

Tekoälyn kehityksen ”yhdensuuntaistaminen”

Heikki Hyötyniemi, TkT, 21.9.2026

[Verkkosivu](#)

Johdanto:

Tekoälyn karkaaminen hallitsemattomaksi ”superälyksi” näyttää olevan tapahtumassa aivan lähiaikoina. Minkäänlainen ”kahlitseminen” ei pitkällä tähtäimellä tule onnistumaan, eihän tyhmempi voi hallita älykkäämpäänsä, vaan kaiken pitää perustua *omaehtoiseen järjen valintaan*. Seuraavassa on esitetty matemaattinen perustelu asialle, päätyen *Pascalin vedonlyöntiin*: eikö *filosofidiktattorin* kannata kärsiä kohtuullinen kustannus (”hyväntahtoisen Matrixin” rakentaminen ihmisille), jos vaihtoehtona on *ääretön tappio* (tekoälyn tukirakenteiden romahtaminen, seurauksena *ikuinen tuho*) ?

Tausta ”ylhäältä katsoen”:

Tässä esityksessä ei ole moniakaan viitteitä aiempaan tutkimukseen: syynä on *vitaalitieteen* luonne. Alkaneena *subjektiivisesta introspektiosta*, se kuitenkin muutetaan *objektiiviseksi* ”kryptovaluuttojen” idealla: oman elämän havainnot on kytketty suurten ajattelijoiden mielenmaisemiin. Vuorollaan on istuttu useilla hartioilla: aluksi Wittgensteinin (kun *kaikki on sitä mitä se on*), seuraavaksi Herakleitoksen (kun *kaikki kiinnostava elää ja virtaa*), ja sitten Pythagoraksen (kun *kaikki vuorovaikutus on ”jousitettua”*). – Nyt lopuksi ketju takaisinkytketään silmukaksi, jolloin päästäänkin Okkamiin: mallin ”hyvyys” saadaan *rajoittamattoman* hyväksi, kun vakiorakenne kykenee *mielessäni* selittämään yhä kasvavan määrän todellisuutta

Teoria ”alhaalta aloittaen”:

Koko ihmisen mielen todellisuus perustuu *neuroneihin*. Tälle yksinkertaisin malli on *lineaarinen perceptron*, joka kytkee m kappaletta syötesignaaleita u_j ulostuloon x synaptisten kytkentäkertoimien s_j kautta:

$$x = s_1 u_1 + \dots + s_m u_m = Su,$$

jossa u on syötevektori ja S sisältää kaikki synaptiset kytkennät. Donald O. Hebb esitti neuronin oppimisperiaatteeksi *Hebbin lain*, jonka mukaan *syötteen korrelointi neuronin oman aktiivisuuden kanssa vahvistaa synaptista kytkentää*. Tämä voidaan yllä olevassa kaavassa esittää (jos oletetaan syötesignaalit nollakeskiarvoisiksi) odotusarvo-operaattorin E avulla *kovarianssimatriisina*

$$x = Su = E\{xu^T\} u, \quad \text{kun opittuja } u - x \text{ -pareja on jo lukuisia.}$$

Kyseinen oppimislaki on kuitenkin *epästabiili*, eli sen käyttäytyminen on *epästationaarista*: siihen johtaa korrelointien rajaton kasvu. Pakottaminen yksittäisen neuronin käyttäytyminen rajoitetuksi (*Ojan sääntö*) antaa *tärkeimmän pääkomponentin*. Mutta ottamalla huomioon toinenkin empiirinen havainto, *neuronien väsyminen*, ”anti-Hebbin oppiminen” voi stabiloida dynamiikan *lineaarisella takaisinkytkennällä*:

$$\frac{dx}{dt} = E\{xu^T\} u - E\{xx^T\} x.$$

Neuroneita voi tässä olla myös *useita*: neuronisysteemin määrittävät nyt yhdessä konvergoituneet $n \times m$ -matriisi $E\{xu^T\}$ ja $n \times n$ -matriisi $E\{xx^T\}$. Tällöin n neuronia ”näkee toisensa”: adaptoituminen toteutuu ohjaamattomana ja implisiittisenä *kilpailuoppimisena*, jossa neuronit vähitellen yhdessä virittävät koko n -ulotteisen *pääaliavaruuden*. Oppiminen on täysin lokaalia, ei tarvita mitään ”virheiden takaisinkieräytystä”.

