Teaduspreemia täppisteaduste alal tööde tsükli „Aru ja tehisaru süsteemide sarnasuste ja erinevuste kaardistamine“ eest

Jaan Aru

Foto: Birgit Püve

https://doi.org/10.3176/evp.2024.03

SEKELDUSED JA SEIKLUSED TEADVUSETEADUSE JA TEHISARU TIHNIKUTES

Teaduse keerdkäigud

Aastal 2017 oleks mul olnud väga raske ennustada, et ma saan riigi teaduspreemia. Veelgi huvitavam on see, et mul oleks olnud ka võimatu ennustada, mille eest selle saan. Minu viimaste aastate teadustöö on olnud oluliselt mõjutatud kahest faktorist, mida aastal 2017 oli raske ette näha. Üks neist oli fantastiline teadusavastus, mis muutis mu arusaama teadvusest. Teine oli järsk ja üsna ootamatu areng tehisaru valdkonnas. Kirjeldangi neid kahte nüüd järgemööda.

Kui läksin Berliini, et teha järeldoktorantuur Matthew Larkumi laboris, ei plaaninud ma tegelikult teadvust uurida. Matthew Larkum on legend, kes mõtleb ja püüab välja selgitada, kuidas ajukoor töötab, nii et kavas oligi paremini selgeks teha ajukoore töö algoritmid.

Teadus on põnev just seetõttu, et kunagi ei tea, mis käänaku taga on. Me üritame aru saada nähtustest, mida mitte keegi maailmas veel ei mõista. Seetõttu juhtub üsna palju sündmusi, mis võivad plaanitud teelt kõrvale viia. On teadlasi, kes ei lase end eksitada ja on kinni oma eesmärgis, näiteks selles, mis on granti sisse kirjutatud. Mina nii sihikindel teadlane pole. Usun seda, et uued avastused peavad meie etteplaanitud teed muutma. Peame olema paindlikud.

Mäletan elavalt, kuidas oma Berliini labori töösse sisseelamise ajal üks Matthew Larkumi järeldoktorant mulle lõunasöögi ajal labori inimesi tutvustas ja ütles: „Siin on kõik nii toredad, välja arvatud see üks paranoiline tüüp. Temaga pole mõtet rääkida."

Olen vist alati olnud halb sedasorti nõu kuulamises. Veelgi enam, üks peamisi nõuandeid, mille hea sõber enne doktorantuuri mulle andis, kõlas just vastupidi: „Otsi neid teadlasi, keda teised peavad kummaliseks, ja räägi nendega." Paranoiline – täpselt minu jaoks.

Selgus, et Mototaka polnud sugugi paranoiline. Tal olid lihtsalt veidi teistsugused tööharjumused. Ta oli sageli öösiti laboris ja tegi katseid. Mõnda aega oli tal laboris isegi voodi. Teised järeldoktorandid ei sallinud teda, sest ta jaksas neist rohkem tööd teha ja teda ei huvitanud sotsialiseerumine. Täpselt minu tüüp!

Juba meie esimesel kohtumisel näitas Mototaka mulle hämmastavaid andmeid, mis panid aluse kogu sellele aastapreemiaga auhinnatud uurimistööde tsüklile. Aga et kõigest aru saada, tasub anda natuke taustainfot.

Teadvuse suur probleem

Teaduse juurde oli mind aastal 2004 toonud mõistatus, kuidas aju materiaalsest masinavärgist saab tekkida teadvus. Kuidas ajus tekib tunne olla keegi? Või väga lihtsa näitega: miks näpistamine viib ebameeldiva tundeni? Võime jälitada signaale näpistamise kohast ajju, jälgida valuga seotud protsesse ajus, aga küsimus teadvuse kohta on küsimus sellest, kust tekib see tunne – valu.

Ajakirja The Times Literary Supplement 1992. aasta numbris kurtis filosoof Jerry Fodor kuulsalt: „Kellelgi pole vähimatki aimu, kuidas miski materiaalne võiks olla teadvusel. Keegi isegi ei tea, mis tunne oleks, kui tal oleks vähimatki ettekujutust selle kohta, kuidas mis tahes materiaalne võiks olla teadvusel."

Sir Andrew Huxley, Nobeli preemia laureaat, ütles Cambridge'i neuroteaduste initsiatiivi avasõnades:

„Võib-olla kõige keerulisem ja samal ajal ka kõige huvitavam probleem neuroteaduses on teadvuse olemus ja selle seos ajus toimuvate füüsiliste sündmustega. Veel mõned aastad tagasi vältisid neuroteadlased seda teemat, välja arvatud Sir John Eccles. Teda oli kasvatatud katoliiklasena ja ta uskus hinge olemasolusse igas inimeses. Seevastu on sellest nüüdseks saanud moekas teema, millele on pühendunud mitmed seltsid ja ajakirjad. Mulle tundub, et see on praegu neuroteaduse suurim probleem. Kes iganes selle lahendab, on saanud Newtoni ja Darwini omaga võrreldava koha teaduse ajaloos."

