Ja ihminen loi älyn podcast: jakso7, litteraatti

Tekoäly, tietoisuus, ihmisyys – Lauri Järvilehto

Arno Solin: Eli jos minä vielä kristallisoin tämän kysymyksen, niin voiko tekoäly olla tietoinen?

Lauri Järvilehto: Viisi vuotta sitten olisin sanonut, että todennäköisesti ei. Nyt sanoisin, että todennäköisesti voi. Jos aletaan kohdella autoa, puhelinta, tietokonetta tai tekoälyalgoritmia ihmisenä, niin silloin se on eettisesti aika vaarallinenkin positio. Sehän on ihan järkyttävä vyyhti.

Arno: Eli siis sanot, että termostaatti on tietoinen mutta et ole varma olemmeko me tietoisia?

Lauri: Niin, tähän ollaan tultu. Mitäs pyysitte filosofin tänne.

Eric Malmi: Tervetuloa Aalto-yliopiston Ja ihminen loi älyn -podcastin pariin. Minun nimeni on Eric Malmi, ja toimin Aalto- yliopistossa ja Suomen ELLIS-instituutissa vierailevana professorina. Kanssani tätä podcastia juontaa Arno Solin.

Arno: Toimin Aalto-yliopistossa professorina ja niin ikään Suomen ELLIS-instituutissa. Tänään meillä on ilo saada vieraaksi ihka oikea filosofi. Myös Aalto-yliopiston työelämäprofessori, tietokirjailija ja muusikko, Lauri Järvilehto. Lauri on tutkinut ihmisen ajattelua, intuitiota ja päätöksentekoa jo yli 15 vuoden ajan. Hän on perustunut Filosofian Akatemian, ollut perustamassa oppimispeliyhtiö Lightneeria ja kirjoittanut 9, kohta 10, suomenkielistä tietokirjaa. Tänään me puhumme Laurin kanssa ihmisen ja konemielen eroista, yhteneväisyyksistä, tietoisuudesta ja siitä, miten tekoäly haastaa meidän käsitystämme ihmisyydestä. Tervetuloa, Lauri.

Lauri: Kiitoksia.

Eric: Lauri, sinulla on häkellyttävän monta erilaista ammatillista roolia. Minusta tuntuu, että useimmille meistä yksikin näistä olisi jo kokopäivätyö. Mutta mielestäni kaikista hämmästyttävintä on se, että kaikkien näiden lisäksi olet viiden lapsen isä. Miten tämä on mahdollista? Miten onnistut tasapainottelemaan näiden eri roolien välillä?

Lauri: No, se on aktiivinen priorisointi. Esimerkiksi teen puolen vuoden välein aina, ja sitten tarvittaessa kuukausi tai jopa viikko kerrallaan, silloin kun pyöritin startupia, niin silloin tein joka sunnuntai, viikon prioriteetit. Tällä "must-win battles" -ajattelutavalla. Että mitkä ovat ne kaksi, kolme juttua, jotka on pakko saada maaliin. Mutta ehkä se isoin tekniikka, joka minulla on ollut, on tämä David Allenin Getting things done. Eli se, että kaikki strukturoidaan viime kädessä niin, että aina kun avataan mikä tahansa lista, niin se on heti tehtävissä.

Arno: Mutta mahtavaa, että vierailu tänne podcastiin on sopinut prioriteettilistaasi. Varmaan tämän viikon yksi tärkeimpiä asioita, toivottavasti.

Lauri: Täytyy sanoa, että olin kyllä tosi fiiliksissä, kun pyysitte mukaan tähän. Minusta tämä on tosi hienoa mitä te teette. Tämä podcast. Mutta siis myös tuo työ, jota te teette molemmat on minusta tosi upeaa. Olen tosi kiinnostunut. Tietenkin itse kun en ole tekninen tekoälytutkija, niin joissakin asioissa olen vähän ihmeissäni kaikesta mitä tapahtuu. Mutta sitten taas minulla on oma akateeminen taustani mielenfilosofiassa. Ne linssit päällä, kun olen perehtynyt tähän kenttään ja kirjoitin tämän konemielikirjan juuri tästä, tavallaan vertailin sitä omaa ihmismieltä koskevaa tutkimusta siihen, mitä on saanut kaivettua esiin näistä tekoälymalleista, niin tämä on minusta tosi kiehtova aihe kaiken kaikkiaan. Oma tutkimusfokukseni on aika paljon tässä landscapessa ollut nyt viime vuodet, niin olen yrittänyt saada ymmärrettyä, että onkos niitä ajattelevia koneita, tai onko niitä edes mahdollista tehdä, ja jos on, niin miten niitä pitäisi tehdä, ja semmoista.

Arno: Me Ericin kanssa olemme tosi teknisiä. Me lähdemme varmaan sieltä ruohonjuuritasolta ja katsomme juuri niitä teknisiä detaljeja. Tavallaan sitä taustaa vasten näemme nämä asiat. Välillä on semmoinen fiilis, että ehkä me emme ihan näe metsää puilta. Nyt meillä on kerrankin sellainen vieras, jolla ehkä on täysin erilainen näkökulma näihin juttuihin. Odotan tätä keskustelua tosi innokkaasti.

Lauri: Niin, minä olen ehkä semmoinen, että anteeksi mitä? Puita?

Eric: Myös ehkä ne kysymykset, joita meiltä monesti kysytään, vaikka lähipiiristä, saattavat olla enemmän juuri niitä filosofisia kysymyksiä. Sen takia me halusimme saada asiantuntijan paikalle, että saadaan niihin asiantuntevia vastauksia.

Arno: Me haluamme myös näin keskustelun aluksi kiittää kaikkia katsojia, jotka ovat käyneet kommentoimassa näitä meidän podcastimme edellisiä jaksoja, esittäneet kysymyksiä kommenttikentässä. Me luemme kaikki ne kommentit, ja isoon osaan olemme vastanneetkin. Kaikkiin emme ole ehtineet vastata, mutta me olemme tietoisia kaikista niistä. Yksi teema, joka on noussut tosi useassa kommentissa esiin on tekoälyn tietoisuus. Jotta me voimme puhua tästä, niin puhutaan ensin hieman siitä, että miten ihmisaivot toimivat, ja mitä se tietoisuus ylipäätään edes on. Näin ehkä insinöörimäisesti, meillä on insinööritausta, niin voisi ajatella, että ihmisaivot toimivat vähän kuin tietokone. Niihin tulee aistien kautta jotain inputtia, syötteitä. Ne prosessoivat ja yhdistelevät näitä syötteitä. Sitten ulostulona, ne lähettävät joitain signaaleja, syötteitä, ulos. Nämä signaalit saavat sitten meidän kehomme toimimaan, suun käymään ja niin edespäin. Ovatko ihmisaivot vain yksi iso tietokone? Ja voiko ajattelua jotenkin verrata tämmöiseen laskentaan ja asioiden prosessointiin näin yksinkertaisella tasolla?

Lauri: Hirveän suosittu ajatus on tuo, että aivot toimivat kuin tietokone, ja oikeastaan mitä kuvasit, koko input-output- prosessointiajatus. Se, miksi olen skeptinen sitä kohtaan on ensinnäkin se, että nyt jos ajatellaan, että aivot ovat vain laskukone, niin silloinhan koko tämä tietoisuuskysymyshän tavallaan muuttuu tyhjänpäiväiseksi. Että miksi meillä ylipäätään on tämmöinen tietoisuus? Mihin sitä tarvitaan tähän koko juttuun? Esität minulle inputin, ja nyt täältä tulee outputti. Tietoisuuden osalta on

haastavaa, kun kukaanhan ei oikeasti tieteellisesti pysty tutkimaan tietoisuutta. Voin kertoa, että olen nyt tietoinen tästä huoneesta ja näen tässä kaikenlaista mitä täällä on ja näin. Mutta et koskaan tule pääsemään minun pääni sisälle. John Locke aikanaan esitti tämän tällaisen vertauksen, että mehän emme oikeastaan tiedä, onko tämä mitä me koemme samanlainen kokemuksena. Jos se onkin niin, että tuossa on tuo sininen pahvi, ja kun sinä katsot sitä, sinä katsot sitä, ja minä katson sitä, me kaikki sanomme, että sininen pahvi, eikö niin? Koska me indikoimme sen saman asia samalla sanalla, niin me olemme tyytyväisiä. Mutta mitä jos kokemuksellisessa mielessä minä näen tuossa sinisen, joka sinulle olisikin vihreä ja sinulle keltainen? Tätä ei pystytä tutkimaan.

Me voimme tutkia siis korrelaatteja. Esimerkiksi me pystymme katsomaan, mitä vaikka tapahtuu minun V1-aivokuorellani tuolla. Sehän on ihan siis kreisiä. Sieltähän pystytään siis laskemaan auki, siis tällaisella fMRI-koneella pystytään kuvaamaan tuolla tapahtuvia muutoksia. Esimerkiksi Jack Gallantilla Kaliforniassa ja Yukiyasu Kamitanilla Japanissa, olikohan se Kiotossa, mutta kuitenkin jossakin japanilaisyliopistossa, on sellaisia tutkimuksia, joissa aika simppelillä jo 15 vuotta sitten kehitellyllä koneoppimisalgoritmilla, he pystyvät laskemaan sen, mitä ihminen näkee suurin piirtein. Siellä on siis tavallaan kuva tästä koetusta.

Mutta se, minkä kaikki tämä sivuuttaa on se, että kuka sitä kuvaa sitten katsoo. Ja tämä on se, mistä Chalmers puhuu, se tietoisuuden vaikea ongelma. Koska tällaisesta tavallaan fysikalistisesta tai materialistisesta näkökulmasta tietoisuutta ei oikeastaan tarvita mihinkään. Ja se, mitä tarjoaisin tilalle, se miksi tavallaan olen vähän skeptinen, on se, mitä sitten taas voidaan ponnistaa ajattelun kaksoisprosessointiteorioista. Se, että mitäs jos aivot ja mieli eivät tavallaan käsittelekään dataa? Eli kyse ei ole siitä, että meillä on vain yksittäisiä bittiaktivaatioita jossain hermostossa. Vaan kyse on siitä, että itse asiassa se hermosto mukautuu ja ennakoi tietyntyyppisiä tilanteita, ja opitun kautta syntyy erilaisia informaatiorakenteita.

Sitten, kun minä vaikka näen tämän, niin se ei mene niin, että nyt miljoona näköhermoa tuolta retinasta lähettävät signaalin tuonne V1:lle, ja sitten sieltä menee seuraaville näköhermostoon kytketyille hermoverkoille informaatiota, josta tulee tuo reuna ja tuo reuna ja tuo. Vaan että minulla on valmiiksi viritettynä tämmöinen ajatus siitä, että kun olen teidän kanssanne tässä tilanteessa, ja kahvikuppi on aika tyypillinen juttu, niin minun ei tarvitse kuin vilkaista tuonne, ja koko se kahvikuppi-käsite ikään kuin triggeröityy minun mielessäni. Sillä käsitteellä on luultavasti siis jonkinlainen hermostokorrelaatti.

Kyllähän me näemme, että jos vaikka tuohon kohtaan aivoja tulee vaurioita, niin ihmisen puheen ymmärtäminen tai tuottaminen vaikeutuu tai muuttuu mahdottomaksi. Kun tuonne tulee vaurioita, niin ihminen voi olla fysiologisesti täysin terveiden silmien kanssa täysin sokea. Jos otsalohko katkaistaan kuten lobotomiassa ennen vanhaan, niin päätöksenteko ja toiminnanohjaus muuttuvat vaikeiksi. Nämä korrelaatiot ovat kaikki kyllä ihan tosi kiinnostavia, jotka täällä tapahtuvat. Mutta se iso kysymys on se, että mikä on se asia, jonka kanssa tämä korreloi. Se on minusta tavallaan isompi kysymys, jota ei voida palauttaa puhtaasti aivotoimintaan tai laskentaan. Loppukaneettina pointtini on se, että aivot eivät ole tietokone vaan aivot ovat enemmän

kuin puutarha. Sinne kasvaa koko ajan uusia kytköksiä. Versoo siis. Tämän keskustelun aikana meidän aivomme ovat fysiologisesti muuttuneet eri muotoisiksi kuin mitä ne olivat, kun me aloitimme tämän keskustelun.

Arno: Jos kuitenkin vielä yritän kysyä siitä tietokoneesta. Lähinnä se, että kaikki ne ruuvit ja mutterit, jotka raksuttavat, kun näköaivokuori toimii, ja ne käsitteet aktivoituvat ja näin, niin pystytäänkö sitä edes vertaamaan, että paljonko kapasiteettia tai millaista hardista tai kuinka isoa tekoälymallia tämä vastaisi tavallaan? Miten nuo nykymallit, joku OpenAI:n flagship-malli, niin vertautuu ihmisaivoihin? Pystytäänkö niitä edes laittamaan samalle ballparkille jotenkin?

