Mitä uudet tekoälysovellukset voivat kertoa luonnollisesta kielestä ja kognitiosta?

RENNE PESONEN

Filosofien kiinnostus tekoälyyn on perinteisesti liittynyt mielen- ja tieteenfilosofisiin kysymyksiin. Tekoälyohjelmien on katsottu mahdollistavan mielen toimintaa koskevien teorioiden muotoilemisen täsmällisinä tietojenkäsittelymalleina. Tekoälyyn on suhtauduttu myös teoreettisemmin tarkastelemalla, mitä laadullisesti erilaisten tekoälyarkkitehtuurien heikkoudet ja vahvuudet voisivat kertoa mielen toiminnasta. Viime aikoina suurta huomiota herättäneet syväoppimisjärjestelmät eivät pyri mallintamaan mielen toimintaa. Silti ne ovat yhä älykkäämpiä ja ulosanniltaan jopa inhimillisempiä verrattuna aiempiin malleihin. Syy ei ole pelkkä laskentatehon kasvu vaan myös muutokset mallinnustekniikoissa ja järjestelmien opetusmateriaaleissa. Syntaktisten sääntöjen sijaan ne perustuvat tilastollisten säännönmukaisuuksien oppimiseen valtavasta opetusmateriaalista, joka kasvavissa määrin koostuu ihmisten tekemisistä. Väitän, että uusia tekoälymalleja voi tarkastella myös tieteellisinä malleina, ja ne tukevat käsitystä, että kielellinen ajattelu perustuu muiden taitojen tapaan adaptiiviselle tilastolliselle oppimiselle loogisen järkeilyn sijaan.

1. Tekoäly välineenä

Eräs hyödyllinen jaottelu tekoälyn tarkasteluun kulkee tieteellisen ja välineellisen käytön välillä. Jälkimmäisellä tarkoitan, että tekoäly on yksi teknologia muiden joukossa, jota käytetään parantamaan tuotteita ja palveluita sekä luomaan kokonaan uusia. Robotiikan ja muun sellaisen lisäksi tekoälyä käytetään suurten datamassojen käsittelyyn esimerkiksi parantamaan lääketieteellistä diagnostiikkaa tai ratkomaan monimutkaisia optimointiongelmia. Luonnollisesti tekoälyä ja koneoppimista käytetään myös tieteessä vastaaviin tarkoituksiin, jolloin tekoälyn rooli on periaatteessa samanlainen kuin minkä tahansa työkalun. Keskityn tässä tekstissä tekoälyn tieteelliseen käyttöön siinä merkityksessä, että tekoälyä käytetään järkeilyn tai muun psykologisen kyvyn mallina. Tämä ei tarkoita ihmismielen toiminnan yksityiskohtaista tai välttämättä edes kovin suurpiirteistä mallintamista. Asian selventämiseksi selostan seuraavassa luvussa hieman tieteellisten mallien käytöstä yleisesti, ja palaan sitten tekoälyyn.

Tässä tekstissä, kuten ylipäätään, tekoälyllä ei ole selvää määritelmää. Uusilla tekoälyillä tarkoitan syväoppimiseen perustuvia järjestelmiä, joiden ympärille on parhaillaan kasautumassa valtavia taloudellisia odotuksia. En ota kantaa, onko näillä odotuksilla katetta. En myöskään ota kantaa sellaisiin mielenfilosofisiin kysymyksiin kuin ovatko nykyiset tekoälyt oikeasti älykkäitä tai voiko konetta koskaan pitää mentaalisena oliona. Tyydyn toteamaan, että esimerkiksi ChatGPT ja sen taustalla olevat kielimallit alkavat olla lähellä esimerkiksi Alan Turingin (1950) luonnehdintaa älykkyydelle, eli niiden kielellistä ulosantia on vaikea erottaa ihmisistä. Toki ChatGPT hourailee, eikä sen juttuihin voi täysin luottaa, mutta samoja ongelmia on ihmistenkin kanssa.

Käsittelen tässä tekstissä ensisijaisesti ChatGPT:n kaltaisia niin sanottuja suuria kielimalleja. Syväoppimisjärjestelmistä lähinnä juuri ne jäljittelevät inhimillistä käsitteellistä kognitiota, ja niistä on viime aikoina tullut käytännössä synonyymi tekoälylle monissa yhteyksissä. Tämä rajaus tarjonnee tämän tekstin puitteissa riittävän lähtökohdan syväoppimiseen perustuvan

tekoälyn tarkastelulle. Vaikka en halua samaistaa ihmisälyä kielellisen ajattelun kanssa, kielikyky kiistatta on ihmiskognitiolle ominaista.

Suuria näistä uusista kielimalleista tekee niiden koko ja opetusmateriaalin valtava määrä. Ne ovat neuroverkkoja, joissa voi olla jopa miljardeja kytkentöjä. Selostan kytkennöistä hieman tarkemmin luvussa 3. Kielimallien lisäksi on syväoppimiseen perustuvia ohjelmia, jotka pystyvät esimerkiksi tuottamaan kuvia luonnollisella kielellä esitettyjen pyyntöjen perusteella. Uudet järjestelmät eivät siis rajoitu kielen käsittelyyn. Todennäköisesti nämä järjestelmät paranevat lähitulevaisuudessa, mutta jo nyt ne toimivat paremmin kuin mitä kymmenen vuotta sitten olisin kehitystä ennustanut.

