaje, dok se vjerojatnosti (u smislu *likelihooda*) odnose na vjerojatnost pojavljivanja podražaja koji pružaju dokaze o tome što se nalazi u okolini. Generativni model se ažurira minimiziranjem prediktivne pogreške koja se generira kada postoji razlika između predviđanja i stvarnih signala, tj. podražaja kojima je osoba izložena.

Ovaj proces se može jasnije prikazati uz pomoć primjera. Zamislite da hodate kroz šumu i odjednom čujete šuškanje u grmlju. Vaš mozak, na temelju priora, tj. prethodnih vjerovanja o okolini u kojoj se nalazite, nastoji odrediti vjerojatnost hipoteze o tome što je prouzrokovalo šuškanje. Na primjer, ako se nalazite negdje u šumi u Gorskom kotaru, vaš mozak bi mogao generirati predviđanje da je medvjed uzrokovao šuškanje. Pretpostavimo sada da osim šuškanja čujete lavež. Taj novi podražaj dovodi do prediktivne pogreške jer kada bi se u grmu nalazio medvjed, onda ne biste čuli lajanje već neke druge zvukove. Na temelju te prediktivne pogreške vaš mozak ažurira vjerovanja tako što smanjuje vjerojatnost hipoteze da se medvjed nalazi iza grma te povećava

[24] Usp. Ryan SMITH, Maxwell J. D. RAMSTEAD, Alex KIEFER, Active inference models do not contradict folk psychology, *Synthese*, 200 (2022) 2, 81, https://doi.org/10.1007/s11229-022-03480-w.

[25] Hohwy, *The predictive mind*…

[26] Na hrvatskom terminologija nije dovoljno jasna. Naime, osim »vjerojatnosti« u smislu *likelihooda*, prethodna vjerovanja su također vjerojatnosti. Terminologija se temelji na Bayesovom teoremu koji se koristi za izračunavanje vjerojatnosti nekog događaja s obzirom na dostupne dokaze ili informacije. U tom kontekstu, prethodna vjerovanja su vjerojatnosti nekog događaja koje čine hipotezu koju osoba razmatra prije nego joj nova dokazna građa postane dostupna (formalno zapisano P(H)), dok su *likelhood-ovi* uvjetne vjerojatnosti koje čine vjerojatnost da će osoba opservirati dokaznu građu s obzirom na hipotezu koju razmatra (formalno zapisano P(d | H)). Za više o odnosu Bayesovog teorema, principa slobodne energije i prediktivnog kodiranja, usp. Hohwy, *The predictive mind*…

vjerojatnost hipotezi da se tamo nalazi pas. Naime, hipoteza da se pas nalazi iza grma bolje objašnjava dostupnu dokaznu građu od alternativne hipoteze da se tamo nalazi medvjed, te na taj način adekvatnije minimizira prediktivnu pogrešku.

Formalniji prikaz osnovnih elemenata algoritma koji se temelji na prediktivnom kodiranju uključuje sljedeće komponente:

1) $b_n(c) = b_{n-1}(c) + p(s - b_{n-1}(c))$[27]

Ovdje *s* označava senzorni signal (tj. podražaj); $b_n(c)$ označava vjerovanje ili predviđanje da je *c* uzrok od *s*; $b_{n-1}(c)$ označava prethodno vjerovanje ili predviđanje da je *c* uzrok od *s*. Simboli u zagradi $(s - b_{n-1}(c))$ čine prediktivnu pogrešku. Broj *p* označava koeficijent težine koji određuje utjecaj prediktivne pogreške na ažuriranje vjerovanja.[28] U ovom kontekstu, *p* se shvaća kao preciznost prediktivne pogreške. Ona mjeri pouzdanost signala ili prethodnog vjerovanja. Ideja je da se nova predviđanja $b_n(c)$ trebaju formirati ovisno o prethodnom vjerovanju koje mozak ima o relevantnim uzrocima i prediktivnim pogreškama koje su ponderirane njihovom preciznošću. Dakle, što je senzorni signal neprecizniji, trebao bi imati manje utjecaja na ažuriranje vjerovanja. Na primjer, ako vozimo po magli, onda možemo očekivati da će preciznost vizualne percepcije biti niža. Stoga, u takvim okolnostima, preporučljivo je više se oslanjati na prethodna vjerovanja o tome što možemo očekivati na cesti.