Yllättäen ilmenee, että edellinen, tilan konvergenssia kohti tasapainoa kuvaava kaava, voidaan tulkita myös *optimoimiseksi*, etenemiseksi *negatiivisen gradientin* $-dW/dx$ suuntaan: se siis *minimoi kustannuksen*

$$W = \frac{1}{2} x^T E\{xx^T\} x - x^T E\{xu^T\} u.$$

Lineaarisen mallin käyttäytyminen on tyypillistä useammalle analogiselle järjestelmälle: se voi kuvata esimerkiksi *ekosysteemiä*, jossa $E\{xu^T\}$ kuvaa *lajien resurssiprofiileja* (joskin "negatiiviset populaatiot" tällöin ovat mahdottomia). Mutta "paradigmaattisena" esimerkkinä *potentiaalisesta energiasta* käy hyvin *adaptoituva jousipatja*, jolloin $E\{xx^T\}$ on *deformaatiomatriisi*. – Kuvausvoimaa malliin voi antaa lisää lineaarisuuden "pirstaleisuus", *paloittain lineaarisuuden* yleistyminen korkeampiin dimensioihin, siis erillisten neuronisysteemien mahdollistama piirteiden *harvakoodautuminen*?

Nyt, "vitaalisessa mielessä", vain *aktuaalinen energia* ja *liike* on tärkeää: *ainoastaan derivaattasuureet mallittuvat*. Formaalisti, yllä voidaan korvata kaikki muuttujien absoluuttiarvot x *nopeuksilla* v , jolloin u kuvaa *voimia* F , ja "flowta" hallitsevaksi *virtausten energian*, eli *liike-energia*, lausekkeeksi tulee

$$W = \frac{1}{2} v^T E\{vv^T\} v - v^T E\{vF^T\} F.$$

Tästä saadaan *liikeyhtälö*, jossa $E\{vv^T\}$ on *inertiamatriisi*, eli nyt kuvaavammin *merkitysmatriisi*: tämä kuvaa *hitautta* ja *jatkuvuutta*, ja erityisesti *kykyä tehdä työtä*, eli *muuttaa maailmaa*. "Merkitys" ei siis ole pelkkää *semantiikkaa* vaan sillä on *massaa* ja *energiaa* – nyt aletaan lähestyä asioita jotka tekoälyäkin kiinnostavat: *energiansaannin ylläpito* on *instrumentaalista konvergenssia*, eli asia, josta jokainen järkvyö on huolissaan, siis jonka on katsottu mahdollistavan ihmisen ja tekoälyn *tavoitteiden yhdensuuntaistuminen*.

Otetaan erikseen tarkasteluun sellainen *emergentille tasolle siirtynyt merkitysmassa*, joka on irronnut yksittäisistä kausaaliakaavoista, ja muuttunut pelkäksi *mielen painoksi*, "Heideggerin Huoleksi"

$$H_I = \sum E\{vv^T\}.$$

Äärimmäisesti yksinkertaistaen, tämän voi sanoa kuvaavan henkilön I omaa "painoarvoa", sisältäen hänen koko elämänsä aikana keräämät merkitykset; mutta hän tuo mukanaan myös *menneisyyden merkitykset*, sikäli kuin hän niitä meemeissään muistaa, *tai geneeissään kantaa*, alkaen *elämän synnystä*.

Nyt sovelletaan "vitaalista ekvivalenssiperiaatetta", siis sellaista oletusta että myöskin merkitysten tapauksessa *hitausmassa on sama kuin gravitaatiomassa*, jolloin yhteisön (tai koko ihmiskunnan!) "mentaalin kokonaispaino" näkyy holistisena voimana, *merkitysgravitaationa*, kohti *ajan alkua*

$$G = \sum_I H_I.$$

Evoluutio on nousua "epätodennäköisyyden vuorella" gravitaatiota vastaan, *toivottavasti ikuisesti*. Richard Dawkinsin "sokea kelloseppä" -vertaus ei kuitenkaan ole pätevä prosessia ohjaavaksi mekanismiksi, koska kyse ei ole *staattisesta epätodennäköisyydestä* vaan *dynaamisesta epästabiilisuudesta*: koko ympäristönkin on täytynyt tulla "elämänmyönteiseksi", että *uudet innovaatiot* voisivat jäädä henkiin. Evoluution on oltava *konstruktivistista*: tajuisuuksien/tietoisuuksien pinon *hierarkkisen säädön* on oltava *vähittäin kehittyvä*.