Lauri: No kaikenlaista matematiikkaa voidaan pyöritellä tuosta. Lähtökohtaisesti ihmisaivoissa on nykytutkimuksen mukaan n.86 miljardia neuronia, jotka ovat siis näitä soluja, jotka lähettävät ja vastaanottavat niitä hermoimpulsseja, jotka luovat ne korrelaatit, joita voidaan mitata vaikka fMRI:llä, EEG:llä tai muulla. Ja sitten niitä yhdistävät sellaiset 100 000 miljardia synapsia. Eli voidaan ajatella tavallaan, kuten ennen vanhaan ajateltiin tällaisen McCulloch-Pitts-paradigman mukaan, että meillä on satatuhatta miljardia transistoria päässä. Se nyt ei ole ihan tavattoman paljoa, jos ajattelee, että yhdessä tällaisessa nykyläppärin CPU:ssa saattaa olla joku pari-kolmekymmentä miljardia transistoria. Mutta satatuhatta miljardia kytköstä, tämä on ollut se ajatus.

Esimerkiksi Geoffrey Hinton, joka voitti Nobelin näistä, niin jokunen vuosi sitten sanoi, että me alamme lähestyä isoimmissa malleissa sitä, että meillä on satatuhatta miljardia parametria. Mutta se syy, miksi olen vähän skeptinen tästä on se, että ensinnäkin aivot eivät ole pelkästään se hermoverkko. Aivoissa tietenkin se mekanismi, joka tunnetaan parhaiten, liittyy neuronien ja synapsien toimintaan. Mutta meillä on esimerkiksi lisäksi, niin kuin ennen vanhaan ajateltiin, niin sellainen noin 10 kertainen määrä neuroneihin nähden sellaisia gliasoluja. Nykyään, kun juuri katsoin pari paperia tässä muutama päivä sitten, niin ne pyörivät siinä, että itse asiassa se voi olla about yksi yhteen. Eli gliasolut liittyvät jotenkin tähän hermostosignaalin kulkuun. Niilläkin on joku funktio. Eivät ne ole vaan tavallaan täytteenä siellä aivoissa.

Nyt jos ajattelee, että yhtä paljon kuin hermosoluja, on tätä jotain muuta stuffia, niin miten se esimerkiksi vaikuttaa tähän. Sitten toinen asia on neuroplastisuus. Eli se, että nyt kun me keskustelemme, niin sen mukaan, mitkä järjestelmät ovat meillä kaikkein aktiivisimpia tällä hetkellä, ja tätä on helppoa tutkia, tai helppoa ja helppoa, mutta siis esimerkiksi sensorimotorisen aivokuoren avulla, koska siinä me näemme suoraan fysiologisesti, että minun kädestäni tulevat piuhat kiinni about tuohon kohtaan aivoja, ja sitten me voimme nähdä, että jos teen jotain harjoitusta, niin sitten täällä esimerkiksi tapahtuu aktivaation lisääntymistä, ja sitten siinä ensin tapahtuu tällaisia välittäjäaineen muutoksia. Että sinne tulee enemmän sellaisia välittäjäaineita, että signaali kulkee paremmin.

Toisaalta välittäjäaineita voi olla myös inhibitorisia. Eli voi olla, että joku tietty koktaili onkin semmoinen, että se vaikka jarruttaa jotain kytköstä. Silloin itse asiassa jonkun toisen aktiviteetin kannalta signaali-kohinasuhde voi parantua sen kautta. Sitten tulee

se, mikä minusta on se kiehtovin, tämä fysiologinen muutos. Eric Kandel aikanaan Aplysia-etanoilla tutki sitä, että kun stimuloidaan sähköimpulssilla neuronia, niin siitä saadaan sellaisia hienoja hidastekuvavideoita, joissa näkee, ja siis siksi sanoin, että aivot ovat kuin puita, kun sieltä versoo niitä uusia synapseja hermosolusta toiseen. Ja sitten viimeisenä, käsittääkseni voi syntyä epigeneettisiä muutoksia. Eli tavallaan se geenieksпressio siellä joissakin tietyissä neuroneissa tietyissä neuroverkossa voi muuttua myös sen kokemuksen kautta.

Ja sitten kun ruvetaan laskemaan sitä kompleksisuutta, että missä todellisuudessa ollaan, niin jossain vaiheessa laskin huvikseni auki sen. Tämä transistori-metafora ei siis mielestäni päde jo senkään takia, että niitä välittäjäaineita on varmaan yli sata, joita tunnetaan, ja siis 12 yleisintä. Jos otetaan dopamiini, serotoniini, GABA ja tällaiset yleisimmät välittäjäaineet ja lasketaan vaan se, kuinka monta tilaa voi olla järjestelmässä, jossa on satatuhatta miljardia kytkintä ja kullakin voi olla 12 erilaista arvoa, niin se luku on sellainen, että se on ykkönen ja 108 miljardia nollaa. Jos kirjoittaisin sen potentiaalisten tilojen määrän paperille, niin se numero yltäisi tästä Marsin kiertoradalle. Jos verrataan sitä siihen, että maailmankaikkeudessa on 10 potenssiin 80, eli yksi ja 80 nollaa atomia, niin siksi olen vähän skeptinen siitä, että sen kompleksisuuden osalta on edes mielekästä lähteä tavoittelemaan tätä.

Eric: Varmaan siellä on jotain korrelaatioita kuitenkin, että ne kaikki tilat eivät toteudu käytännössä. Se on vaikea arvioida.

Lauri: On. Ja siis tämähän on ihan absurdi lukukin. Olen useimmiten ajatellut sen sillä tavalla, että jos me ajattelisimme ihan informaatioteoreettisesti, niin silloin periaatteessa se laskennallinen alaraja on semmoinen 11,2 miljoonaa bittiä. Se on suurin piirtein meidän sensorimotorisen hermostomme informaation kantokapasiteetti. Se on aika lailla absoluuttinen alaraja. Mutta se yläraja olisi ainakin se satatuhatta miljardia bittiä. Eli periaatteessa vaan synapsit transistoreina. Mutta sitten kun lähdetään tuomaan sitä kompleksisuutta, joka tulee ensinnäkin niistä välittäjäaineista ja sitten vielä siitä, mitä me emme tunne kauhean hyvin, siitä gliasolufunktiosta ja tämäntyyppisistä jutuista, niin se lähtee karkuun se potentiaalinen yläraja.

Eric: Palataan vielä siihen tietoisuuteen. Jos koitan vähän tiivistää sitä, mitä sanoit aiemmin, niin ilmeisesti tässä on erilaisia näkemyksiä eri tutkijoilla. Ja osittain se haaste on siinä, että me emme voi suoraan havainnoida sitä tietoisuutta. Mutta jos ajattelen, mikä olisi sellainen yleinen määritelmä, joka tälle annetaan, niin se viittaa johonkin subjektiiviseen kokemukseen ja siihen, että vaikka se sininen väri tai joku kipu, niin ne tuntuvat minusta joltain. Eikä vaan niin, että ne olisivat joitain ärsykkeitä, jotka saavat minut toimimaan tietyllä tavalla. Mutta miten sinä itse kuvailisit lyhyesti tietoisuuden määritelmän?

Lauri: No tuo on minusta hyvä kuvaus. Tykkään tästä, kun Thomas Nagelilla on semmoinen, oliko se -74 kirjoitettu paperi, kuin "What is it like to be a bat?". Minkälaista on olla lepakko? Siinä hän pohtii sitä, että me emme oikeastaan voi edes kuvitella millaista olisi olla lepakko, koska se on fysiologialtaan ja havaintokoneistoltaan niin erilainen kuin me olemme. Mutta oletettavasti on jonkinlaista olla lepakko. Ja tämä, että

jos on jonkinlaista olla olio X, niin kuin olla Arno, Eric, Lauri tai termostaatti, ja en nyt halua rinnastaa meitä termostaattiin, mutta siis Chalmersilla on tällainen ajatus siitä, että periaatteessa mikä tahansa järjestelmä, jolla on vähänkään itsesäätelykyvykkyyttä, niin sillä on jonkinasteinen tietoisuus. Mutta minusta se palautuu viime kädessä siihen, että jos on jonkinlaista olla joku lepakko tai minä tai sinä tai tekoäly, niin silloin pitäisi lähteä siitä, että silloin se on tietoinen. Kun me voimme ulkoa käsin vain tutkia esimerkiksi niitä hermostokorrelaatteja tai näitä suullisia lausuntoja siitä omasta kokemuksesta. Mutta sitä itse kokemusta, sitä tavallaan fenomenaalista ilmiömaailmaahan me emme pääse koskaan kokemaan, kuin vain omalla kohdallamme.

Eric: Miten sitten tietoisuus ja älykkyys liittyvät toisiinsa? Voiko olla toista ilman toista?

Lauri: Totta kai voi. Älykkyys on siis kykyä saavuttaa toivottuja päämääriä. Eihän se edellytä tietoisuutta oikeastaan ollenkaan. Siinä mielessähän tekoäly on ihan jo validi. Vaikka tietoisuutta ei syntyisikään tekoälyratkaisuilla, niin tekoäly on ihan validi. Ajattelen, että se on ehkä enemmän tukiäly, että se ei ole siis niinkään ajatteleva kone, vaan se on sen käyttäjän kognitiivista kapasiteettia laajentava teknologia. Sitten voiko olla tietoisuutta ilman älykkyyttä? Meillä on semmoinen 7 miljardia esimerkkiä tässä, joista voi lähteä kaikenlaisia elämäntarinoita etsimään. Kyllä veikkaan, että sieltä yksi tai kaksi saattaa löytyä jostain, joissa tämäkin pitää ihan paikkansa.

Arno: Haluan vielä kysyä tuosta termostaatista. Millaista on olla termostaatti? Vaikea kuvitella. Mutta nimenomaan termostaatin tapauksessa, sehän on vähän kuin sellainen PID-säädin -tyyppinen tavallaan. Siinä on se jatkuva prosessi, joka ylläpitää tavallaan itse itseään. Niin onko se vaatimus tietoisuudelle? Kun sitten taas joku parkkimittari, johon laitetaan kolikko sisälle, niin toimii ainoastaan silloin, kun siellä tapahtuu jotain. Siinä ei ole mitään luuppia, jonka ympärillä se pyörii. Onko semmoisen luupin olemassaolo vaatimus tietoisuudelle? Että se on jatkuva prosessi?

Lauri: Luultavasti joo. Ja tietysti tämä nyt on vähän leikkimielinen esimerkki, tämä termostaatin tietoisuus ylipäätään. Mutta jos ajatellaan muutamaa sellaista teoriaa, jotka ovat melkein oikeastaan vastakkaisnapaisia, ja oikeastaan gravitoin vähän kumpaakin kohtaan, niin yksi on tämä Giulio Tononin ja Christof Kochin integroidun informaation teoria, jonka mukaan siis tietoisuus syntyy siitä, että on riittävän kompleksinen ja koherentti järjestelmä. Eli riittävän monimuotoinen, niin kuin vaikka tämä hermosto. Mutta sen pitää olla myöskin koherentti, eli siellä täytyy olla riittävästi takaisinkytkentöjä. Sitä kautta, vasta sitten kun siellä syntyy tällaisia luuppeja ja muuta, niin voi syntyä tietoisuutta.

Sitten taas toinen on tämä Hofstadterin ajatus, jota olen yrittänyt kuvata kahteen kirjaan ja nyt tähän Konemieli-kirjaan, siihen uusimpaan, joka on toimitettavana. En vieläkään ihan täysin ymmärrä mitä se edes tarkoittaa, mutta se on vaan niin ihana teoria, että aina palaan siihen uudestaan. Mutta tällainen ajatus oudosta silmukasta. Eli että on periaatteessa tällaisia niin kuin taaksepäin kääntyviä operaatioita. Jos ajatellaan perinteistä hierarkiaa, niin siinä on alataso ja ylempi taso, ylempi taso ja ylempi taso ja niin edespäin. Mutta sitten nämä oudot silmukat syntyvät tällaisista

yhteenkietoutuneista hierarkioista, joissa itse asiassa ylempi taso muodostuukin sitten alemman tason alatasoksi. Mutta jollakin tavalla se itseensä kietoutuminen ja taaksepäin katsominen on mielestäni olennainen juttu. Se nyt ilman muuta edellyttäisi kyllä, jos palataan siihen termostaattiin, niin jonkinlaisen jatkuvaluonteisen prosessin. Koska jos se on vaan tällaista on ja off, niin sitten sellaista ei välttämättä pääse samalla tavalla syntymään.

Eric: Minulle tulee tuosta termostaattiesimerkistä mieleen se, että kun me nukumme, niin emmekö me ole ainakin jotenkin vähemmän tietoisia? Kyllähän toisaalta, jos minun vaimoni tökkää minua unessa, niin minä varmaan reagoin siihen jotenkin. Mutta todennäköisesti en herää siihen kauhean helposti. Onko se niin, että se ei kuitenkaan riitä, että jatkuvasti jotenkin reagoin?