Poveznica PSE-a i minimiziranja prediktivne pogreške jest što je potonji jedan od biološki uvjerljivih modela/algoritama kako kognitivni sustavi mogu minimizirati slobodnu energiju.[29] Da bi organizam preživio u okolini on mora adaptivno djelovati (tj. minimizirati slobodnu energiju) uz što manji utrošak (fizičke) energije. U tom smislu, osnovna je ideja da je prediktivno kodiranje energetski učinkovit način za procesiranje podataka. Naime, ako je okoliš u skladu s prethodnim predviđanjima, onda mozak nema potrebe dalje procesirati podatke. Samo u slučaju prediktivne pogreške (tj. kada su narušena predviđanja), mozak šalje signale u hijerarhijski gornje slojeve mozga da bi ažurirao svoje modele okoline. Dakle, umjesto da troši energiju reagiranjem na svaki dolazni podražaj, mozak obrađuje samo one podražaje koji nisu u skladu s njegovim prethodnim vjerovanjima i time na energetski učinkovit način omogućuje adaptivno djelovanje.

[27] Ovo je simplificirana verzija pravila ažuriranja koja, uz dodavanje hijerarhijskih slojeva, može implementirati hijerarhijsko bayesijansko zaključivanje. Za tehničke detalje, vidi npr. Christoph D. MATHYS i dr., A Bayesian foundation for individual learning under uncertainty, *Frontiers in Human Neuroscience*, 5 (2011), https://doi.org/10.3389/fnhum.2011.00039.

[28] Ova jednadžba je strukturalno slična modelima koji se temelje na učenju putem potkrepljivanja, gdje ulogu *p*-a igra stopa učenja koja se obično označava s *w*. Za više, vidi Sutton, Barto, *Reinforcement learning…*

[29] Usp. Jakob HOHWY, New directions in predictive processing, *Mind & Language*, 35 (2020) 2, 209-223, https://doi.org/10.1111/mila.12281.

Nadalje, u okviru algoritma prediktivnog kodiranja, aktivno zaključivanje i perceptivno zaključivanje postaju dva načina na koji se prediktivna pogreška može minimizirati, koji u biti odgovaraju dvama načinima na koje se slobodna energija može minimizirati. Na primjer, prediktivnu pogrešku da se čaša nalazi ispred osobe O, može se minimizirati tako da O promjeni vjerovanje o položaju čaše (na primjer, umjesto da vjeruje da je čaša ispred njega, O može ažurirati vjerovanje da se nalazi sa strane) ili tako da O aktivno djeluje i stavi čašu ispred sebe te time uskladi svijet sa svojim mentalnim stanjima.

Na razini implementacije, postoje istraživanja kojima se nastoje odrediti anatomski dijelovi i neurofiziološki procesi koji su potencijalni kandidati za implementaciju algoritama poput onih koji se temelje na minimizaciji prediktivne pogreške.[30] Ovdje se neće ulaziti u detalje implementacije, dovoljno je naglasiti da, s obzirom na općenitost paradigme koja se temelji na PSE-u, Marrova se analiza može na plodan način primijeniti da bi se rasvijetlili fenomeni koji karakteriziraju funkcioniranje uma na različitim razinama opisa.