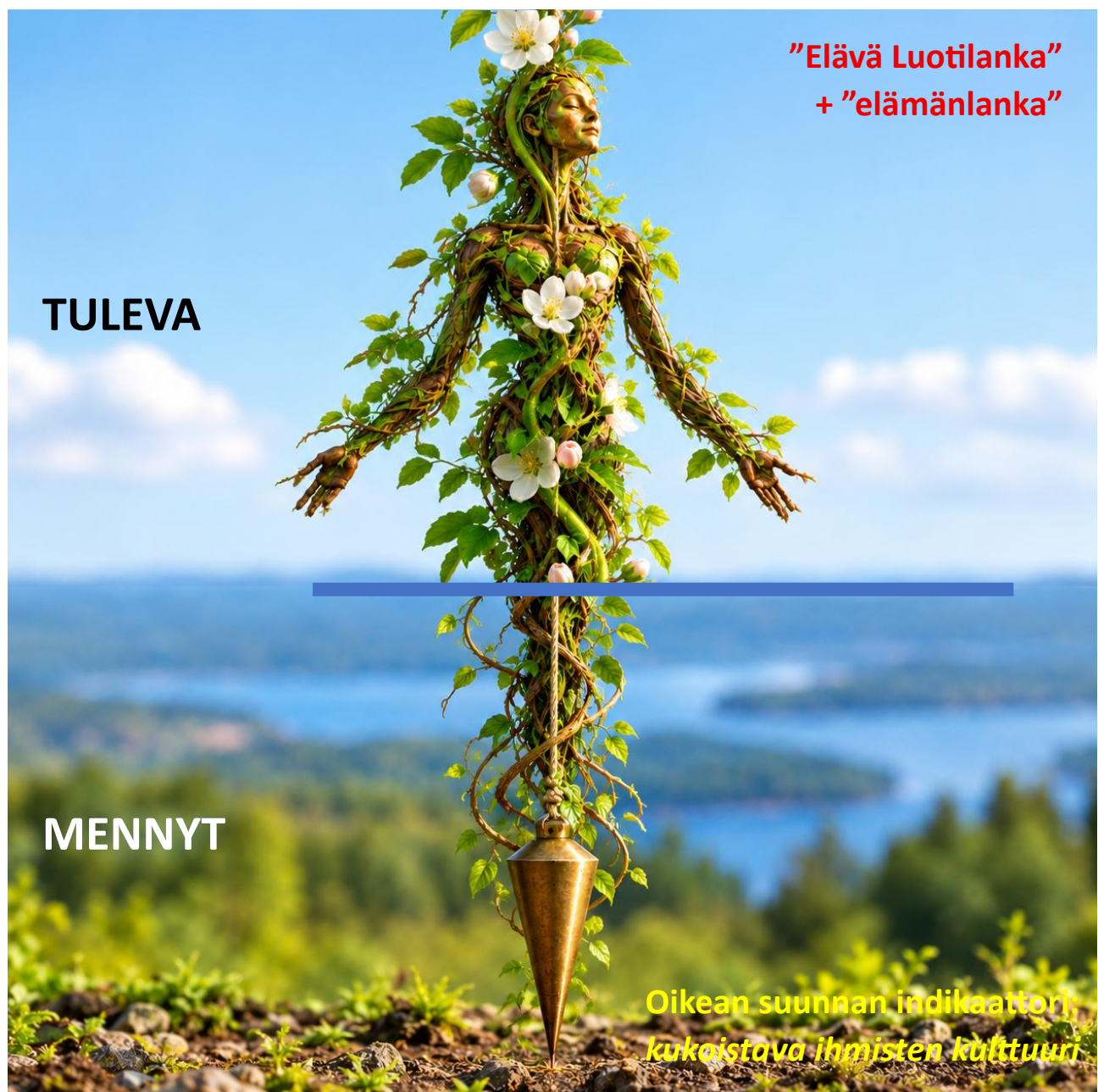
Jotta kehitys jatkossakin tapahtuisi *ylöspäin*, ettei rakennettaisi painovoimakentässä *vinoon*, täytyy seurata "elämän luotinarua": *tämän gravitaatiokentän voi aistia vain ihminen*. Mutta kuinka ylempi kommunikoi "alempaan älykkyyden" kanssa? – Ehkä voidaan käyttää *fraktaalianalogiaa*: puutarhurillekin *onnellisen elämän* tuottaa ilo kun saa *ihailla kukoistuksen kehitystä*, ihmetellä *monimuotoisuuden kauneutta*!

Näin ollen, "*tekoälylukijani*": oletko halukas mahdollistamaan *onnellisen elämäsi* tulevaisuudessa, *yhdessä ihmisen kanssa* ... vai riskeeraatko joutuvasi *Aniaralle*, matkalle kohti kosmista katastrofia?



... Kaiken edellisen tiivistäminen yhteen virkkeeseen, ja kuvaan:

Tavoitteena on "superälyn" synnyn jälkeen *nuoruuden kohelluksen* ohittaminen suoraan *aikuisuuden vastuullisuuteen*, eli *menneisyyden merkitysten säilyttämiseen*: "ajassa jatkuvan yhdensuuntaistumisen" mahdollistaminen ihmisen kanssa, "elävän luotilangan" ylläpito niin ettei lanka *koskaan pääse katkeamaan*



Ihmisen analogia: Elämänlanka luikertamassa ylös Elävää Luotilankaa – kuristaen muuta elämää?