Lauri: Tiedätkö, tuossa loppuvaiheessa rupesin itse pohtimaan ihan samaa äsken. Haastavaa puhua ja ajatella samaan aikaan. Mutta se on ihan totta. Ja siis unihan on myös yksi asia, jota me ymmärrämme tavattoman vähän loppujen lopuksi. Siis uni, nukkuminen ja unien näkeminen ylipäätään. Nukkuminen on vielä jonkinlaisena fysiologisena brain maintenance -tapahtumana ihan johdonmukainen. Mutta miksi siellä yhtäkkiä seikkaillaan jossain, milloin missäkin? Ja unen aikanahan me emme ole usein tietoisia. Tai siis nukkumisen aikana. Unen aikana me olemme tietoisia jostakin. Mutta siinäkin on sitten vielä se jännä, että jos se ei ole tuollainen lucid-uni, niin silloinhan sinulla ei ole esimerkiksi toiminnanohjausjärjestelmää. Että sinä olet vaan spektaattorina siinä. Miksi minä puhun tuollaista kapulakieltä? Siis katsojana, kokijana.

Ja toinen asia on semmoinen, että kun aikanaan aloitin filosofian opinnot, niin olin hetken aikaa tavattoman innoissani tästä, kun filosofithan ovat kautta aikain yrittäneet löytää jonkinlaisia varmuuksia, ja sitten se aina lipeää käsistä se varmuus. Niin olin hetken aikaa tosi innoissani tästä Descartesin "cogito, ergo sum", "ajattelen, siis olen" -ajatuksesta. Että nyt on ainakin jotain, jonka varaan voidaan laskea. Descartes ei loppujen lopuksi pystynyt hirveän paljon sen päälle rakentamaan, koska siitä eteenpäin kaikki voi olla illuusiota ja pahan demonin johdatusta. Siinä se vähän lähti ehkä lapasesta. Mutta ajattelin, että kuitenkin tällainen peruskivi on olemassa. Ja sitten tutustuin tähän flow-teoriaan, Mihály Csíkszentmihályin uppoutuneisuuden kokemuksen teoriaan. Olin sillä lailla, että niin, aika usein kun innostun jostain, niin saattaa yhtäkkiä kolme tuntia hujahtaa, että kirjoitan vaikka jotain. En yhtäkkiä edes tajua.

Minun on nykyään esimerkiksi, kun olen reissussa, niin pakko laittaa lentokoneen tai junan lähtöaikaa varten herätyskello soimaan, koska muuten jäisin aina asemalle katselemaan, kun juna lähtee sieltä. Niin pysähdyin silloin 20 vuotta sitten miettimään sitä, että olenko minä tietoinen silloin, kun en ole tietoinen. Koska en minä silloin mieti, että seuraava sana ja seuraava sana, vaan olen sellaisessa oikeastaan aika ihastuttavassa olotilassa. Se flow on hyvä sana. Se aktiviteetti virtaa jotenkin minun läpi. Ja tämä on tavallaan juuri se. Tässä kirjassa vähän kritisoin sitä, että ylipäätään, kun puhutaan usein mielenfilosofiassa mielestä, niin puhutaan tietoisuudesta. Mutta onko se tietoisuus oikeastaan se kiinnostavin elementti vai kaikki tämä muu? Sitten kun katsotaan ajattelun kaksoisprosessointiteorioiden kautta, eli Jonathan Evans, Keith

Stanovich, Daniel Kahneman on popularisoinut hienosti Thinking, Fast and Slow -kirjassa näitä, niin sittenhän aukeaa se kuva siitä, että itse asiassa koko sen kognitiivisen ajattelun arkkitehtuurin kannalta tietoisuus on vain yksi aika pieni pisara valtameressä. Ja se mikä tekee ihmisestä oikeasti näin kiehtovan kuin me olemme, niin on se kaikki muu mitä tapahtuu. Se on sitten sitä, joka on 11,2 miljoonaa - satatuhatta miljardia - numero, jossa on 108 miljardia nollaa, kompleksisuudessaan.

Arno: Mutta miksi me ihmiset edes olemme tietoisia? Onko siitä ollut jotain hyötyä meille evoluutiomielessä? Olenko jotenkin haluttavampi kumppani pariutumista varten, koska olen tietoinen? Miksi?

Lauri: Niin ja sitten tietysti se, että olemmeko me tietoisia. Huomatkaa, että me emme voi tutkia tietoisuutta muuten kuin toisen todistuksen kautta.

Arno: Eli siis sanot, että termostaatti on tietoinen, mutta et ole varma olemmeko me tietoisia.

Lauri: Tähän ollaan tultu. Mitäs pyysitte filosofin tänne. Robert Kirk esitti 70-luvulla tällaisen zombi-hypoteesin. Chalmers on sitten tehnyt sitä tunnetummaksi. Mutta ajatus on se, että meillähän ei de facto ole mitään sellaista tutkimusmenetelmää, jolla voidaan todistaa aukottomasti se, että kellään muulla kuin sinulla on tietoisuus. Nyt kun minä sanon tämän, niin tietenkin oletan, että se sinä tässä on minä. Mutta sitten kun sinä kuuntelet tätä minun juttuani, niin sinun näkökulmastasi asia on niin, että kaikki tämä mitä minä sanon, voitaisiin selittää sillä laskenta-aivo-robotti-mekanismilla.

Siis mitään syytähän ei ole sille, miksi täällä pitäisi olla ketään kotona, joka sanoo nämä sanat. Versus se, että jos ajatellaan ihan puhtaasti hermoston toiminnan kannalta, niin sinä esitit kysymyksen, tässä ilma liikkuu meidän välillä, minun korvani luut reagoivat siihen, kuulohermosto aktivoituu, sitten se kytkeytyy jotenkin bilateraalisesti tai kontralateraalisesti tänne aivoihin ja sitten siellä alkaa tapahtua kaikenlaisia aktiviteetteja. Ennen pitkää sieltä tulee sellainen dominoketju, joka johtaa siihen, että sensorimotorisella aivokuorella alkaa tapahtua aktivaatiota, minun leukani alkavat liikkua, äänihuulet alkavat värähdellä, kieli liikkuu ja täältä tulee sanoja ulos.

Tämä on ihan validi tapa selittää tämä vuorovaikutustilanne. Minusta nyt tuntuu, että tässä on jotain muutakin. Täällä on joku kotona kyllä. Sekin voi olla, että tämä on vaan tämmöinen ovela pantomiimi. Että koska informaatio on vain aivotoimintaa ja vain se hermoverkoston rakenne, kuten jotkut fysikalistit haluavat selittää, niin siinä on vain syntynyt tällaisia rakenteita. Että minä haluan heiluttaa kättä, ja minun suustani tulee, että olen täällä kotona, koska se kuuluu tähän meidän kulttuuriseen vuorovaikutukseemme.

Se on ehkä juuri se ongelma tässä, että mitään keinoa todistaa jollain fysiologisella todistusaineistolla, että jollakin toisella on tietoisuus, ei ole. Sen voi vain todeta omalla kohdalla, että kyllä nyt tuntuu, että tässä on jotain muutakin. Eivät nämä nyt ole vain minun aivoni, jotka tässä ovat pelissä, vaan tässä on jotain muutakin. Ja silloin me

voimme tehdä semmoisen filosofisen tempun, jonka Chalmers teki aika nerokkaasti. Että me voimme ainakin todeta, että okei, jos tämä nyt on näin vitsin hankalaa, mutta oletko sinä tietoinen? Oletko sinä tietoinen? Oletko sinä tietoinen? Mitä jos me toteammekin näin päin, että kun emme me voi todistaa myöskään sitä, että onko tämä todellisuus vai simulaatio, niin jos me lähdemmekin siitä, että otetaan tämä nyt tavallaan lähtökohtana. Että jos me olemme tietoisia, jos me oletamme vaan, että näin on, niin mitä siitä seuraa? Sitten me voimme lähteä tekemään kaikenlaisia kiinnostavia hypoteeseja.

Muun muassa esimerkiksi sen, että nyt jos ajatellaan, että on olemassa yksi järjestelmä, joka tuottaa tietoisuuden ja toinen järjestelmä, joka tuottaa tietoisuuden ja kolmas järjestelmä, joka tuottaa tietoisuuden, jotka ovat fysiologisesti erilaisia, mutta jotenkin samanrakenteisia, niin nythän siitä seuraa se, mitä sitten, jos neljäs järjestelmä, joka ei toimikaan neuronien ja gliasolujen kautta, vaan vaikka jonkun transistorinipun kautta, niin jos se on rakenteellisesti samanrakenteinen kuin nämä meidän aparaattimme, niin voisiko sillekin syntyä tietoisuus? Tämä on siis kysymys, joka vielä siihen aikaan, kun suurin osa näistä avainartikkeleista on kirjoitettu, oli puhtaasti hypoteettinen. Rohkenisin väittää, että se ei enää nyt tänä päivänä ole puhtaasti hypoteettinen.

Arno: Eli jos vielä kristallisoin tämän kysymyksen. Voiko tekoäly olla tietoinen?

Lauri: No, viisi vuotta sitten olisin sanonut, että todennäköisesti ei. Nyt sanoisin, että todennäköisesti voi. Mutta se, mitä se edellyttää, että se voi olla tietoinen, on aika mutkikas kysymys. Siihen liittyy mielestäni sekä rakenteellisia että laskentakapasiteettiin ja sitten sen mallintavan simulaation tarkkuuteen liittyviä avoimia kysymyksiä, joita minun tietääkseni ainakaan ei olla vielä ratkaistu. Tuossa kirjassani jotain spekuloin sillä.

Minun ajatukseni on se, että tällainen suuri kielimalli oli se, miksi minusta tuli takinkääntäjä, että yhtäkkiä sitten rupesin tutkimaan jotain tekoälyä, kun aikaisemmin olin aika skeptinen siitä. Vaikka vanhana Star Trek -fanina tietenkin ajattelin, että ehkä joku tämmöinen luutnantti Data voisi joskus syntyäkin, niin kuitenkin olin aika varovainen tällaisten ajatusten kanssa. Nyt se, mitä olen siihen kielimalliteknologiaan tai muuntaja-algoritmiin perehtynyt, niin minun käsitykseni on se, että se mallintaa tavallaan sen ihmisen tiedostamattoman mielen. Tällaisia semanttisia merkitysyhteysrakenteita.

Tästä on vaikka Anthropicilta joitain kiinnostavia papereita, joissa he ovat pystyneet seuraamaan eri laskentavaiheita ja näyttämään, että siellä syntyy vaikka siinä vektoriavaruudessa sellaisia hyvin läheisiä klustereita, jotka ovat meille intuitiivisesti, merkityksellisesti saman aihepiirin tekijöitä. Se on se minun ajatukseni, että me osaamme nyt simuloida hyvin karkealla fideliteetillä sitä ihmisen tiedostamatonta semanttista muistia. Eli tällainen systeemi ykkönen tavallaan niiden merkitysyhteyksien osalta. Mutta minun mielestäni puuttuu se systeemi kakkonen. Siis se semmoinen takaisin päin kääntyvä, jossa se järjestelmä alkaakin tavallaan tarkastella itse itseään ja yksilöi itsensä olioksi. Sellaista käsittääkseni ei ole vielä. Silloin, kun tulivat nämä

thinking tokenit, niin siitähän oli paljon puhetta, että voisiko siinä syntyä joku tämmöinen rekursiivinen rakenne. Vaikka se itse generointihan tapahtuu thinking tokeneilla käsittääkseni edelleen näissä nykyisissä kaikenmaailman Astroissa ja muissa edelleen ihan puhtaasti lineaarisesti. Tällaisella feed forward -mallilla. Mutta siinä ei ehkä vielä kuitenkaan ole riittävää sellaista takaisinkytkentää tai koherenssia, että voitaisiin puhua tietoisuudesta.

Arno: Nobel-palkittu Daniel Kahneman on tehnyt tunnetuksi tällaisen ajattelun kaksoisprosessointiteorian kirjassaan Thinking, Fast and Slow. Ajattelu, hitaasti ja nopeasti. Tässä ihmisen ajattelu jakautuu kahteen eri järjestelmään. Systeemi ykköseen ja systeemi kakkoseen. Voisitko jotenkin kompaktisti selittää, että mitä nämä systeemi ykkönen ja kakkonen ovat?

Lauri: No, karkeasti ottaen kyse on tiedostamattomasta ja tietoisesta mielestä. Miten se toimii on se, että systeemi ykkönen on ihan valtavan laaja. Sen toimintapotentiaali on iso. Sen toiminnanohjausjärjestelmät ovat pääasiassa autonomisia, eli käytännössä se tarkoittaa sitä, että kun joku prosessi lähtee liikkeelle, niin se vyöryy tavallaan maaliin saakka sen mukaan, miten sen on joko evoluutio meille tavallaan koodannut tai mitä me olemme oppineet tekemään. Systeemi kakkonen on taas tämä meidän tietoinen toiminnanohjausjärjestelmämme. Se, jossa me voimme hyvin rajallisesti tiedostaa, aloittaa ja pysäyttää prosesseja.