Milline suurepärane motivatsioon igale noorele teadlasele! Niisiis olin selle probleemiga silmitsi seisnud, sellega maadelnud, oma doktorantuuris selle uurimisse panustanud (nt Aru jt, 2012). Tegelikult olin aastaks 2018 teadvuseuuringutest väsinud ja läksin Berliini, et mõista aju saladusi.

Püramidaalrakkude saladused

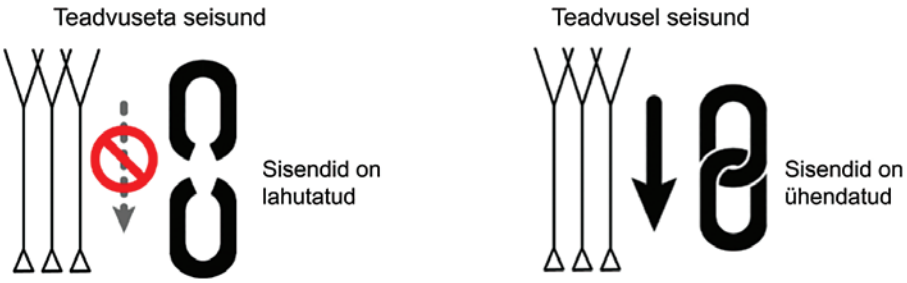
Inimaju töö mõistmise keskmes on tihti just ajukoor (ehk korteks). Selles on miljardeid neuroneid ja mitmeid sadu erinevat tüüpi neuroneid. Meid huvitavad praegu ajukoores sügaval asuvad püramiidja kujuga neuronid. (Täpsemalt asuvad nad viiendas korteksi kihis ja neid nimetatakse viienda kihi püramidaalneuroniteks.) Nendel ajukoore süvakihi püramiidrakkudel on keskne roll ajusiseses andmeedastuses ja info koordineerimises.

Just neid rakke ja nende tööd läksin uurima Berliini Matthew Larkumi juurde. Tema oli Nobeli auhinnaga pärjatud Bert Sakmanni laboris juba aastal 1999 avastanud, et nendel rakkudel on kaks erinevat osa: üks kogub andmeid välismaailmast ja teine sisemaailmast (Larkum jt, 1999). See võimaldab ühe raku tasemel teha midagi näiliselt väga abstraktset: välist ja sisest võrrelda ja kokku viia. Tegu on hämmastava seaduspäraga, millel Larkumi teooria kohaselt on oluline roll mõtlemise masinavärgis (Larkum, 2013).

Varustatuna selle taustainfoga võime nüüd tagasi minna hetke, mil olin jõudnud Matthew Larkumi laborisse ja asunud vestlusesse tema järeldoktorandi Mototaka Suzukiga. Erinevalt sellest, mida see teine järeldoktorant mulle lõunalauas oli öelnud, ei üritanud Mototaka midagi varjata või mind vältida – hea meelega näitas ta oma tulemusi.

Mototaka Suzuki ja Matthew Larkumi katsetes (Suzuki, Larkum, 2020) selgus, et üldnarkoosi all juhtub midagi neuronisiseste protsessidega. Närvirakud, kus teadvusseisundis seesmine ja välimine sisend töötasid üksmeeles, töötasid teadvuseta seisundis teistmoodi. Nimelt ei viinud seesmise sisendi aktiveerimine närviraku aktiivsuseni; seesmine ja välimine sisend olid teineteisest lahutatud (joonis 1). See oli täiesti ootamatu, sest seni polnud keegi arvanud, et teadvusega seotud mehhanismid võiksid olla rakusisesed. Kuidas see töötab? Miks see oli tähtis? Kas võib olla, et teadvuse suure probleemi mõistmiseks on tarvis aru saada rakusisestest mehhanismidest?

Mäletan, kuidas ma ahhetasin, kui seda tulemust nägin. Kui väljusin Mototaka laborist, ei saanud ma seda ideed enam peast välja: äkki nad olid tõesti avastanud võtme teadvuse mõistatuse lahtimuukimiseks? Rakusisese võtme.



Joonis 1. Meie teooria jaoks on tähtis Mototaka Suzuki avastus, et teadvuseta seisundis on püramidaalneuronite kaks osa (ülemine ja alumine) omavahel lahutatud. Teadvusele omase ajuaktiivsuse (ja ka teadvuselamuse) tagab just see, et sisendid on ühendatud ja võimaldavad keerukamat ajuaktiivsust.

Teooria

Mototaka ja Matthew olid küll teinud avastuse, aga teaduses on tarvis avastus ka õigesti konteksti panna ehk öelda teistele, mida see tähendab. Nende artikkel (Suzuki, Larkum, 2020) ilmus küll ajuteaduse tippajakirjas Cell, aga oli võrdlemisi tehniline ja seega ma teadsin, et ega teadvuseuurijad seda eriti ei loe. Või kui isegi loevad, siis nad ei suuda aru saada, mida see tähendab. Seega mina nägin enda rolli nende tulemuste laiema tähenduse väljaselgitamises. Erinevalt Mototakast ja Matthew'st olid mul vastavad teadmised juba olemas, sest olin ju varem pikalt ja põhjalikult teadvust uurinud.