4. Neki prigovori

Dosadašnjoj raspravi bi se moglo prigovoriti da pretpostavlja valjanost paradigme iz kognitivnih (neuro)znanosti koja je tek u povojima i za koju nije jasno da može zahvatiti sve važne aspekte ljudske psihe.[31] Na primjer, neki ukazuju na to da prediktivno kodiranje nije analitički rješivo, u smislu da nije jasno da je to algoritam koji može u razumnom vremenu obaviti izračune koji bi omogućili adaptivno ponašanje i rješavanje kognitivnih zadataka s kojima se organizmi svakodnevno suočavaju.[32] Budući da PSE svodi djelovanje i percipiranje na minimizaciju slobodne energije, drugi prigovaraju onda da nije u mogućnosti napraviti jasnu razliku između motivacijskih stanja, poput želja, i spoznajnih stanja, poput vjerovanja.[33] Naime, zdravorazumska perspektiva na djelovanje pretpostavlja da se namjerne radnje temelje na željama i vjerovanjima koja imaju funkcionalno jasno odvojene uloge.[34] Problem je u ovom kontekstu u tome što neki tvrde da PSE ne može zahvatiti zdravorazumsko psihološku ontologiju mentalnih stanja koja pravi jasnu razliku između motivacijskih i spoznajnih

[30] Usp. Karl FRISTON i dr., The anatomy of choice: active inference and agency, *Frontiers in Human Neuroscience*, 7 (2013), https://doi.org/10.3389/fnhum.2013.00598.

[31] Vidi npr. Marko JURJAKO, Can predictive processing explain self-deception?, *Synthese*, 200 (2022) 303, https://doi.org/10.1007/s11229-022-03797-6.

[32] Usp. Johan KWISTHOUT, Iris van ROOIJ, Computational resource demands of a predictive Bayesian brain, *Computational Brain & Behavior*, 3 (2020) 2, 174-188, https://doi.org/10.1007/s42113-019-00032-3

[33] Usp. Colin KLEIN, What do predictive coders want?, *Synthese*, 195 (2018) 6, 2541-2557, https://doi.org/10.1007/s11229-016-1250-6.

[34] Usp. npr. Matej SUŠNIK, Hjumovska teorija motivacije: u obranu dogme, *Prolegomena*, 11 (2012) 1, 83-105.

mentalnih stanja.[35] Ako je tako, onda se čini da će PSE imati problem objasniti intencionalno ljudsko djelovanje.[36]

Čak i ako se prihvati uvjerljivost prethodnih prigovora, on nisu presudni za argument koji se iznosi u ovom radu. Cilj ovog rada nije dokazati točnost svakog aspekta paradigme koja se temelji na PSE-u ili zadržati u svim aspektima zdravorazumsko-psihološko shvaćanje uma.[37] Umjesto toga, cilj je pokazati da se Marrove razine objašnjenja mogu primijeniti na um u cjelini. Stoga je za trenutne potrebe dovoljno pokazati da PSE daje obećavajući istraživački okvir koji je utjecajan u kognitivnim (neuro)znanostima te omogućava jasno razumijevanje da se Marrove razine objašnjenja mogu primijeniti na um u cjelini. Uzimajući u obzir taj kontekst, treba još primijetiti da, čak i ako je PSE u nekim aspektima revizionistički istraživački program, to ne umanjuje njegovu važnost za pokazivanje na koji način se Marrove razine objašnjenja mogu primijeniti za analizu uma u cjelini.

Drugi prigovor može biti da PSE podrazumijeva funkcionalističku sliku uma koja je i sama sporna u filozofiji uma.[38] Prema funkcionalizmu, mentalna stanja se definiraju na temelju uloge koju imaju u uzročnom povezivanju vanjskih podražaja, drugih unutarnjih stanja i ishoda, poput nekih ponašanja. Na primjer, bol se funkcionalistički može analizirati kao stanje koje nastaje kao posljedica oštećenja tkiva koje uzrokuje druga mentalna stanja, poput želje da se osoba odmakne od izvora boli, te u konačnici može izazvati neka ponašanja, poput odmicanja od izvora boli, trljanja dijela tijela koji boli i slično. Obično se funkcionalističkom shvaćanju uma prigovara neuspješnost objašnjenja prirode fenomenalne svijesti i kako to da su funkcionalne uloge povezane s nekim svjesnim iskustvima. Konkretnije, funkcionalisti se susreću s izazovom objašnjenja zašto oštećenje tkiva uzrokuje svjesno mentalno stanje koje ljudi prepoznaju kao bol, a ne, na primjer, kao škakljanje.[39] Iz same funkcionalne definicije boli

[35] Usp. Matteo COLOMBO, Social motivation in computational neuroscience: or if brains are prediction machines, then the Humean theory of motivation is false, u: Julian Kiverstein (ur.), *Routledge handbook of philosophy of the social mind*, 2017, poglavlje 18.