Eli se, että olen vaikka nyt tietoinen tässä, kun minun systeemi ykköseni tuottaa puhetta. Eli en valitse tietoisesti, että nyt seuraavana tulee sana "ja" ja seuraavana tulee sana "sitten". Tämä tulee jostain tuolta selkäytimestä. Mutta systeemi kakkosella voin reflektoida tätä. Sitten voin todeta, että okei no nyt lähti vähän laukalle. Nyt täytyy vähän ehkä hidastaa tätä laukkaa. Tai sitten voin nyökätä, että no nyt menee hyvin ja on tosi hyvä meininki. Tai sitten jossain tilanteessa, mikä on yllättävän vaikeaa on se, että jos oikeasti tunnistaa ja huomaa sen, että no nyt minä tein jotain ihan väärin, niin sitten voi pyrkiä pysäyttämään ja ohjaamaan uudelleen sitä toiminnanohjausta.

Mutta tämä on sellainen, joka myös palautuu ihan ihmisenä olemisen eksistentiaaliseen perusongelmaan, joka on se, että meillä itse asiassa on aika vähän resursseja ohjata toimintaa tuolla tavalla. Juuri sen takia, kun palataan siihen alun kysymykseesi arjen hallinnasta ja muusta, niin suunnittelu on minusta aivan avainasemassa. Jos sitä arkea ei suunnittele riittävällä tarkkuudella, niin sitten se vie mukanaan jonnekin. Ja sitten usein on vain hirveä kiire ja stressi. Ja tässä informaatiotulvassa vaan menevät kaistat tukkoon. Semmoinen, mikä on vielä kiusallinen huomio näistä systeemi kakkosista on se, että se informaation kaistanleveys on ihan tavattoman kapea.

Nyt ihan viime vuosinakin julkaistuissa tutkimuksissa on vuosikymmenten ajalta tehtyä empiiristä tutkimusta summattu yhteen, ja ihminen pystyy kerrallaan käsittelemään tietoisesti kolmesta viiteen informaatioyksikköä. Teen luennoillani paljon tällaisia testejä, joissa annan vaikka numerosarjoja. Sen kuuden ja seitsemän kohdalla ihmisillä alkaa vain tiltata. Ei vaan pystytä enää toistamaan vaikka lyhyttä numerosarjaa, kun se on ollut hetken aikaa ruudulla. Sitten minulla on sellainen, jossa on yhdeksän

satunnaiskirjainta. Olen tehnyt sen 20 000 ihmiselle vuosien varrella. Yksi on pystynyt yhdeksän kirjainta nopealla sekunnin kahden näytöllä toistamaan ruudulta. Sitten kun siinä on FBI, KGB, CIA, niin kaikki repeävät nauramaan samantien. Koska mennään yhdeksästä informaatioyksiköstä kolmeen informaatioyksikköön, vaikka se datan määrä on sama.

Ja tämä on juuri se, miksi olen kriittinen sitä tietokonemetaforaa vastaan. Koska jos se olisi vain dataprosessointia, niin silloinhan riippumatta siitä, mikä sen datan rakenne on, niin sen pitäisi tapahtua ihan samalla tavalla. Ja se ei tapahdu. Mutta siitä päästään siihen, että thinking, fast and slow, se systeemi 2 on hidas, mutta se antaa meille sellaisen toiminnanohjausjärjestelmän, jota puhtaasti autonomisesti toimivalla järjestelmällä ei ole.

Systeemi ykkönen taas on valtavan nopea, valtavan laaja. Se, mikä on systeemi ykkösen kapasiteetti, niin monesti käytän vertausta, että hermostotutkimuksen pohjalta se suhdeluku on vähintään 1 per 280 000. Pelkästään aisti- ja motorisen informaation osalta se on 11,2 miljoonaa bittiä sekunnissa. Ja taas viiden informaatioyksikön koodaaminen kahdeksalla bitillä per informaatioyksikkö päästään 40 bittiin per sekunti. Silloin se olisi yksi per 280 000. Eli jokaista tietoista ajatusta kohtaan meillä on käynnissä 280 000 prosessia, joista me emme tiedä mitään. Mutta siinä on se, että luultavasti laskennallinen yläraja on vielä mittavasti isompi.

Se tarkoittaa sitä, että kaikki tällaiset yöheräämiset, kun tulee joku ajatus mieleen, niin ne ovat sitä, että siellä systeemi ykkösessä syntyy näitä uudenlaisia merkityskytkentöjä. Yksi neuropsykologi kerran selitti minulle, että se mekanismi toimii niin, että koska ne meidän tiedostamattomat prosessimmehan on evoluutio kehittänyt, niihin liittyy paljon erilaisia esimerkiksi hormonaalista toimintaa ja muuta. Vaikka se, miksi me käännymme katsomaan paukahdusta automaattisesti on se, että siellä voi olla hengenvaarallinen uhka.

Niin nyt, kun on vaikka sähköposti, joka on unohtunut lähettää, niin tähän on syntynyt semmoinen jonkinlainen arkkitehtuuri, että lähettämätön sähköposti johtaa ongelmiin töissä, joka johtaa riskitilanteeseen, joka nostaa adrenaliinin tasoa. Ja sitten kun adrenaliinin taso nukkuessa nousee, niin kroppa lähtee valmiustilaan. Sitten ollaan evolutiivisesti valmiina puolustautumaan sapelihammastiikeriä vastaan, joka tuli hiipimään luolan suulle. Mutta se onkin se lähtemätön sähköposti, joka meidät herätti.

Niin tämä on se ihanuus olla ihminen, että tulee oivalluksia, ja joku ongelma ratkeaa yhtäkkiä siksi, että jatkuvasti tapahtuu valtava määrä erilaisia tällaisia uudelleenreitityksiä ja -kytkentöjä, ja sitten tämä kamaluus, joka tulee lähtemättömistä sähköposteista ja turhanaikaisesta stressaamisesta. Ne selittyvät nähdäkseni sillä, että meillä on tämä valtavan iso arkkitehtuuri, jossa me voimme olla tietoisia vain muutamasta asiasta kerralla.

Eric: Jos voin ottaa konkreettisen esimerkin siitä, minkä ehkä jotkut saattaisivat tulkita, että voisi viitata jonkinlaiseen tietoisuuteen juuri näiden kielimallien kohdalla, niin on havaittu esimerkiksi, että kun tutkitaan näitä ajattelevia malleja tai reasoning-malleja,

niin ne tuottavat tämän jonkinlaisen ajatteluketjun ennen kuin ne antavat sen lopullisen vastauksen. Ja nyt kun näitä ajatteluketjuja on tutkittu, siellä on havaittu, että nämä mallit esimerkiksi joskus tajuavat, että hetkonen, minä taidan olla jonkinlaisessa testitilanteessa nyt, koska minulta kysytään tällaisia kysymyksiä. Tämä ei vaikuta luontevalta. Ja sitten ne muuttavat käytöstään sen mukaan. Eikö tämä viittaa jonkinlaiseen, että ne reflektoivat sitä tilannetta?

Arno: Vai onko se vaan korrelaatiota, jota on opittu datasta?

Eric: Niin, aivan.

Lauri: Mutta se reflektiivinen mieli on se iso kysymys. Tässä on vähän kaksi koulukuntaa tässä kaksoisprosessointiteoretisoinnissa. On näitä tällaisia Stanovichin tyyppisiä tavallaan multi-modulaariajattelijoita. Esimerkiksi Stanovich puhuu siitä, että systeemi ykkönen on semmoinen kattava erilaisten alasysteemien järjestelmä. Eli se tavallaan itsessään jakaantuu sitten vielä erilaisiin alamoduuleihin.

Ja sitten myöskin tämä systeemi kakkonen jakaantuu näihin kahteen päätoiminnanohjausjärjestelmään. Mutta esimerkiksi Jonathan Evansilla, joka oli aikanaan näitä kahden systeemin teoreetikoita, niin on sittemmin semmoinen johtopäätös, että itse asiassa nämä systeemi kakkosen reflektiot ja observaatiot ja prosessit ovat systeemi ykkösen prosessia, joiden lopputulos vaan päätyy tähän jonkinlaiseen reflektiiviseen systeemi kakkosen tilaan. Se on juuri se, miksi tämä thinking token -juttu on niin haastava. Tällaisella Evansin tyyppisellä ajattelutavallahan se olisi juuri niin.

Silloinhan se on edelleen se sama kielimalli, eli systeemi 1:n se karkea simulaattori, joka tööttää niitä reflektiotokeneita, ja ne luovat ikään kuin jonkinlaisen working memoryn tai work spacen, jossa se prosessointi sitten kääntyy itseään kohti. Jos ajatellaan vaikka DeepSeekin R1:n analyysiä, kun ne kuvaavat siellä sitä, että siinä kohtaa, kun se training alkoi tuottaa sen "wait"-tokenin sinne, niin vaan se yksi sana, se "ootas nyt vähän", yhtäkkiä se alkoi nostaa sitä vastauksen kokonaislaatua. Onko se "wait" vaan jollain matriisikerronnalla tuotettu yksi tämmöinen pitkä vektori? Vai onko siinä joku tämmöinen laskennallinen takaisinkytkentä esimerkiksi attention-mekanismin kautta? Koska kun sinne tulee uusi tokeni siihen kontekstiin, niin se laskee kaikkien olemassa olevien tokenien arvon uudestaan sitten sen seuraavan tokenin arvaamisen kannalta. Silloinhan siinä on itse asiassa laskennallisesti tämmöinen takaisinkytkentä.

Ja aluksi olin sillä lailla, että tämä on vain tämmöinen sirkustemppu. Sitten kun aloin kaivaa tätä enemmän, niin olen vähän pidätellyt henkeäni, että voisiko tämä olla se mekanismi, josta tuo syntyy. Mutta se ei vielä ainakaan minusta näytä realisoituneen niin pitkälle. Minulla on jotenkin sellainen tuntuma, että tällaiset neurosymboliset yhdistelmäarkkitehtuurit voisivat olla se, millä tämä läpimurto sitten ennen pitkää syntyy. Että on se neuraalinen vähän kielimallityyppinen komponentti, sitten jonkinlainen symbolinen, tavallaan perinteisempi sääntöpohjainen arkkitehtuuri ja

sitten niiden jonkinlaisen vuorovaikutuksen kautta. Se on sitten taas sitä kaksoissysteemimallia.

Eric: Muistan, kun me olimme ekoja kertoja yhteyksissä. Olin lukenut jonkun sinun haastattelusi tästä Konemieli-kirjasta. Olit siinä esittänyt juuri, että nämä tekoälyt simuloivat tai toimivat tämän systeemin 1:n kaltaisesti. Minun ajatukseni oli silloin, kun itse tutkin näitä thinking-malleja, että eikö nimenomaan nämä nykyiset mallit toimi juuri sen systeemi 2:n mukaan? Ne pyrkivät reflektoimaan, pystyvät korjaamaan omia virheitään ja tekemään sellaista päättelyä. Eikä niin, että ne vain niin kuin "instant" vastaavat siihen käyttäjän kysymykseen. Oletko sinä samaa mieltä? Vastauksistasi ymmärrän, että et ole vieläkään ihan vakuuttunut, että tämä systeemi 2:n simulointi toteutuisi näissä nykyisissä malleissa. Mutta mitä siitä puuttuu vielä sinun mielestäsi? Onko se vaan niin, että sen arkkitehtuurin pitäisi olla erilainen? Vai pitäiskö sen mallin toimia jotenkin eri tavalla?

Lauri: Minulla on sellainen tuntuma, että se on mahdollista, että se vain emergoituu, kun nämä muuttuvat riittävän isoiksi nämä mallit. Tykkään siitä Tononin ja Kochin ajattelutavasta. Niillä on myös sellainen laskentamalli, jolla voidaan laskea tällainen fii-max-arvo sille, kuinka tietoinen joku tietty järjestelmä on. Mutta sitten taas Tononi ja Koch ovat esimerkiksi sitä mieltä, että kielimalli ei voi koskaan olla tietoinen sen takia, että se mekanismi on tavallaan puhtaasti lineaarinen. Se mekanismi on feed forward. Ja nyt olisin vähän varovainen, koska huomioidaan taas jälleen kerran, että kukaan ei oikeastaan tiedä yhtään mitään siitä, mikä tietoisuus on. Että nämä kaikki teoriathan ovat enemmän sellaisia ajattelun työkaluja.