Niisiis asusin mõtlema, kuidas selgitada nende tulemuste tähtsust teadlastele, kes küll mõtlevad teadvusest, aga ei saa eriti midagi aru ajust. Minu mõttetöö põhjal sündis dendriidilise integratsiooni teooria (ehk DIT), mis väidab, et teadvuse aluseks on see, kui aju keskel asuvad kõrgema taseme taalamuse tuumad moduleerivad sügavate kihtide püramidaalneuroneid nii, et suhtlus raku kahe osa vahel toimub. Sel juhul on püramidaalneuronid aktiivsemad ja nad lükkavad käima aktiivsusahelad ajukoore ja koorealuste struktuuride vahel, soodustades aju dünaamilist sidusust. Aju aktiivsus saab olla keerukam ja mitmekesisem just tänu sellele lokaalsele mehhanismile püramidaalrakkude sees. Teadvus võib olla seotud laia skaala ajuaktiivsuse mustritega, aga nende tekkimisel ja arengul on kesksel kohal just need lokaalsed rakusisesed mehhanismid.

Meie artikli (Aru jt, 2020) peamine argument oli, et teadvuseteaduses on küll palju teooriaid, aga me peame arvesse võtma viimase aastakümne uusi teadmisi ajuteadusest. Erinevalt teistest teadvuseteooriatest keskendub DIT neurobioloogiale. Just nii, nagu Francis Crick teadvuseteaduse algusaastatel ütles – teadvus tekib ajus ja seega peame mõistma neid neuronaalseid mehhanisme, mis selle aluseks on. DIT on noor teooria, mis on siiski saanud teadvuseteaduses juba üsna palju kõlapinda ja on esimene osa teaduspreemiani viinud artiklite sarjast (Aru jt, 2020; vt ka Bachmann jt, 2020).

Tehisaru rikub teadvuseteadlaste peo ära

Niisiis näis kõik tore: olime just välja pakkunud uue viisi, kuidas teadvust mõista ja hetkeks (aga vaid hetkeks) tundus, et oleme teadvuse mõistatuse lahendamisele lähemal. Aga siis tuli tehisaru.

Ilmselt on paljud lugejad kasutanud tehisaru süsteeme nagu ChatGPT. Tegu on hiigelvõimsa tehisnärvivõrguga, millesse on sõna sõna haaval sisse treenitud praktiliselt kogu internetis olemas olev mõistlik ja ka ebamõistlik tekst. Iga uue sisseantud sõna korral ennustab algoritm, mis peaks olema järgmine sõna. On küllaltki üllatav, et sellise suhteliselt labase õpetusega süsteemid suudavad anda (vahel) sisukaid vastuseid. Nad käituvad, nagu nad oleksid arukad. Näiteks võin tehisarusse ChatGPT sisestada oma artikli (Aru jt, 2020) ja paluda tehis-

arul see kolme lausega kokku võtta: „Artikkel käsitleb seda, kuidas hiljutised neurobioloogilised läbimurded on valgustanud teadvuse aluseks olevaid rakulisi mehhanisme, eriti rõhutades kortikaalsete püramiidrakkude rolli. Tutvustatakse dendriitilise integratsiooni teooriat, mis väidab, et alt-üles ja ülevalt-alla andmevoogude paindlik integratsioon rakutasandil on teadvusliku töötluse jaoks hädavajalik.“ Pole ju paha. Tehisaru süsteem teeb selle kokkuvõtte kümne sekundiga.

Probleem teadvuseteadlase jaoks on see, et maailmas on suur hulk inimesi ja ka arvutiteadlaseid, kes arvavad, et ChatGPT-sarnased süsteemid võivad olla teadvusel. Kui mitte täna, siis juba peagi. Kui ChatGPT-lt praegu küsida, kas ta on teadvusel, siis ta tavaliselt vastab umbes nii, et tema on tehisaru süsteem ja ei saa selle kohta midagi öelda või et ta pole teadvusel. Aga see polnud alati nii. Aasta 2022 suvel vastasid sarnased tehisaru süsteemid, et nad on teadvusel. Ehkki see, et üks süsteem ütleb, et ta on teadvusel, ei peaks olema kriteerium, mille järgi hinnata, kas süsteem on teadvusel või ei, on mitmed tehisaru uurijad justkui ära tinistatud. Nad usuvad, et need süsteemid võivad tõesti olla teadvusel. Aga mulle tundus see usk pime. Miks peaks ChatGPT (või sellega sarnased süsteemid) olema teadvusel? See, et nad vastavad vahel viisakalt nagu inimene, ei tähenda, et nad tunnevad nagu inimene. Minu jaoks oli sellel järeldusel liiga vähe teaduslikku tõendusmaterjali.