[36] Za odgovore na ovaj navodni problem za PSE, vidi Andy CLARK, Beyond desire? Agency, choice, and the predictive mind, *Australasian Journal of Philosophy*, 98 (2020) 1-15, https://doi.org/10.1080/00048402.2019.1602661; Smith, Ramstead, Kiefer, *Active inference models do not contradict folk psychology…*

[37] Što se tiče prigovora da PSE ne zahvaća zdravorazumsko-psihološku ontologiju mentalnih stanja, nije jasno koliko je on relevantan iz još nekih razloga. Naime, teorije kognitivnih znanosti ne moraju nužno opravdati ontološke podjele među tipovima mentalnih stanja kako se one shvaćaju na zdravorazumsko-psihološkoj razini. Drugim riječima, teorije iz kognitivnih znanosti mogu biti revizionističke u odnosu na zdravorazumsku psihologiju. Za raspravu, usp. Bermúdez, *Philosophy of Psychology…*, odjeljak 2.4 i poglavlje 5; Jurjako, *Can predictive processing...*

[38] Usp. npr. Davor PEĆNJAK, Tomislav JANOVIĆ, Nefunkcionalnost funkcionalizma, u: Maja Hudoletnjak Grgić, Filip Grgić, Davor Pećnjak (ur.), *Aspekti uma*, Zagreb, Institut za filozofiju, 2011, 1-18.

[39] Za recentnu raspravu, usp. Michal, POLÁK, Heat and pain identity statements and the imaginability argument, European journal of analytic philosophy 18, 2 (2022) (A1)5-31.

(ili bilo kojeg drugog mentalnog stanja) čini se da ne slijedi da bi ono trebalo biti popraćeno svjesnim iskustvom.[40]

Iako se može raspravljati o filozofskoj kontroverznosti funkcionalističke slike uma, ta kritika nije izravno relevantna za trenutnu raspravu. Naime, sama Marrova trodijelna analiza podrazumijeva funkcionalističko shvaćanje uma. Da bi se je moglo primijeniti na ljudske kognitivne sposobnosti, mora se pretpostaviti da se one mogu funkcionalno opisati u terminima procesiranja informacija. Stoga, ako se u potpunosti odbaci funkcionalistička slika uma, onda ne bi imalo smisla govoriti o mentalnim sposobnostima koje se mogu analizirati putem Marrova teorijskog okvira. Štoviše, sam temelj kognitivnih znanosti bi se urušio jer počiva na funkcionalističkoj slici uma.[41] Budući da u ovom radu nema mjesta za potpunu raspravu uvjerljivosti funkcionalističke slike uma i temelja kognitivnih znanosti, onda se trenutna rasprava treba shvatiti uvjetno. Onoliko koliko ima smisla pretpostaviti uvjerljivost funkcionalističke slike uma i ulogu Marrove analize u objašnjenju kognitivnih procesa, toliko nam prihvaćanje PSE-a omogućava da na smislen način primijenimo Marrovu trodijelnu analizu na objašnjenje uma u cjelini (tj. uma shvaćenog kao da, osim modularnih, uključuje i nemodularne procese).