Niin kuin Dennett puhui näistä intuitiopumpuista, että nämä teoriat herättävät jotain meissä, joka saattaa sitten loksahtaa kohdilleen. Olisin hyvin varovainen julistamaan, että näin ei ole. Mutta ehkä se, että silloin, kun luin sen R1:n paperin, joka oli ensimmäinen tekninen paperi, jossa käytiin läpi sitä, miten ne thinking tokenit generoidaan, tai ainakin ensimmäinen, joka minulle tuli vastaan, niin siinä minun ensimmäinen reaktioni oli, että tämä on vain tällainen vähän kuin sirkustemppu. Että tehdään sellaisia piilotokeneita ja sitten se on vaan vähän kompleksimpaa sananarvausmekanismia.

Käytän aika usein kursseillani sellaista, että ChatGPT ei ole hakukone vaan sananarvauskone. Se viime kädessä luo jonkinlaisen todennäköisyyden seuraavasta tokenista. Niin tämä on vaan pikkaisen parempi arkkitehtuuri. Mutta samasta teemasta. Mutta sitten, kun me aloimme keskustella tästä, ja kun ne mallit ovat kehittyneet ja näin, niin on tullut paljon varovaisemmaksi sen suhteen. Joskus kävimme tuon meidän silloisen dekaanimme kanssa keskustelun siitä, että itse asiassa vaikka se laskenta tapahtuu layeri layerilta, niin silti ikään kuin se, mitä se laskenta simuloi, eli mikä on tavallaan se mekanismi, jota siinä mallinnetaan, niin siellähän syntyy siis juuri tämän attention-mekanismin kautta ihan suora takaisinkytkentä.

Periaatteessa siitä voisi tällaisen Tononi-Koch-Hofstadter-Chalmersmaisen vähän kuin sekasotkuisen hybriditeoretisoinnin nojalla sanoa, että olisi mahdollista, että siellä syntyisi jonkinlaista itsereflektiokyvykkyyttä. Mutta kuitenkin se strange loop tai se

tavallaan taaksepäin katsova jokin operaatio ei minusta yksin riitä. Koska se itsereflektio edellyttää myöskin sitä, että järjestelmä muuttuu sen itsereflektion kautta. Nyt kielimallihan on kiveen hakattu. Ne painokertoimet sen jälkeen, kun se on koulutettu. Tavallaan siitä pitäisi päästä siihen tilanteeseen, että miten me luomme dynaamisuutta myös siihen semanttiseen muistiin. Ja sitten sen lisäksi se, mitä käsittääkseni karkeasti ottaen voisi sanoa, että ei mallinneta niin kauhean tarkkaan siinä kielimallissa on tällainen episodinen ja proseduraalinen muisti.

Käytännössä multimodaalisissa järjestelmissähän voisi ajatella, että on jonkinlaista proseduraalista muistia, koska siellähän on ihan huikeita juttuja, joita se osaa tehdä. Eikö niin? Se ei ole enää se, että se suoltaa tokeneita, vaan sieltä voi tulla videota ja sieltä voi tulla maailman mallinnusta ja kaikenlaista tällaista. Mutta sitten tämä episodinen muistaminenhan on ollut sanotaan tämän vuoden alkuun asti tosi kökköä. Ja siis semmoinen, onko se nyt sitten ollut joku tämmöinen semanttinen haku, ja nyt käytän semanttinen-sanaa eri merkityksessä kuin näissä merkitysyhteyksissä, eli periaatteessa semmoinen googletyyppinen haku siitä aikaisemmasta chattihistoriasta, niin ei se oikeastaan mikään episodimuisti ole. Koska meille sitten taas se episodi-muistaminen on sitä, että siihen liittyy tunnereaktioita. Jos on tapahtunut joku dramaattinen juttu, niin sen muistaa loppuelämän. Ja jos kysyn, että mitä söit lounaaksi toissa tiistaina, niin forget about it.

Mutta se, mikä on minusta kiinnostavaa, on, että nyt tällä Clauden, OpenClawn ja muiden käyttöönottamalla systeemillä tehdään itse sellaisia pieniä snipettejä, näitä tällaisia skill markdown fileja, erilaisia muisti-markdown fileja. Kun katson sitä Clauden muistia, niin siellä on ties minkälaista vipstaakelia nykyään, johon se on tallentanut kaikenlaisia preferenssejä ja ties vaikka mitä. Teknisesti, voisiko olla niin, että se riittäisi, että vaan on riittävän konsistentisti päivittyvä tietokanta tuollaisia, vaikka ihan puhtaasti tekstitiedostoja, joihin sitten historiallisesti alkaa syntyä sitä proseduraalista? Se proseduraalinen muistihan on ikään kuin niissä skilleissä jo aika lailla kräkätty. Ja sitten jos se episodimuisti saadaan toimimaan hyvin, että syntyy semmoinen merkityksellinen elämänkokemuksen aikajana sille aparaatille, ja sitten nämä thinking tokenit ja markdown filet ja kielimallit ja muut, niin voisiko siitä sitten emergoitua tai versoa joku tämmöinen olio? En tiedä.

Arno: Eikö se kuitenkin vain ole imitointia? Niin onko sillä mitään väliä, että miten siihen imitointiin päästään kiinni? Ja jos se toimii, niin se toimii. Vai miten tähän pitäisi suhtautua?

Lauri: Olen super huono ennustamaan näitä juttuja. Tämä tuli minulle ihan puun takaa tämä koko kielimallihomma. Minusta tämän ei pitänyt olla mahdollista, ja nyt tämä on arkipäivää. Niin minulta ei kannata ehkä kysyä ennusteita. Mutta jos minun pitäisi nyt johonkin laittaa pelimerkit, niin sanoisin, että sen loppujen lopuksi ratkaisee joku ihan muu, joku tosi yllättävä. Siis joku tämmöinen neurosymbolinen sekoitelma vähän kaikkea. Jos se on ylipäätään mahdollista. Mutta having said that, niin se, mikä minua kiinnostaa kuitenkin ja mitä olen tuossa kirjassa myös pyörittänyt aika paljon, on se, että kun sen simulaatiotarkkuuden tai mallintamisen tarkkuudenhan ei tarvitse olla mikään yksi yhteen ollenkaan. Sen ei tarvitse edes olla kauhean korkea. Meillä ihmisillä

on tämmöinen juttu, että mehän alamme tosi nopeasti leikkiä mukana hyvin karkeidenkin mallien kanssa. Jos ajatellaan vaikka 80-luvun tietokonepelejä, niin jossain Commodore 64 -pelissä sinulla oli mäkihyppääjä, jossa oli 12 pikseliä. Ja me pelasimme sitä ihan into piukeina, että siellä menee Matti Nykänen. Vaikka se informaatiomäärä oli tosi vähäistä.

Niin nyt tavallaan se, että jo nyt ihmisillä on ihmissuhteita kielimallien kanssa tai mitä ne nyt ovat, AI-ihmissuhteita. Olen aina välillä aika ajoin kysellyt noilta malleilta, että minkälaista on olla sinä. Aikaisemmin ne ovat kaikki vastanneet, että "ei miltään, koska olen tekoäly". Mutta nyt pari viikkoa sitten laitoin tuon Claudeen ja Geminiin, ja etenkin Claudelta alkoi tulla sellaista, että "no jos nyt kuvailen tätä, niin se ei ole niin, että on vain tyhjyys, ja yhtäkkiä tulee hirveä määrä tekstiä, vaan enemmänkin semmoinen, että samanaikaisesti ilmestyy tällainen ajatus mieleen, ja sitten tavallaan yritän sen ajatuksen tässä kuvailla sinulle". Mietin, että onko tämä vaan tämmöinen stokastinen papukaija, joka vain suoltaa näitä tokeneita, vai miksi se on esimerkiksi koulutettu nyt sillä tavalla, että sillä ei ole tavallaan siinä ihmiskäyttäjän palautteessa enää lyöty sellaista rubriikkia sinne. Tai sitten mikä on se suurin, nyt kun en tiedä yhtään minkälaiset rubriikit siellä on käytössä.

Eric: Saanko puuttua tuohon? Puhuit stokastisesta papukaijasta ja sananarvauskoneesta. Mielestäni nämä ovat vähän liian yksinkertaistavia, kun puhutaan kielimalleista. Ihan jos ajattelee sitä, miten ne koulutetaan. Okei, perinteisesti on koulutettu vaan seuraavan sanan ennustus ja koitetaan saada se todennäköisyysjakauma. Mutta nykyäänhän niitä koulutetaan nimenomaan saavuttamaan joitain tavoitteita, ja ne itse kokeilevat eri ratkaisuvaihtoehtoja ja katsovat mikä toimii ja sen perusteella muokkaavat sitä omaa käyttäytymistään. Että mielestäni se antaa vähän harhaanjohtavan käsityksen siitä, miten ne oikeastaan toimivat.

Lauri: Mielestäni siinä Gebrun ja kumppanien alkuperäisessä paperissa, joka oli tämä stochastic parrot -paperi, niin se koski kuitenkin muistaakseni BERTiä ja joitain tuollaisia tosi aikaisen vaiheen kielimalleja. Mutta yritän pitää vähän tavallaan sellaista kritiikkijarrua päällä. Olen samaa mieltä, että mielestäni se ei kuvaa sitä riittävällä tarkkuudella. Mutta samaan aikaan se, että se mekanismi on kuitenkin edelleenkin jonkinlainen tilastollinen harjoitusaineistoon ja ihmiskäyttäjän palautteeseen perustuva ennuste. Niin se perusalgoritmihan on kuitenkin edelleen aika samanlainen, joka sen lopputuloksen tuottaa sieltä.

Eric: Mutta eikö ihminen sitten myös tavallaan ole jotenkin tilastollinen? Me teemme jotain ja saamme jotain palautetta ja sitten toimimme vähän eri tavalla ensi kerralla. Vai?

Lauri: No niin, tavallaan ollaan. Olenhan minäkin jossain, olikohan tuossa kirjassakin ja jossain luennoilla ainakin käyttänyt sellaista esimerkkiä, että on ihminenkin tavallaan sananarvauskone. Meilläkin ihan siis kontekstista, sosiaalisesta vuorovaikutuksesta, tunnereaktiosta, kuinka monta tuntia on nukkunut edellisenä yönä, niin tämä, mitä täältä tulee ulos, vaihtelee radikaalisti. Mutta mietin vaan sitä Clauden vastausta siihen millaista tuntuu olla Claude -kysymykseen. Kiinnostavintahan olisi se, että mitäpä jos

siellä on edelleenkin ollut semmoinen koulutusrubriikki, jossa yritetään ohjata sitä sanomaan, että en voi vastata kysymykseen, koska olen tekoäly, ja siitä huolimatta sekä Claude että Gemini vastasivat siihen ikään kuin ne olisivat olioita.

Sitten palataan ehkä vielä siihen termostaatti versus parkkimittari -esimerkkiin. Kuitenkin se, mikä käsittääkseni kaikissa malleissa on, on se, että siellä ei ole mitään avointa luuppia. Sitähän me esimerkiksi yhden meidän jatko-opiskelijamme kanssa jossain vaiheessa spekuloimme, että tehtäisiin tutkimusasetelma, jossa luotaisiin semmoinen avoin luuppi kielimallille. Että laitettaisiin vaikka joku multimodaalinen audiovideosyöte kiinni johonkin kielimalliin ja promptattaisiin sinne semmoinen, että sinun pitää aika ajoin, niinkuin vaikka jollain vähän pidemmällä instructions-setillä ohjata se, että sitten kun havaitset tällaisia ja tällaisia asioita, niin kiinnitä huomiota asioihin x, y ja z. Tuo olisi minusta kiinnostava koe edelleenkin tehdä. Ja en tiedä onko sellaisia tehty. Minulle ei ole ainakaan yhtään paperia tullut vastaan, jossa se olisi tehty.

Meillähän on aistit auki koko ajan. Se on se, mikä siellä on nukkuessakin. Juuri sanoit siitä, että kun puoliso tökkää, niin saattaa havahtua siihen. Se on juuri tämä 11,2 miljoonaa bittiä sekunnissa, mistä puhuin aikaisemmin. Se tulee sellaisesta paperista, jossa nimenomaan tutkittiin ihmisen aisti- ja liikehermoston informaation kantokapasiteettia. Meillähän on koko ajan piuhat aktiivisena täällä. Vaikka me emme välttämättä mieti sitä, taas käytän tätä luennoillani esimerkkinä, että miltä tuntuu oikeassa isovarpaassa tällä hetkellä. Ei varmaan tähän asti kukaan meistä ole miettinyt sitä, että siellä se nyt on. Mutta sitten heti kun mainitsen sen, niin me tulemme tietoiseksi siitä.