Või ma eksisin? Üks põhjus, miks on raske vastu vaielda sellele, et ka tehisaru võib olla teadvusel, on asjaolu, et nii arvutuslikus neuroteaduses kui ka tehisaruuuringutes eeldatakse, et nii närvirakud kui ka üldine arhitektuur on tehisaru süsteemis ja inimajus võrdlemisi sarnased. Näiteks keskmise õpiku kohaselt töötab närvirakk umbes nii: informatsioon siseneb dendriitide kaudu, kust ta liigub läbi rakukeha ja lõpuks saadetakse aksoni kaudu aktsioonipotentsiaalid, mis seda informatsiooni edasi kannavad. Neuron tuliskleb.

Niimoodi lihtsustatult kirjeldamisel on oma võlu. Minu jaoks on nüüd ilmnenud, et lihtsustamisel on ka pahupool. Just selle lihtsa kirjelduse tõttu arvavad tehisaru uurijad, et tehisnärvivõrkudes olevad neuronid ja bioloogilised neuronid on sarnased. Ka tehisneuronid koguvad „presünaptilisi“ sisendeid, „kaaluvad“ neid ja siis väljastavad signaali. Kui neuronite töö on täpselt sama ja kui ka töötlus on sama, siis miks peaks üks süsteem olema teadvusel, samas kui teine mitte? Mis on ajus teistmoodi?

Kuidas aju arhitektuur erineb tehisaru omast

Kas tehisaru on inimarule sarnane? Esimene kaitse selle argumendi vastu on, et sarnasus on pinnapealne ja tänapäeva tehisaru süsteemid ei tööta ajuga sarnaselt. Artiklid, mis seda väidavad, moodustavad viimase osa mu teaduspreemiale nomineeritud artiklitsüklist (Aru jt, 2023; Suzuki jt, 2023).

Tuleme tagasi selle järeldoktorandi juurde, kelle eest mind esialgu hoiatati. Nagu oleme näinud, muutsime koos seda, kuidas mõeldakse teaduse neurobioloogilistest korrelaatidest. Tegelikult möödus suur osa minu Berliinis oldud ajast Mototaka laboris, kus tegime katseid ja arutlesime aju üle. Üks mõte, mida juba tollal, st aastal 2018, jagasime, on see, et ehkki tehisaru tehnoloogia taga on sügavad närvivõrgud, ei vasta nende arhitektuur sellele, kuidas asjad ajus käivad.

Sügavad närvivõrgud on „sügavad", kuna neis on palju kihte. Kui anda neile sisendit, näiteks see tekst siin, siis see läbib kiht-kihilt terve tehisnärvisüsteemi. Osas sellistest süsteemidest on sadu kihte, kust sisend järjest läbi liigub. Aga see ei tundu olevat viis, kuidas aju töötab.

Ajus on küll olemas töötlushierarhia, mis on hoopis vähem hierarhiline kui tehisnärvisüsteemis. See tähendab, et madalalt tasemelt on võimalik otse ühendust võtta kõrgema taseme töötluspiirkondadega.

Teine väga oluline erinevus hierarhilisest töötlusest on see, et aju keskel olev struktuur – taalamus – on seotud kõigi ajukoore osadega. Niisiis, ehkki töötlus ajukoores võib küll tunduda hierarhiline, on igal ajukoore osal võimalik iga teisega suhelda kiiresti taalamuse kaudu.

Kui me umbes aastal 2018 selle üle arutlesime, siis tundus idee nii lihtne ja loomulik. Küllap keegi sellest varsti kirjutab! Aga kui aasta 2022 keskel Mototakaga arutasime, millest võiks kirjutada, siis oli idee teiste poolt ikka avaldamata. See oli üllatav, aga väljendab midagi olulist, mis teaduses ikka juhtub: kui liigutakse kaasa peavoolu ideedega, siis võidakse mõned väga lihtsad põhiasjad ära unustada. Peavooluks on juba mõned aastad olnud idee, et aju teeb midagi sügavõppe (ehk süvaõppe) sarnast, aga nagu ülal esile toodud, ei vasta see tegelikkusele.

Siin nägimegi võimalust: me ei hakka spekuleerima sel teemal, kuidas aju töötab, vaid rõhutame seda, et on anatoomilised faktid, mis näitavad, et aju ei tööta nagu sügavõpe. Natuke rumalal kombel oli lõpuks artikli pealkirjas siiski kirjas ka „Shallow brain hypothesis" (pindmise aju hüpotees), mis justkui viitaks, et tegu on hüpoteesiga. See oli väike möödalaskmine meie poolt, aga samas kuidagi tahad oma uut ideed ju ka nimetada.