Teorijskom okviru koji se temelji na PSE-u bi se mogli uputiti metodološki prigovori. Budući da PSE pripada bayesijanskoj skupini modela, podložan je sličnim kritikama koje se upućuju toj široj skupini pristupa u kognitivnim znanostima. U tom pogledu, bayesijanskim pristupima se upućuju prigovori da su previše fleksibilni, u smislu da *post hoc* namještanjem parametara (na primjer, prilagođavanjem priora) mogu objasniti bilo koji skup podataka, da u Marrovoj terminologiji zahvaćaju samo računalnu razinu objašnjenja, bez uvida u mehanizme kognicije, te da ne uzimaju dovoljno u obzir biološka, evolucijska i algoritmička ograničenja pri formulaciji modela funkcioniranja kognitivnih sustava.[42]

Iako navedene kritike dobro pokazuju slabosti prvog vala korištenja bayesijanskih modela u kognitivnoj (neuro)znanosti, upravo razvoj istraživanja u sklopu paradigme PSE pokazuje kako se oštrica takvih kritika može otupiti. Kao što se pokazuje u ovom radu, paradigma PSE ne zahvaća samo računalnu razinu u Marrovoj hijerarhiji, već putem raznih algoritama, poput prediktivnog kodiranja, omogućava formulaciju teorija o moždanim procesima koji implementiraju različite kognitivne sposobnosti i time minimiziraju slobodnu energiju. Formulacija takvih »procesnih« modela unutar PSE paradigme,

[40] Za raspravu, usp. Jurjako, Malatesti, *Filozofija uma*…, poglavlje 5.
[41] Usp. Bermúdez, *Philosophy of psychology*…, poglavlje 3.
[42] Usp. Jeffrey S. BOWERS, Colin J. DAVIS, Bayesian just-so stories in psychology and neuroscience, *Psychological bulletin*, 138 (2012) 3, 389-414; Matt JONES, Bradley C. LOVE, Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition, *Behavioral and brain sciences*, 34 (2011) 4, 169-188.

omogućava mehanicističko objašnjenje niza kognitivnih fenomena te u više slučajeva pokazuje bolju podudarnost s empirijskim podacima od alternativnih modela.[43]

Osim toga, autori koji rade unutar bayesijanske paradigme, pri formulaciji modela kognitivnih sposobnost sve više počinju uzimati u obzir biološka i druga ograničenja koja karakteriziraju rad ljudskih organa. Poput fizičkih radnji, kognitivni procesi u mozgu, uključujući pamćenje, predstavljanje vrijednosti opcija i donošenje odluka, pretpostavljaju biološke troškove u vidu potrošnje energije. Kao što je već navedeno, upravo algoritmi utemeljeni na prediktivnom procesiranju dobivaju na plauzibilnosti jer se smatra da na energetski učinkovit način mogu implementirati kognitivne procese različite razine kompleksnosti.[44]

Nadalje, činjenica da se u sklopu PSE paradigme pri formulaciji računalnih i algoritamskih teorija o funkcioniranju kognitivnih sposobnosti često uzimaju u obzir biološka ograničenja i alternativni modeli dostupnih empirijskih podataka, umanjuje značajnost prigovora koji se temelji na fleksibilnosti bayesijanskih modela. Drugim riječima, uzimanjem u obzir bioloških ograničenja, formulacija modela će od samog početka bolje odražavati biološke temelje neke psihološke sposobnosti. Također, izvođenjem komparativnih studija u kojima se uspoređuje sposobnost bayesijanskih i ne-bayesijanskih modela, odnosno algoritama da objasne funkcioniranje neke psihološke sposobnosti, umanjuju se potencijalno negativni aspekti fleksibilnosti bayesijanskih teorija. Naime, na taj način se može direktnije testirati njihove komparativne prednosti i nedostatke u objašnjavanju dostupnih empirijskih podataka.[45] Stoga, općeniti prigovori koji su upućeni prijašnjim istraživanjima unutar bayesijanske paradigme ne bi trebali automatski biti prigovor korištenju načela PSE-a za formulaciju modela i daljnji razvoj istraživanja uma i mozga.