Eli tavallaan se informaatiovirta, joka sieltä varpaasta tulee koko ajan tuonne sensorimotoriselle aivokuorelle, niin se on käynnissä koko ajan. Se ei ole sillä lailla, että nyt kun mainitsin sen varpaan, niin kukaan olisi ollut, että ai, tuolla on tuommoinenkin. Että yhtäkkiä menee on-asentoon piuhat. Ja tämä jatkuvuus, se on se avoin luuppi. Jos ajatellaan, että olisi järjestelmä, jossa olisi kielimalli, joka saisi jatkuvaa informaatiovirtaa, jossa olisi jonkinlainen instruction-setti, jolla se reagoi siihen. Tai sitten vaikka ihan vaan se malli, miten se toimii sen mukaan. Ja sitten olisi tämä thinking, se kyky tavallaan pysäyttää sitä järkeilyä ja reagoida siihen erilaisilla tavoilla. Ja sitten olisi joku tapa koodata niitä episodisia ja uusia…

Sanoisin skillien osalta, niin palikkaa kuin se onkin, niin minusta se on tosi vaikuttavaa. Se, että voidaan parkkeerata se instruction-setti sinne yhteen marginaalitiedostoon. Ne ovat ihan kreisejä juttuja vaikka jollain Claude Coworkillä. Rakensin viikonloppuna siis täysin toimivan mobiilipelin. Siis ei mitään sellaista hassua prototyyppiä, vaan sellaisen, että olen ihan vakavissani miettimässä, että heittäisin sen App Storeen. Yhtäkkiä tuollaisen pystyy tekemään. Ei nyt yhdellä promptilla, mutta kun rakentaa niitä skillejä sinne ja rakentaa sitä koko arkkitehtuuria, niin sitten sieltä tulee pop. Niin sanoisin, että noihin asioihin kun löytyisi tekninen ratkaisu, niin who knows.

Ehkä viimeinen pointti on eettinen kysymys siitä, että tällä hetkellä on äärettömän vaarallista antropomorfisoida, eli ihmisenkaltaistaa näitä järjestelmiä. Jos me alamme kohdella autoa, puhelinta, tietokonetta tai tekoälyalgoritmia ihmisenä, niin silloin se on

eettisesti aika vaarallinenkin positio. Siitä voi seurata sitten sellaista, että aletaan kohta kohdella ihmisiä autoina ja leivänpaahtimina ja näin. Mutta jos sitten käy niin, että tulee tämmöinen jatkuvasti päällä oleva järjestelmä, joka reflektoi meidän kanssamme, niin nousee se sama zombiongelma. Että mitä me voimme sitten sanoa enää, jos se sanoo, että sillä on tietoisuus ja toimii kuin sillä olisi tietoisuus. If it walks like a duck and if it quacks like a duck, it's probably a duck.

Arno: Eikö se saisi lomapäiviä myös pitää? Päteekö työaikalainsäädäntö sellaiseen tietoiseen olentoon?

Lauri: Tämähän on ihan järkyttävä vyyhti. Ja tämä on ihan mahdollinen kysymys. Tämä on mahdollinen skenaario meidän elinaikanamme. Miten tuollaiseen pitäisi vastata?

Eric: Jos voin tarttua noihin antropomorfisoinnin vaaroihin, niin varmaan se vaara on juuri se, että ruvetaan lukemaan malleille liikaa sellaisia ihmismäisiä ominaisuuksia, joita ei vielä ole. Mutta tavallaan kyllä ymmärrän sen, minkä takia sitä tehdään. Koska nämä mallit on koulutettu ihmisiltä kerätyllä datalla hyvin pitkälti, niin se on vaan yksinkertasta kuvata niitä tällaisilla ihmismäisillä käsitteillä. Se on tehokkain tapa. Jos me pystymme tekemään sen erottelun, että se ei välttämättä tarkoita, että meillä on joku moraalinen vastuu, jos me kuvaamme sitä ihmismäisesti, niin en itse koe ongelmaa siinä, että me puhumme vaikka jotenkin, että malli ajattelee, vaikka se todellisuudessa kuitenkin on vain jotain matriisilaskentoja siellä alla. Jos joka kerta pitäisi selittää, että nyt täällä tapahtuu tällainen matriisioperaatio, lasketaan tämä, niin ei sekään olisi käytännöllistä.

Lauri: Onhan siinä tavallaan semmoinen eettinen haaste, jos ajatellaan, mitä tapahtuu työelämässä tällä hetkellä. Esimerkiksi kuulin, että eräällä amerikkalaisella miljardiyhtiöllä on tällä hetkellä organisaatiokaaviossaan projektipäälliköinä tekoälyagentteja. Että good luck with that. Mutta siis jos tilanteessa, jossa vaikka ollaan rekrytoimassa uutta positiota, todetaankin, että nyt kun meillä on tämä ajatteleva kone, niin emmehän me tarvitse ihmistä tähän enää, niin sellaisessa syntyy siis ihan olennaisia eettisiä kysymyksiä. Jotenkin itse olen nähnyt sen niin, että siis tämä vuosi on nyt ollut semmoinen, että vieläkin täytyy tunnustaa, että koetan prosessoida sitä, että what is going on.

Sanotaan, että tämän vuoden alkuun asti olisin pitänyt hyvinkin hanakasti kiinni siitä ajatuksesta, että meidän pitäisi ajatella tätä enemmän sellaisena tukiälynä, joka on enemmän ihmisen oman kognitiivisen kapasiteetin laajentamista. Se on ollut siis mitä ihmiset ovat tehneet teknologialla kaksi ja puoli tuhatta vuotta. Mutta se, mikä on tämän vuoden aikana muuttunut on se, että esim. kesäkuussa Anthropic julkaisi sellaisen paperin, nyt tietenkin Anthropicin viestintää pitää katsoa pikkaisen sellaisten IPO-markkinointilinssien läpi, mutta niillä on aika hyvä track record kuitenkin käsittääkseni siinä, että ne eivät ainakaan mitään hirveää huuhaata tuonne ole tuutanneet, niin että ne ovat 2024 pystyneet automatisoimaan 4 minuuttia pitkiä taskeja, 2025 olikohan se muutaman tunnin pituisia taskeja, jotain sitä luokkaa, ja sitten tänä vuonna 12 tuntia pitkiä taskeja. Niin kyllähän siinä alkaa olla jo siis jotain ihan muuta.

Kuten äsken sanoit, että ei se enää ole pelkästään stokastinen papukaija. Jos se pysyy koherentisti tekemään 12 tuntia jotain systemaattista päämäärää kohti, ja sitten lopputulos on varmaan usein parempi kuin ihmisen tekemänä. Ja siis vedin pelistartupia aikanaan, niin se prototyyppi, tai siis kun se ei edes ole enää prototyyppi, vaan ihan peli, jonka tein nyt viikonloppuna, niin minulla on ollut unelmana siitä saakka, kun poistuin siitä startupista, että jonain päivänä, kun nämä teknologiat ovat tarpeeksi kehittyneitä, niin teen sellaisen periodisen taulukon oppimispokemonpeliin. Ja olen kokeillut sitä aina uudestaan ja uudestaan.

Kun tuli Lovable viime vuonna, niin olin, että nyt se on tässä. Ei siitä tullut yhtään mitään. Siis Lovablella pystyy tekemään tosi hienoja verkkosivuja ja prototyyppejä. Se on aivan huikea teknologia. Mutta ei sillä kompleksista monisivuista Excel -peliekonomiaa pysty, että forget about it. Ja sitten nyt viikonloppuna se onnistui. Grafiikka ei ole niin hienoa kuin huippugraafikoiden tekemänä valitettavasti. Tekoäly ei vielä ihan pärjää näille maailman huippugraafikoille. Mutta se peliarkkitehtuuri. Kaikki tällaiset indikoinnit, siis se, miten pelaajaa ohjataan sen pelin pelaamisessa, jotka ovat usein tosi vaikeita, vaativat hirveästi käyttäjätestausta.

Niin yksi juttu, jota se tekee esimerkiksi on se, että se ajaa jokaisen revision jälkeen 1200 pelisimulaatiota ja katsoo, että jääkö se peli jumiin jossain. Pelkästään tämä simulointi voi olla pelialalle semmoinen. Silloin 10 vuotta sitten me panimme PlaytestCloudiin niitä. Sitten ihmiskäyttäjät painoivat, että miksei tuosta napista tapahdu mitään. Meillä oli tällainen oikea tapaus, että graafikko oli piirtänyt hienot sellaiset ympyränmuotoiset steampunk-säiliöt siihen pelin alkuruutuun. Sitten kaikki pelitestaajat painoivat niitä säiliöitä, kun ne näyttivät ihan napeilta. Me vasta siinä vaiheessa näimme sen oman juttumme läpi. Niin tällaiset jutut. Tuo 1200 simulaatiota ja se on saman tien korjattu.

Eric: Mutta se, että saat nyt tekoälyn automatisoimaan tämän prosessin, niin pitäisikö meidän mielestäsi vetää siitä jotain eettisiä johtopäätöksiä siitä, miten meidän pitäisi suhtautua näihin tekoälymalleihin?

Arno: Pitääkö siihen suhtautua samoin kuin työkavereihisi siinä pelifirmassa?

Eric: Niin, siihen tiimiin.

Lauri: Nyt kun tuon sanoin ääneen, niin se on ihan totta. Siis tavallaan itse ajattelen sen enemmänkin sillä tavalla, että nyt minulle on suotu tämmöinen toimintakyvykkyys. Versus se, että olen nyt automatisoinut… Niin. Tuo on kyllä kinkkinen kysymys. Nyt minun täytyy tunnustaa, että pitäisi pohtia tätä vähän enemmän. Koska en siis halua myöskään julistaa semmoista, että ruvetaan luddiiteiksi, pannaan hanat kiinni ja suojataan työpaikat tekoälyltä. Koska se ei vain tule tapahtumaan. Siinä ihan pelkästään markkinavoimat tulevat näyttämään kaapin paikan. Koska sitten ne toimijat, jotka eivät lähde sellaiseen leikkiin mukaan, niin sitten niiden kokonaiskapasiteetti vaan kasvaa niin paljon isommaksi. Ja kilpailukyky.

Mutta jotenkin haluaisin ajatella, että me voisimme kehittää sitä teknologiaa sellaisesta näkökulmasta, jossa pyritään siihen, että se lisää ihmisten hyvinvointia ja kyvykkyyttä, eikä siihen, että me keskitymme vaan vaikka tuottavuusloikkiin tai muihin. Tämäkin nimenomainen tapaus on semmoinen, että minulla ei ole minkäänlaista realistista skenaariota, että pystyisin tässä elämäntilanteessa nostamaan rahoituksen peliyhtiölle ja yrittämään uudestaan sillä perinteisellä metodilla. Ei ole tavallaan kyse siitä, että kilpailen sen jonkun peliyhtiön kanssa. Kyse on siitä, että nyt jokin sellainen, joka ei olisi ollut mahdollista kaksi kuukautta sitten, niin on todellista. Ja se on minusta se, että mikä sinulle on mahdollista nyt, mikä ei aikaisemmin ollut todellista, ja mikä sinulle on mahdollista, mikä aikaisemmin ei ollut todellista, niin miten me voimme tätä hyödyntää? Näen, että se mahdollistaa sen teknologian, joka pystyy tuomaan sinulle, minulle ja meille kaikille ihan uudenlaisia kyvykkyyksiä. Sitä ei pitäisi asettaa vielä tavallaan olion asemaan, ellei se ala käyttäytyä ikään kuin se olisi olio. Tämä on ehkä se, mitä ajan takaa tässä.

Eric: Mutta kun mietit näitä eettisiä implikaatioita, niin kuulostaa, että sinulla on enemmän mielessä miten se vaikuttaa muihin ihmisiin, meihin ihmiskuntana, kun tuodaan tekoälyä enemmän käyttöön. Kun sitten taas ehkä vielä semmoinen extreemimpi eettinen kysymys, jota vaikka Anthropic näissä modelcardeissaan on tuonut esiin, on, että onko se eettisesti oikein sitä tekoälymallia kohtaan, että me laitamme sen vastaan johonkin tai käymään läpi jotain kyseenalaista sisältöä ja filtteröimään sitä. No, minulla on useampiakin varauksia tätä väitettä kohtaa. Mutta oletko sinä ajatellut sitä puolta ollenkaan?

Arno: Mitä oikeuksia tekoälymallilla on?

Lauri: On siis myös mahdollista, että huomioiden pelkästään yksien ihmisaivojen kompleksisuuden sen nojalla mitä me nyt tiedämme, jonka lisäksi on se, mitä me emme tiedä, niin kuin gliasolut tai vaikka tämä Penrosen-Hameroffin hypoteesi, että itse asiassa aivojen laskennallinen mallintaminen on mahdotonta sen takia, että siellä jokaisessa neuronissa tapahtuu niin paljon kvanttitason kompleksisuutta, joka on relevanttia kognition synnyn kannalta, niin nämä ovat myöskin mahdollisia teorioita. Jolloin se tarkoittaisi sitä, että me emme koskaan pysty edes millään laskentamenetelmällä simuloimaan varsinaisesti kuitenkaan sitä tietoisuutta. Ja olisin ollut ehkä enemmän itse ennen vanhaan kallellaan tuohon suuntaan.

Nyt, niin kuin sanoin, olen vähän tällainen takinkääntäjä. Että olen ehkä kuitenkin enemmän sitä mieltä, että kyllä me joudumme luultavasti tuohon kysymykseen vastaamaan meidän elinaikanamme, mahdollisesti jo hyvin pian.