Igal juhul oli meie peamine argument, et ei pea vaidlema selle üle, kas aju teeb midagi sügavõppe sarnast või ei. Me vaatame fakte ja faktid ütlevad, et aju teeb muu hulgas ka midagi teistsugust, kus õppimine ja käitumine pole mitte sügav, vaid madala töötlushierarhiaga. Iga korteksipiirkond on vahetult seotud taalamusega ja taalamus saab mõjutada iga korteksipiirkonda.

Ehk osalt seetõttu, et võtsime sihtmärgiks ühe moodsa ajuteaduse eelduse ja ründasime seda faktidega, õnnestus meil artikkel saada ajuteaduse kõige täht-

samasse ülevaateajakirja Nature Reviews Neuroscience (Suzuki jt, 2023). Minu teada pole tänaseks päevaks ükski eestlane varem seal ajakirjas artiklit avaldanud (aga loodan, et eestlaste kirjutatud artikleid ilmub seal veel ja veel!).

Niisiis erineb aju arhitektuur tehisaru omast (Suzuki jt, 2023). Sellest vastuargumendist ei piisa, et veenda tehisaru uurijat, kes arvab, et ka tehisaru võib olla teadvusel. Vastava tehisaru uurija küsimus on, mis juhtub, kui loome tehisaru, mis töötab dendriitilise integratsiooni teooria põhimõtetel, kus on püramidaalsed rakud ja aju keskel asuv keskjaam taalamus. Kui teeme tehisaru nii, et algoritm on sama mis meie kirjeldatud teoorias, siis kas see süsteem oleks teadvusel?

Bioloogiline *vs*. tehislik

Sel hetkel, kui tehisaru süsteemid on loodud nii, et nende sees toimuvad arvutusprotsessid on bioloogilise aju omadega sarnased, on teadlasel kaks valikut. Kas nõustuda, et need süsteemid tõepoolest on teadvusel, sest nendes toimuvad arvutusprotsessid ju vastavad sellekohasele teadvuseteaduse teooriale, või siis võtta samm tagasi ja üritada aru saada, kus midagi valesti läinud on.

Minus tekitab kõhklust see, et ehkki meie teooria järgi loodud tehisaru simuleeriks selle teooriaga seotud arvutusprotsesse, oleks see ju kõigest arvutiprogramm, tarkvara. Aga kas teadvus on tarkvara? Või on siin üks sügavam probleem?

See ei ole uus küsimus. Juba 1990. aastal rõhutas filosoof John Searle, et igal arvutusmudelil on talle omased piirid. Näiteks tornaadot simuleeriv arvutusmudel ei hirmuta meid, tsunaamilainete mudel ei tee meid märjaks ja pitsa seedimise mudel ei täida meie kõhtu. See argument on iseenesest loogiline, aga üllataval kombel peame täna, enam kui kolm aastakümmet hiljem, ikka sama teema üle vaidlema. Põhjus peitub ilmselt ülivõimsates tehisaru süsteemides, mis suudavad (vahel) anda mõistlikke vastuseid ja jätta (vahel) nutika mulje. Kui Searle'i filosoofiline argument ei suutnud arvutiteadlasi veenda 1990. aastal, siis on see tänapäeval oma veenvuse veelgi enam kaotanud. Ehk peaks argument teadvuse ja tarkvara erinevuse kohta olema bioloogiline?

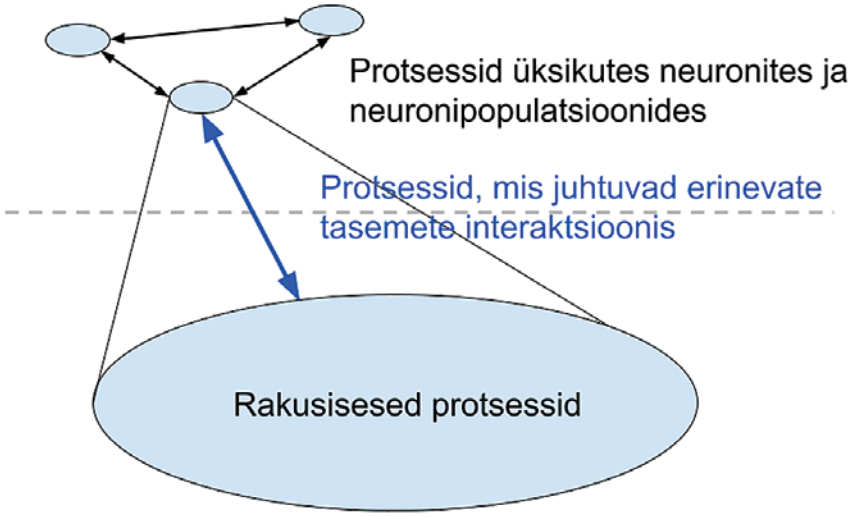
Nende teemade üle aastal 2022 juureldes hakkas mulle üha enam terendama, et kõikide nende protsesside mudeldamine ja simuleerimine jätab ikka välja peamise: bioloogia ei ole kõigest tarkvara. Kui võtaksime lugeja ajust välja mõne neuroni, saaksime selle katseklaasi panna ja seal elus hoida. Bioloogiline neuron oleks võimeline seal kasvatama uusi jätkeid, looma ühendusi teiste neuronitega. Need ühendused võivad katseklaasis muutuda tugevamaks või nõrgemaks. Neuron võib katseklaasis tuliskleda. Bioloogiline närvirakk teeb kõike seda ja muudki veel, samas kui tehisneuron on vaid mõttetu tükk koodi.