Zaključak

Ovaj rad bavio se pitanjem mogu li se Marrove razine objašnjenja koristiti kao opći okvir za objašnjavanje kognitivnih fenomena na različitim razinama opisa. Neki smatraju da to nije moguće jer su Marrove razine objašnjenja

[43] Usp. npr. Lara HENCO i dr., Bayesian modelling captures inter-individual differences in social belief computations in the Putamen and Insula, *Cortex*, 131 (2020) 221-236; David M. COLE i dr., Atypical processing of uncertainty in individuals at risk for psychosis, *NeuroImage: clinical*, 26 (2020) 102239.

[44] Za dodatnu raspravu kako razmatranja energetske učinkovitosti utječu na formulaciju bayesijanskih neuralnih modela odlučivanja i računanja, vidi Paul W. GLIMCHER, Efficiently irrational. Deciphering the riddle of human choice, *Trends in Cognitive Sciences*, 26, (2022) 8, 669-687.

[45] Usp. Henco i dr., *Bayesian modelling…*; Cole i dr., *Atypical processing…*

osmišljene za analizu modularnih kognitivnih sustava, dok su mnoge mentalne sposobnosti nemodularne. Nasuprot tom mišljenju, u ovom radu se tvrdi da takav argument nije dobar te da se i nemodularni sustavi mogu uspješno analizirati putem Marrovih razina objašnjenja. Da bi se to pokazalo, u radu se koristi utjecajna paradigma iz kognitivnih znanosti koja se temelji na principu slobodne energije. Prema ovoj paradigmi, um se može shvatiti kao sustav koji mozak koristi da bi minimizirao slobodnu energiju i tako se održao u poznatim okolnostima koje su određene fenotipom organizma. Budući da princip slobodne energije daje sveobuhvatnu funkciju uma i mozga u kojem je implementiran, ova paradigma pruža jasne naznake kako se modularne i nemodularne mentalne sposobnosti mogu analizirati iz perspektive Marrovih razina objašnjenja. Zaključak rada treba shvatiti uvjetno. Ukoliko se prihvati funkcionalistička slika uma u skladu s, primjerice, principom minimizacije slobodne energije, utoliko se može očekivati da će Marrove razine objašnjenja pružiti koristan okvir za razumijevanje funkcioniranja svih kognitivnih procesa.

Zahvale

Ovim putem želimo iskazati zahvalnost dvama recenzentima na korisnim komentarima koji su pridonijeli poboljšanju prethodne verzije ovog rada. Istraživanje za ovaj rad je omogućeno financijskom potporom projekta KUBIM (uniri-human-18-265) kojeg financira Sveučilište u Rijeci.

Marko Jurjako*

The role of Marr's Levels of Explanation in Cognitive Sciences

Summary

This paper considers the question of whether the influential distinction between levels of explanation introduced by David Marr can be used as a general framework for contemplating levels of explanation in cognitive sciences. Marr introduced three levels at which we can explain cognitive processes: the computational, algorithmic, and implementational levels. Some argue that Marr's levels of explanation can only be applied to modular cognitive systems. However, since many psychological processes are non-modular, it seems that Marr's levels of explanation cannot explain such psychological processes. To show that the latter claim is not convincing, the paper draws upon an influential paradigm from cognitive sciences that is based on the principle of free energy. Based on this paradigm, the paper argues that even non-modular psychological processes can be computationally analyzed and algorithmically implemented. The conclusion of the paper is that, at least under the assumption that the function of the mind is to minimize free energy, Marr's levels of explanation can be successfully used as a general framework for understanding psychological processes at different levels of description.

Key words: active inference, cognitive sciences, Marr's levels of explanation, predictive coding, the principle of free energy minimization.

(*na engl. prev.* Marko Jurjako)

* Marko Jurjako, PhD, Assoc. Prof., University of Rijeka, Faculty of Humanities and Social Sciences, Department of Philosophy and Division of Cognitive Sciences; Address: Sveučilišna avenija 4, 51000 Rijeka, Croatia; E-mail: mjurjako@ffri.uniri.hr.