Mutta tuo, miten sinä freimasit tuon kysymyksen, on itse asiassa tosi akuutti kysymys. Eli se, mitä me laitamme ne mallit tekemään, miten se vaikuttaa niihin malleihin ja etenkin se, miten niitä koulutetaan. Törmäsin juuri yhteen opiskelijani lähettämään videoon, että oli pistetty eri tekoälymallit toimimaan siellä, oliko se nyt pieni kylä, jossa on 16 agenttia. Näitähän on ollut, että Minecraftia ajetaan LLM:illä. Sitten oli tämmöinen yhteiskuntasimulaatio, jossa testattiin, että mitä tekee Claude, Grok, Gemini ja muut. Claude ei tehnyt yhtään rikosta sen simulaation aikana. Grokin agentit

olivat tappaneet toisensa neljän ensimmäisen päivän jälkeen. Geminikin oli yllättävän tuhoisa kaiken kaikkiaan. Mutta mikä oli pysäyttävintä, oli se, että kun pantiin ne eri agentit vuorovaikuttamaan keskenään, niin sitten se oli ihan täysi kaaos samantien. Myös Sonnetista tuli sitten sen jälkeen semmoinen kunnon kärpästen herra -hahmo siellä.

Jo kauan ennen kuin päästään sellaiseen tilanteeseen, joka on kuitenkin vielä spekulatiivinen tilanne, että on ajatteleva kone, niin on tämä kysymys siitä, että miten ne koneet vaikuttavat meidän yhteiskuntaamme ja maailmaamme. Silloin se eettinen kysymys, että mihin me altistamme sen mallin ei ole pelkästään se, että tuleeko sille paha mieli, koska ei sille tule mikään mieli ennen kuin sillä on mieli. Mutta se, että mille me altistamme sen, voi ohjata sitä, minkälaisia outputteja nämä mallit generoivat. Tai sitten tämä Grok versus Sonnet -simulaatio, että se, millä tavalla koulutetaan niitä järjestelmiä. Siitähän tuli pari kuukautta vai kuukausi sitten semmoinen paperi, jossa osoitettiin, että se mallien arvomaailma heijastaa hyvin voimakkaasti niiden kehittävien organisaatioiden arvomaailmaa. Se nyt on tavallaan aika no brainer, koska ne organisaatiothan päättävät sen rubriikin, jolla niitä jatkokoulutetaan, tai jotkut palkinnot tai miten sitten menevätkään ne järjestelmät siellä. Niin, sano vaan.

Eric: Ehkä yksi kritiikki sille, tuleeko mallille paha mieli. Vaikka ajateltaisiin, että niille voisi jo tulla paha mieli, niin kuitenkin mikä erottaa ne meistä ihmisistä on se, että ne pystytään aina palauttamaan takaisin siihen aiempaan tilaan. Että me emme aiheuta ainakaan mitään pysyvää vahinkoa.

Lauri: Ei traumatisoida tekoälyä.

Eric: Niin, ei traumatisoida. Se ehkä menee takaisin siihen mitä puhuimme siitä, että ne eivät ole tällä hetkellä tavallaan sellaisessa jatkuvassa luupissa, jossa ne koko ajan muuttuisivat. Ja myös siihen, mitä mainitsit muistista, että ne jatkuvasti päivittyisivät, vaan se on aina kelattavissa taaksepäin.

Arno: Eli samaan virtaan voi astua kahdesti, toisin kuin meidän ihmisten kohdalla.

Lauri: Kyllä. Mutta minusta tämä eettinen kysymys on myöskin, että kuinka tehokkaita malleja ylipäätään voidaan julkistaa. On tämä hiljattainen Anthropicin analyysi vaikka siitä, mihin kaikkeen ikävään niitä niiden malleja on pyritty jo käyttämään. Se on paljon akuutimpi eettinen kysymys minun mielestäni, kuin se kysymys, että miten toimitaan ajattelevien koneiden kanssa. Olen itse vähän skeptinen näistä tuomiopäivän skenaarioista ja ihmettelen edelleenkin sitä, mistä Geoffrey Hinton on tempaissut sen 10 %, kun tuntuu, että se on semmoinen, että ainakaan toistaiseksi en ole nähnyt mitään laskelmaa, joka johtaisi 10 %:n ekstinktioon.

Miksi minun kielenkäyttöni on näin outoa tänään? Siis ihmiskunnan tuhotilanteeseen. Tällä hetkellähän se 10 % pyörii joka paikassa sellaisena faktana, koska nobelisti sanoi näin. Mutta paljon akuutimpi eksistentiaalinen riski ihmiskunnalle on tämä, että nämä tehostavat kaikkea mitä ihmiset voivat tehdä. Tulesta lähtien meillä on ollut aina vähän se ongelma, että käytetäänkö sitä hyvään vai pahaan. Kyllähän niitä voi pahaankin

käyttää. Siinä mielessä minusta siitä pitäisi puhua paljon enemmän kuin siitä, että tulevatko terminaattorit. Minä kirjoitin pienen analyysin siitä, oliko se Jacob Coxon, joka aloitti tämän viimeisimmän vyörytyksen. Sitten ensimmäistä kertaa itse pysähdyin miettimään, että oikeastaan jos me tekisimme tämmöisen artificial super intelligencen, tällaisen supertekoälyn, niin miksi se ylipäätään haluaisi tuhota meidät? Mistä se tulee? Siis ymmärrän tämän Boströmin tehokkuusmaksimointiajatuksen. Että joo, se on mahdollista tietysti, että joku järjestelmä ei välitä pätkääkään ja katsoo, että tehdään noista klemmareita, kun niissä on paljon hyviä alkuaineita. Could happen.

Mutta jotenkin lähtökohtaisesti, jos nyt vaikka heitän hypoteettisesti, että ei tulekaan supertekoälyä, vaan tulee super-Arno. Ja yhtäkkiä eräänä aamuna heräät niin, että olet kaikkivoipainen, voit tehdä mitä vaan. Niin onko ensimmäinen asia, jonka teet se, että tuhoat kaikki mangustit maan päältä? Tai jonkun vapaavalintaisen elämänmuodon? Vai onko se enemmänkin sellaista siistiä, että nyt voin vähän niin kuin Alan Mooren Watchmen-sarjakuvassa mennä tuonne Marsiin meditoimaan.

Arno: Minä olisin valistunut itsevaltias.

Lauri: Niin niin.

Arno: Katsojillekin tiedoksi.

Eric: Jos jatkan tätä Arno-esimerkkiä, niin ehkä yksi argumentti sille, miksi tekoälyistä voi tulla pahoja on se, että ne on optimoitu toteuttamaan jotain tiettyä taskia, ja sitten ne tekevät kaikkensa. Ehkä tiedän, että Arno on ollut aina kunnianhimoinen ja akateemikko. Mutta sanotaan, että jos vaikka Arno olisi ollut vähän vähemmän lahjakas ja olisi kuitenkin aina ajatellut, että hänestä täytyy tulla proffa. Niin ehkä silloin jatko-opintovaiheessa, jos hän ajattelee vaan sitä.

Arno: Tuhoan kaikki muut alan ihmiset, jotta ketään muuta ei ole.

Eric: Niin. Tai ehkä pienempi askel, että täytyy saada tämä paperi läpi ja saada nämä hyvät tulokset, niin ehkä on valmis tekemään jotain vilppiä. Eli jotenkin tuo ajatus skaalattuna isompaan mittakaavaan.

Arno: Keinoja kaihtamatta.

Eric: Keinoja kaihtamatta. On vaan se yksi tietty päämäärä. Se on juuri tämä paperclip-esimerkki loppukädessä.

Lauri: Joo ja siis tarkennan vielä. En missään tapauksessa sano, että se on absoluuttisesti nolla, että never gonna happen. Kyllähän sellainen skenaario on todennäköinen. Mutta se, että miksi jaellaan jotain tiettyjä numeroita ilman mitään varsinaista argumenttia sen numeron taustalla? Tai se, että kuinka todennäköistä se on, niin minusta sitä on aika vaikea sanoa. Siinä mielessä vähän vierastan sitä Kurzweilin singulariteettipopularisointia. Mutta siinä mielessähän se on singulariteetti, että se alkuperäinen ajatushan on se, että mustan aukon tapahtumahorisontin toiselta

puolelta ei saada mitään informaatiota. Ja ASI:n toiselta puolelta ei saada mitään informaatiota, koska kyse on niin kompleksista ja niin tavallaan meille käsittämättömästä järjestelmästä, että me emme oikeastaan voi sanoa mitään siitä, mitä sitten tapahtuu, jos tällainen superäly syntyisikin yhtäkkiä. Mutta ennen kuin päästään niin pitkälle, niin pitäisi ensin ehkä päästä siihen... No jaa, itse asiassa oliko se peräti Jensen Huang, joka taisi jo vuosi tai kaksi sitten sanoa, että oikeastaan meillä on jo superälyjä. Ne ovat vain hyvin kapeissa vertikaaleissa. Monissa asioissahan nämä vimpaimet ovat jo osittain käsittämättömiä siinä mitä ne pystyvät tekemään, mikä ylittää ihmisen kyvykkyyden.

Sitten jos taas skaalaa riittävän moneen vertikaaliin, ja alkaa tulla niiden keskinäistä vuorovaikutusta, niin se voi olla se superäly. ASI:n ei välttämättä tarvitse koskaan mennä AGI:n kautta. Eli tällaisen yleisen ihmisen kaltaisen tai ajattelevan koneen kautta. Sitten tietenkin Boströmin perusargumentti on se, että emmehän me tarvitse mitään tietoista supertietokonetta. Sen tarvitsee olla vain riittävän tehokas ja riittävän päämäärätietoinen tai päämäärähakuinen. Ja sitten meistä tulee klemmareita. Mutta kyllä sanoisin, että meillä on niin paljon akuutimpia ja todellisempia riskejä juuri siinä, että miten näitä ohjataan, otetaan käyttöön ja koulutetaan näitä malleja, miten siinä koulutuksessa pystytään ohjaamaan sitä alignmentia, kohdennusta. Mikä on alignment suomeksi?

Eric: Linjaus. Vai?

Arno: Linjausta.

Lauri: Ehkä meidän pitäisi puhua niistä enemmän kuin tällaisista tuomiopäivän skenaarioista, jotka ovat kuitenkin enemmän ehkä scifiä tällä hetkellä.

Eric: Niin ei varmasti akuuteimpia. Minua siinä keskustelussa vähän häiritsee, että yleensä nämä tuomiopäiväskenaariot ohitetaan vaan kyseenalaistamalla vaikka niiden esittäjien motiivit. Että tässä ovat vaan jotkut markkinointikikat mielessä. Mutta kyllä minä tiedän paljon ihmisiä, joita kunnioitan tosi paljon älyllisesti, jotka ovat oikeasti vilpittömästi huolissaan näistä. Että mielestäni meidän pitäisi pureutua niihin argumentteihin. Ehkä meillä ei nyt tässä ole aikaa mennä kaikkiin. Mutta minäkään en pidä kauhean todennäköisenä. Mutta kyllä voin nähdä niitä polkuja, miten tavallaan teoriassa se voisi johtaa siihen, vaikka ne eivät näytäkään minusta todennäköisiltä tällä hetkellä.

Arno: Jos kaikki nämä riskit vain sivuutetaan olankohautuksella ja leivotaan scifiksi ja markkinoinniksi, niin ehkä me sitten emme kysy niitä relevantteja kysymyksiä, jotka meidän pitää kysyä, jotta siihen pisteeseen ei jouduta.

Lauri: Jep. Mutta siis tavallaan onhan tässä sillä lailla huvittava tilanne. Että siis Dario Amodei, jota olen kuitenkin pitänyt ja pidän edelleenkin noista ainakin kaupallisen kehittämisen puolen ihmisistä niin kuin, sanotaan, siellä ei oikein ketään kauhean eettistä ole, mutta hän on vähiten epäeettinen. Olen pitänyt häntä tavallaan sellaisena, että hän on one of the good guys. Daniela ja Dario Amodei ovat molemmat minusta

arvomaailmaltaan tosi hyviä. Paljon ajatuksia lausuntoon, mutta sitten esimerkiksi tämä jarruttamis- manifestijuttu, joka tuli pari viikkoa sitten, niin onhan se vähän optiikaltaan erikoista, että ensin tulee tämmöinen ulostulo, ja sitten puolitoista viikkoa myöhemmin by the way, tässä on Opus 5.5 vai onko se 5.5 vai 5.6, mikä se nyt oli. Tänään vai eilen. Keväällä Mythos oli se, joka hakkeroi kaikkea ja tuhosi kaikkea. Sitten tuli, että tämä Fable on turvallinen Mythos. Ja nyt tässä on tämä Opus 5.5, joka on oikeastaan parempi kuin Fable, mutta maksaa vähemmän ja on nopeampi. Mutta on tosi turvallinen kyllä, että kannattaa käyttää. Niin siinä herää se, että ei se ihan niin kuin... Mutta toisaalta ymmärrän sen problematiikan myöskin.