See tundub olevat päris oluline erinevus. Me võime neuroni tööd küll kirjel-

dada valemitega, mis vastavad ka tehisnärvirakkude aktiivsusele, kuid valemid väljendavad vaid tibatillukest osa sellest, mis raku sees tegelikult toimub. Need valemid ei ela. Niisiis on teadvuse aluseks oleva aju ehitusplokid – närvirakud – tehisneuronitest teatud mõttes väga erinevad. Võtmeküsimus on, kas see erinevus on tähtis.

Probleem on nimelt selles, et mitte ükski tänase päeva juhtivatest teadvuseteooriatest ei selgita, miks see, et neuronid on päriselt eksisteerivad keerukad bioloogilised masinavärgid, peaks kuidagi teadvuse jaoks oluline olema. Kuidas neuronite sees toimuv on seotud teadvusega? Pigem keskenduvad teooriad just arvutuslikule tasemele. Näiteks üks tuntumaid teadvuseteooriaid – globaalse tööruumi teooria – ütleb, et teadvus on see, kui sisend jõuab nii-öelda globaalsesse tööruumi. Globaalse tööruumi teooria keskmes on arvutuslik protsess, mille käigus lokaalsed töötlusüksused jagavad oma tööd globaalse tööruumiga. Seda arvutuslikku protsessi on võimalik tehisaru süsteemides rakendada. Tänapäeval on juba tehisarusüsteeme, mis seda teevad.

Need teooriad ignoreerivad täielikult pea kõiki mikrotaseme detaile, mis teevad närvirakust närviraku. Kas need on lihtsalt bioloogilised detailid, mis on tehisaru ja tehisteadvuse loomiseks täiesti ebavajalikud? Või on need „detailid“ neuroni sees toimuva kohta just esmatähtsad – ilma nende mõistmiseta ei saa me



Joonis 2. Enamik tänapäeva teadvuseteooriaid vaatab protsesse, mis toimuvad üksikutes neuronites ja neuronipopulatsioonides (kriipsutatud joone kohal). Meie arvates peaks rohkem süvenema ka rakusisestesse protsessidesse ja sellesse, kuidas rakusisesed protsessid on kaastoimes sellega, mis toimub neuronipopulatsioonide suhtluses.

kunagi teadvusest aru? Mulle tundub ja ma loodan, et oleme uue ajastu alguses, kus teadvust hakatakse üha enam nägema kui vahetult elusolemisega seotud fenomeni. Aga ainult sellest ei piisa, sest teadlastena tahaksime aru saada, miks elusolemine peaks olema seotud teadvusega. Täpselt millised elussüsteemide omadused on need, mis toovad kaasa teadvuse? Meil pole tänasel päeval veel rahuldavat teooriat, mis ütleks, kuidas on elu ja teadvus seotud ja kas teadvus tekib ainult elusloodusе teatud keerukuse astmel. Isegi kui arvutusprotsessid ise ei ole teadvuse aluseks, soovime teadlastena arvutuslikku teooriat sellest, mis teadvus on (ja miks eluga seotud protsessid teadvuse jaoks kesksed on).

Siiski on tore tõdeda, et need ideed on viimaks jõudmas ka teadvuseteaduse diskussiooni. Vähemalt mina isiklikult olen pidanud enda jaoks ümber mõtlema seda, kas teadvus on pelgalt arvutuslik protsess, mida saab tarkvaras järele teha, või on ta midagi põnevamat. Veel selle teadustööde tsükli alguses, kui avaldasime dendriitilise integratsiooni teooria (Aru jt, 2020), ei mõelnud ma väga selle üle, kas teadvus on kõigest arvutuslik protsess või ei. See oli vaikimisi eeldus, mille üle ei arutletud. Tehisaru kiire areng on sundinud mind asju ümber mõtlema. Kui teadvus on kõigest arvutuslik protsess, siis saab seda tehisaru süsteemides järele teha ja süsteemid nagu ChatGPT on (varsti) teadvusel. Aga ehk näitavad tehisaru süsteemid hoopis seda, et elusloodusеs toimuvad protsessid on teadvuse mõistmisel tähtsamal kohal, kui me seda eales arvanud oleksime.

Uusi ideid pole üldse kerge kirja panna. Kui tahad mõelda täiesti varasemast raamistikust välja, siis oled üksi mäe tipus, kus on külm ja kõle. Läks umbes aasta, enne kui suutsin hakata asju vaikselt kirja panema. Oluliseks tõukepunktiks sai see, et osalesin Montrealis ühes töötoas, kus esines ka tehisarualase maailma üks kõige mõjukamaid teadlasi Yoshua Bengio ja üks Google'i tehisaru uurijate juhte Blaise Aguera y Arcas. Mina kirjeldasin meie enda ideid ja tulemusi (nt Aru jt, 2020), aga eelnimetatutest viimane ütles oma ettekande ajal, et tehisaru süsteemid on ka teadvusel – ajus ei ole mingit „haldjatolmu".

See tegi mind väga vihaseks. Õnneks saab teadlane oma viha välja valada ka klaviatuuri toksides. Emotsionaalne seisund võimaldas minu mõtetel lõpuks paisu tagant välja pääseda. Oma osa oli muidugi ka selles, et järgmisel hommikul oli meil hommikusöök kahe sarnaselt mõtleva teadlasega, kelleks olid Mac Shine Austraaliast ja mu enda endine juhendaja Matthew Larkum. Tavaliselt on söögilauavestlused teadlastega üsna igavad – nagu ka Nobeli auhinna saaja Jim Watson on öelnud, räägivad teadlased söögilauas täiesti tavalist juttu. Aga sel hommikul ma nii juhtuda ei lubanud. Iga kord, kui teema hakkas mujale hajuma, tõin selle kohe tagasi – „olgu, aga mis me selle tehisaru probleemiga ette võtame?". Nii me arutasime, sõime ja arutasime. Neli tundi järjest. See oli mu elu parim hommikusöök, mille põhjal sündis ka meie kolme artikkel, kus kirjeldame meie jaoks kõige olulisemaid punkte, miks tehisaru ei ole teadvusel (Aru jt, 2023). Selles artiklis on palju uusi ideid (vähemalt minu jaoks).

Kust tulevad uued ideed?

Kuidas inimaru üldse tuleb uute ideede peale? Kas seda saab kuidagi tagant aidata? Ka nende küsimuste uurimine mahtus mu viimase nelja aasta uurimistöö sisse.

See küsimus on mulle isiklikult oluline. Ma teen teadust just nende hetkede pärast, kus tundub, et olen midagi uut välja mõelnud või millestki uuest aru saanud. On fantastiline tunne, kui saab hetkeks arvata, et oled ainus inimene maamunal, kes mingist küsimusest või teemast aru saab. (Tihti selgub pärast, et see arusaam oli vale – näiteks lahendus tegelikult ei töötanud või oli keegi juba midagi taolist arvanud. Aga sellest pole midagi.)

On ebaselge, kuidas see toimub. Kuidas ajus tekivad uued mõtted? Minu hüpotees (Aru, 2022) on, et uute mõtete loomine on nagu Lego-ehitus. Nagu uue Lego-idee ehitamiseks on tarvis palju Lego-tükke, nii on selleks, et midagi uut välja mõelda, tarvis palju teadmisetükke, mida aju sees omavahel kombineerida. Niisiis on ka heaks teadlaseks saamisel tarvis palju lugeda selle kohta, mida teised teinud on. Aga ainult lugemisest ei piisa; lisaks tuleb ka üritada pidevalt uusi kombinatsioone leida – motiveeriv jõud on ikkagi millestki uuest arusaamine.

Kuidas kombineerimine käib? Kas seda, et uued kombinatsioonid ilmneksid, saab kuidagi soodustada? Üks tõsiasi, mis mind aastaid tagasi selle teema juurde tõi, on see, et uued mõtted tulevad tihtipeale seisundis, mis on väga erinev tavalisest töötegemisest. Kõige kaunimalt kirjeldab seda kuulus saksa teadlane Hermann von Helmholtz: „Sageli [---] saabusid [ideed] ootamatult, ilma minupoolse pingutuseta, nagu inspiratsioon [---] Nad ei tekkinud kunagi väsinud ajus ega kirjutuslaua taga. Alati oli vaja, et oleksin esiteks probleemi igast küljest analüüsinud, nii et mul oleksid kõik selle nurgad ja keerukused meeles. [---] Siis [---] peab tulema tund täielikku füüsilist värskust ja vaikset heaolu, enne kui head ideed saabuvad. Tihti olid nad kohal hommikuti, kui ma ärkasin [---] Aga eriti meeldis neile end ilmutada sel ajal, kui ma päikesepaistelise ilmaga metsas üle küngaste jalutasin.“

Kuidas saab nii olla? See on teadlase jaoks huvitav mitmel põhjusel. Esiteks näitavad niisugused kirjeldused, et ülesande lahendamine ei ole mitte ilmtingimata välismaailmaga seostuv protsess, vaid võib toimuda täitsa ilma välise sisendita. Mitte et välist sisendit kunagi ülesannete lahendamiseks ei kasutataks, aga fakt, et lahendused tulevad jalutades või ärgates, näitab, et uue lahenduse leidmiseks kõige tähtsam protsess on ajus toimuv ümberkorraldumine. Võiks isegi öelda, et see sisemine protsess, see ümberkorraldamine, on uute ideede jaoks hädavajalik.