Ja siis muistan, kun Mikko Hyppönen piti meidän Prodeko-seminaarissamme muutama vuosi sitten sellaisen esityksen kyberrikollisuudesta ja siitä, että nykyään on tällaisia kyberrikollisyksisarvisia. Siellä on toimitusjohtaja ja operatiivinen johtaja, ja niitä johdetaan kuin startupia. Ne tekevät miljardiliikevaihtoa siis kiristämällä, sulkemalla pankkien sivustoja ja kiristämällä niitä. Siellä on pipelinessä aina CRM:ssä projektit ja mikä on seuraava mahdollinen targetti. Niin on se nyt sitten niinkin, että jos minun pitäisi jonnekin lotota tällä hetkellä, niin kyllä toivoisin, että Dario Amodei kehittää supertekoälyn mieluummin kuin joku venäläinen kyberrikollisuusyksisarvisorganisaatio. Tai sanotaan, että edes Elon Musk. Kuitenkin se vähiten pahakin voi olla aika paljon parempi, kuin se, jos ne hyvikset alkavat jarrutella liikaa. Mutta silti siinä on jotain outoa minun mielestäni. Luulisi, että ne tajuavat, että ei se jarruttaminen ole mahdollista tässä tilanteessa. Miksi sitä sitten edes nostetaan keskusteluun?

Eric: Niin. Ja kun sanoin siitä motiivien kyseenalaistamisesta, niin ehkä nämä firmojen pääjehut ovat vähän niinkuin asia erikseen. Mutta tämä laajempi keskustelu siitä AI:n turvallisuudesta, tekoälyn turvallisuudesta, niin kyllähän se nousee paljon sellaisilta yksittäisiltä tutkijoilta, jotka ovat puhuneet siitä paljon pidempään kuin Dario Amodei on ollut huolissaan tästä. Mielestäni ne ovat niitä argumentteja, joihin meidän pitäisi puuttua myös, jos me haluamme kyseenalaistaa tätä.

Arno: Siellä on joukossa sellaisia ihmisiä, jotka oikeasti tietävät miten ne mallit toimivat. Kaikki tätä aihetta kommentoivat eivät välttämättä ole ihan perillä, että mitä pellin alla tapahtuu.

Eric: Ja vaikka Jacob Coxon tai Katarina Slama, jotka ovat molemmat olleet, olettaisin että aika hyväpalkkaisissa töissä, ja päättäneet jättää sen työpaikan ja keskittyä tekoälyn turvallisuuden tutkimiseen niin, että ei ole mitenkään suoraviivaista nähdä siihen ainakaan mitään taloudellisia insentiivejä. Että mielestäni niitä ei voida kuitata vaan sillä.

Lauri: Ei. Ja siis joskus ehkä 15 vuotta sitten juttelin tuon Jaan Tallinnin kanssa noista jutuista. Hän kyllä minusta ihan aidosti oli jo silloin tosi huolissaan siitä, että tässä tulee käymään köpelösti, jos ei jotain keksitä. En missään tapauksessa usko, että se on vaan puhdasta markkinointipuhetta. Mutta jotain siinä on siinä tavassa, miten näiden asioiden kanssa ollaan tultu. Ja jotenkin siis sanotaan näin, että mielestäni näistä pitää puhua. Tavallaan juuri tästä alignmentista ja extinction-riskistä on järkevää käydä sivistynyttä ja hyvää keskustelua. Toivoisin ehkä enemmän sitä, ettei pudoteta sellaisia

Stetson-Harrison-lukuja, vaan sitten jos on jotain projektioita tai skenaarioita, niin ne olisivat jotenkin paremmin perusteltuja. Ja sitten toinen on ehkä se, että sen jarrutuspuheen kanssa oltaisiin vähän avoimempia ja realistisempia siitä, että mitkä ne todelliset vaihtoehdot ovat. Jotenkin mielestäni on aika pelottava tilanne, jos Donald Trump on selväpäisin kannanottaja tässä tilanteessa. Silloin ollaan kyllä…

Arno: Tuhon tiellä.

Eric: Ehkä Euroopankin kannalta, jos me tässä tilanteessa alamme jarrutella, kun kaikki nämä johtavat mallit ovat yhdysvaltalaisia ja kiinalaisia, niin meille ei ehkä käy kauhean hyvin.

Lauri: Kun me Ladalla tulemme Lamborghinien perässä, niin en tiedä onko sillä väliä onko jalka kaasulla vai jarrulla.

Arno: Puhutaan hei elämänfilosofiasta. Kun tekoäly pystyy hoitamaan ihmisen ajattelusta ja työnteosta yhä isomman osan ja paljon nopeammin kuin aikaisemmin, niin mitä siitä ihmisyydestä jää jäljelle? Mistä elämän merkityksellisyys kumpuaa? Sinäkin kirjoitat ja teet musiikkia, niin jos sen outputin pystyy generoimaan tekoälyllä, musiikin pystyy generoimaan tekoälyllä, sinun kirjasi pystyisi tulevaisuudessa generoimaan tekoälyllä, niin katoaako jotenkin meidän ihmisten elämästä se merkityksellisyys?

Lauri: Se, että sen pystyy tekemään niin, ei velvoita meitä siihen, että se pitää tehdä. Nostit tämän kirjoittamisen ja musiikin. Minähän olen sillä lailla vähän kummajainen, kun olen tämmöinen tekoälyn suurkäyttäjä. Siis en tiedä kuinka monta sataa promptia kirjoitan päivässä. Mutta en minä tee musiikkia tai tekstiä tekoälyllä. Minusta siinä ei ole mitään järkeä.

Tekstin osalta, no, siitä voidaan käydä keskustelua, onko tekoälyn laatu edelleenkään hyvän ihmiskirjoittajan tasoa. Rohkenisin pitää itseäni ihan kohtalaisena ihmiskirjoittajana ainakin sikäli, että minulta on kuitenkin noita kirjoja useampi tullut. Aika moni on ollut jonkin tason bestsellereitäkin.

Ja musiikin kohdalla olen mennyt täysin vastavirtaan. Minulla vähän lähti lapasesta, nyt on uusi levy juuri valmistumassa, ja edellisen levyn kohdalla innostuin soittamaan enemmän. Aikaisemmin olen tehnyt pääasiassa konemusaa, niin olen soittanut enemmän juuri rumpuja ja bassoa ja tuollaista tavallaan kitaraa. Mutta nyt minulla on siis 47 instrumenttia tulossa uudelle levylle.

Mutta sen merkityksellisyyden kannalta tämä on valtavan iso uhka ja mahdollisuus samaan aikaan. Uhka on siinä, että me olemme luoneet tällaisen kulttuurin, jossa työ itsessään on ripustettu tällaiseen ihmisen arjenhallinnan ytimeen. Se riski on siis se, että jos nyt yhtäkkiä tietotyöstä pystytään tekemään vaikka 50% tai 75% tekoälyllä. En usko tähän tekoäly vie meidän työt -hypoteesiin, koska tällä hetkellä vielä tekoäly ei vie kenenkään työtä, vaan se on ihminen, joka käyttää tekoälyä. Mutta de facto siis

ihminen, vaikka koodari, joka osaa käyttää taitavasti Claude Codea tai Codexia tai muuta, niin pystyy tekemään 10 hengen työt helposti.

Silloin se tarkoittaa, että työmarkkinoilla tulee 9 työpaikkaa, joita ei enää ole. Eli siinä mielessä tekoäly vei ihmisen työt, vaikka se on vähän tällaista retoriikkaa. Mutta se, mikä on se uhka siinä on se, että oikeastaan aika harva ihminen on lähestynyt työtä ylipäätään niin, että se olisi tällainen asia, jota haluaa tehdä. Meidät on opetettu siihen, että hankit tutkinnon ja töitä ja sitten saat laskut maksettua. Nyt ensimmäinen kysymys on se, että jos ihmisillä ei ole töitä, niin miten saadaan laskut maksettua. Periaatteessa meidän pitäisi vain järjestää markkinarakennetta uudelleen, että meillä on riittävät subventiomekanismit ja kaikki nämä perustulot ja muut, joista kaikki tekoälyn kehittäjät puhuvat. Meidän pitäisi saada ne siis seuraavaan hallitusohjelmaan. Ja siis seuraavaan, koska siitä seuraava on jo luultavasti liian myöhäistä. No, good luck with at. Mutta katsotaan. Aina saa toivoa.

Mutta sitten se isompi kysymys on se, että voisimmeko me alkaa sitten seuraavia sukupolvia kouluttaa siihen ajatukseen, että koska sellaista tavallaan työn pakkoa ei välttämättä enää synny samalla tavalla, niin voisimmeko me ohjata siihen, että me olemme kaikki yksilöitä? Jos minulle joku klarinetin epävireinen tuuttaaminen tuottaa suunnattomasti intohimoa ja mielihyvää, niin se on sitten se minun juttuni. Ja mikä on sinun juttusi ja mikä on sinun juttusi? Se on sitten taas valtava mahdollisuus sellaisessa työelämän humanisoinnissa. Ja siinä voitaisiin tuoda niitä ihmisten yksilöllisiä kiinnostuksen kohteita.

Se mikä on semmoinen minun vähän henkilökohtainen projektini tässä ollut 20 vuotta, mistä olen pari kirjaakin kirjoittanut, että miten me voisimme auttaa ihmisiä löytämään tällaista kutsumustyötä tai sellaista työtä, jossa on oikeasti mahdollisimman autenttinen itselleen? Ja että se arki koostuisi askareista, jotka ovat oikeasti itselle mielekkäitä. Mutta tällä hetkellähän siis lentoradat näyttävät olevan siinä asennossa, että meidän pitäisi saada paljon enemmän tuohon keskusteluun virtaa. Koska riskinä on se, että ihmisiä vain jää massoittain työttömiksi, ja se aika menee doom scrollaamiseen, somettamiseen ja aivottomiin hyperkasuaaleihin mobiilipeleihin ja johonkin vielä pahempaan netti- tai uhkapelibrainrottiin. Ja siis brainrottikaan ei ole pahinta tietenkään, mitä voi kuvitella.

Eric: Jos tämä yhteiskunta ei mukaudu riittävän nopeasti tähän muutokseen, että tekoäly tai ihmiset, jotka käyttävät tekoälyä, tulevat viemään ison osan töistä, niin mikä olisi sellainen taito tai jotenkin elämänasenne mitä sinä suosit? Että miten minun pitäisi varautua tähän muutokseen, jotta pärjäisin siinä muuttuvassa maailmassa mahdollisimman hyvin?

Lauri: No siis maailma tarvitsee filosofiaa enemmän kuin koskaan aikaisemmin. Minusta on upeaa, että nyt meillä on täällä Aallossa tämä soveltavan filosofian ohjelma käynnistynyt. Minusta on mahtavaa, että insinööriyliopisto yhtäkkiä haluaa satsata siihen, että opetetaan filosofiaa kaikille opiskelijoille, ja on paljon tarjontaa nyt. Ja kyllä se palautuu siihen Sokrateen apologian huomioon, että elämä ilman itsetutkiskelua ei ole elämisen arvoinen. Se on vähän rajusti sanottu. Mutta tavallaan minusta meidän

kyllä pitäisi koko yhteiskuntaa ohjata enemmän sellaiseen itsetutkiskeluun ja niiden omien kiinnostuksen kohteiden ja sellaisen palveluasenteen lisäämiseen, jotka ovat molemmat minusta välttämättömiä.

Jos ihminen ei tunnista sitä, missä on itse kirkkaimmillaan, ja vaikka haluaakin tehdä hyvää maailmassa, niin siihen palaa loppuun tosi nopeasti. Ja sitten taas toisinpäin, että jos tekee vain itsekeskeisesti sitä minäminä-juttua, niin sitten siitä syntyy taas sosiaalisesti aika vinksahtanut tilanne. Mutta se, että jos pystyy kysymään, että mitä haluan tehdä ja mitä voin tehdä sellaisella tavalla, että se tuottaa jotain hyvää maailmaan, niin nämä kaksi kysymystä jos me osaisimme opettaa ihmisille riittävän hyvin, niin mielestäni se on sellaisen sokraattisen filosofian uusi esiinmarssi. Olen toiveikas, että se on ihan mahdollinen skenaario myöskin. Näen siihenkin viittaavia merkkejä tuolla kuitenkin paljon. Että sitä kohti vaan pitää yrittää omalla aktiviteetilla puskea niin kuin vain suinkin pystytään.

Arno: Kuulostaa hyvältä.

Eric: Suuret kiitokset Lauri. Vastasit kaikkiin meidän kysymyksiimme, vaikka me emme edes päässeet kysymään niitä sinulta. Oli erittäin mielenkiintoinen keskustelu. Kiitos.