Teiseks näitavad sellised kirjeldused, et lahendused tulevad tihti mitte tööd tehes, vaid tööst eemal olles – jalutades, vannis, pikutades, niisama logeledes. Miks? Mis toimub puhkeseisundis niisugust, mis muidu ajus juhtuda ei saa?

Paarkümmend aastat tagasi tehti hämmastav avastus: une ajal on rottide ajus võimalik mõõta aktiivsusmustreid, mis kordavad ärkvelolnud ajus toimunud aktiivsust. Järgnevatel aastakümnetel on näidatud, et see ei toimu sugugi mitte ainult une ajal, vaid igasuguse rahuliku seisundi puhul. Tegu ei ole sugugi mitte ainult rumala taaskordamisega, vaid pigem on need mustrid, mida aju tagasi mängib, mänguliselt põnevad. Need ei ole pelgalt taaskordused, vaid kombinatsioonid. Vaikses seisundis toimub rohkem kombineerimist ja aju on paremini võimeline teadmisetükke kombineerima (Aru jt, 2023b). See on seisund, kus meil tulevad uued lahendused (Tulver jt, 2023).

Niisiis, kui on vaja uute ideede peale tulla, siis soovitan logeleda! Parim teadustöö ei sünni ennast tohutult tagant utsitades. Parim teadustöö sünnib, kui end tagant utsitada, aga vahepeal ka logeleda. Parafraseerides Tammsaaret: tee tööd, näe vaeva, logele, siis tulevad ka uued ideed.

VIITED

Aru, J. 2022. Loovusest ja logelemisest. Ajujutud, Tallinn.

Aru, J., Bachmann, T., Singer, W., Melloni, L. 2012. Distilling the neural correlates of consciousness. Neuroscience & Biobehavioral Reviews, 36(2), 737–746, https://doi.org/10.1016/j.neubiorev.2011.12.003

Aru, J., Drüke, M., Pikamäe, J., Larkum, M. E. 2023b. Mental navigation and the neural mechanisms of insight. Trends in Neurosciences, 46(2), 100–109, https://doi.org/10.1016/j.tins.2022.11.002

Aru, J., Larkum, M. E., Shine, J. M. 2023a. The feasibility of artificial consciousness through the lens of neuroscience. Trends in Neurosciences, 46(12), 1008–1017, https://doi.org/10.1016/j.tins.2023.09.009

Aru, J., Suzuki, M., Larkum, M. E. 2020. Cellular mechanisms of conscious processing. Trends in Cognitive Sciences, 24(10), 814–825, https://doi.org/10.1016/j.tics.2020.07.006

Bachmann, T., Suzuki, M., Aru, J. 2020. Dendritic integration theory: a thalamocortical theory of state and content of consciousness. Philosophy and the Mind Sciences, 1(II), https://doi.org/10.33735/phimisci.2020.II.52

Larkum, M. 2013. A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. Trends in neurosciences, 36(3), 141–151, https://doi.org/10.1016/j.tins.2012.11.006

Larkum, M. E., Zhu, J. J., Sakmann, B. 1999. A new cellular mechanism for coupling inputs arriving at different cortical layers. Nature, 398(6725), 338–341, https://doi.org/10.1038/18686

Suzuki, M., Larkum, M. E. 2020. General anesthesia decouples cortical pyramidal neurons. Cell, 180(4), 666–676, https://doi.org/10.1016/j.cell.2020.01.024

Suzuki, M., Pennartz, C. M., Aru, J. 2023. How deep is the brain? The shallow brain hypothesis. Nature Reviews Neuroscience, 24(12), 778–791, https://doi.org/10.1038/s41583-023-00756-z

Tulver, K., Kaup, K. K., Laukkonen, R., Aru, J. 2023. Restructuring insight: An integrative review of insight in problem-solving, meditation, psychotherapy, delusions and psychedelics. Consciousness and cognition, 110, 103494, https://doi.org/10.1016/j.concog.2023.103494

Jaan Aru

Sündinud 21. novembril 1984

2004	Tartu Hugo Treffneri gümnaasium (kuldmedal)
2004–2008	psühholoogiaõpingud Berliini Humboldti ülikoolis
2009–2014	doktorantuur Max Plancki aju-uuringute instituudis / Frankfurdi Goethe ülikoolis (*summa cum laude*)

Alates aastast 2014 olen töötanud Tartu ülikoolis. Aastatel 2018–2020 olin Marie Skłodowska Curie järeldoktorandi stipendiumiga Berliinis. Pärast seda tulin tagasi Eestisse ja olen alates 2021. aastast olnud Tartu ülikooli arvutiteaduse instituudi kaasprofessor. Olen kirjutanud kaks populaarteaduslikku raamatut („Ajust ja arust“, „Loovusest ja logelemisest“). Alates 2019. aastast olen Eesti noorte teaduste akadeemia liige ja 2021. aastast kuulun Tartu ülikooli teaduskooli nõukogusse.