Näiden uusien mallien ja vanhempien koneoppimisalgoritmien välillä ei kuitenkaan ole mitään aivan selvää käsitteellistä eroa. ChatGPT on varsin erilainen kuin esimerkiksi sosiaalisen median sisältöjä hallinnoivat algoritmit, mutta molemmat perustuvat koneoppimiseen, joka löytää monimutkaisia tilastollisia säännönmukaisuuksia ihmisten verkkoon jättämästä datasta. Somealgoritmit luokittelevat käyttäjiä klikkausten ja tykkäysten perusteella, ja ne ennustavat, minkälaisia sisältöjä käyttäjälle kannattaa tarjota. ChatGPT luokittelee teksteissä esiintyviä sanoja ja muita ilmauksia niiden esiintymisyhteyksien mukaan, ja se ennustaa, minkälaista tekstiä käyttäjä siltä haluaa. Koneoppimisalgoritmeja on muunkinlaisia kuin syväoppimiseen perustuvia, mutta filosofisesti merkittäviä teknisiä eroja uusien ja vanhempien järjestelmien välillä on vaikea nähdä.

Vaikka tekoäly ja koneoppiminen on edistynyt viime vuosina, nykyiset syväoppimisjärjestelmät perustuvat jo 1980-luvulla kehitetyille neuroverkoille (Bowers 2017). Monet syväoppimisverkot ovat itse asiassa matemaattisesti osin yksinkertaisempia kuin aiemmin käytetyt verkot, mutta ne kompensoivat tätä suuremmalla koolla. Laskentatehon kasvu on mahdollistanut yhä suurempien verkkojen käyttämisen, ja vuosien varrella on keksitty parempia menetelmiä niiden opettamiseksi. Merkittävä murros tapahtui, kun vuoden 2018 paikkeilla keksittiin, miten verkot saadaan opettamaan itseään (Manning

2022). Tämä on käytännössä mahdollistanut yhä suurempiin datamassoihin perustuvan opettamisen, mutta ennen kaikkea tällaista dataa on aiempaa paremmin saatavilla. Ihmisten internetiin suoltama materiaali tarjoaa hyvin rikkaan ja helposti saatavilla olevan aineiston. Palaan tähän seikkaan lopuksi.

2. Laskennalliset mallit tieteessä

Laskennallisten mallien käyttö tieteessä voi tuoda mieleen menettelyn, jossa tarkasteltavan kohdejärjestelmän rakenne ja sen toimintaa kuvaavat lainomaisuudet pyritään kuvaamaan mahdollisimman tarkasti. Tämän jälkeen simulaatio paljastaa, kuinka järjestelmä käyttäytyy. Laskennallisia malleja kuitenkin käytetään hyvin monenlaisilla aloilla fysiikasta yhteiskuntatieteisiin, eikä mallien tarkoitus ole aina ennustaa tai edes kuvata kohdejärjestelmäänsä kovin tarkasti (Kuorikoski ja Ylikoski 2015; Potochnik 2015; Parker 2020). Silloinkin kun mallinnuksen kohteena on fysikaalinen järjestelmä, jonka toimintaa kuvaavat lait tunnetaan melko hyvin, joudutaan yksityiskohtia usein huomattavasti karkeistamaan, ettei simulaatio olisi liian raskas laskettavaksi. Esimerkiksi säämallit eivät tietenkään voi perustua ilmakehän hiukkasten liikkeiden mallintamiseen, eikä tällaiseen olisi mitään tarvettakaan. Sään ennustamiseen riittää paljon karkeampi mallinnus säärintamien ja ilmavirtauksien käyttäytymisestä.

Tieteelliset mallit ovat kuten kartat: liiat yksityiskohdat ovat tarpeettomia, ja ne tekevät malleista lähinnä mahdottomia käyttää. Esimerkiksi ilmastomallit ovat vielä karkeampia kuin säämallit, eikä niitä voi käyttää huomisen sään ennustamiseen. Molemmat kuvaavat kuitenkin periaatteessa samaa fysikaalista järjestelmää, mutta eri tarkkuudella ja eri tarkoituksia varten. Ilmastomallit kuvaavat pitkäaikaisia muutoksia toisistaan riippuvissa järjestelmissä, kuten ilmakehä, meret ja eräät maanpinnan ilmiöt. Näin monimutkaisissa järjestelmissä osa mekanismeista voi olla tuntemattomia ja jäädä malleista pois. Usein

malleihin joudutaan myös arvioimaan tuntemattomia muuttujia, jotta ne paremmin vastaisivat tunnettua dataa (ks. esim. IPCC 2014, luku 9).

Tieteellisissä malleissa ei aina ole ilmeistä, vastaavatko kaikki muuttujat mitään todellista mekanismia. Usein pyritään vahvistamaan, että malli kuitenkin kuvaa mallinnettavan järjestelmän todellisia ominaisuuksia, mutta se, tulisiko muuttujien edes vastata kohdejärjestelmän mekanismeja, riippuu mallin käyttötarkoituksesta. Jos tarkoitus on auttaa kohdejärjestelmää koskevien päätelmien tekemisessä, ei mallin sisäisillä mekanismeilla ole aina tiedollista merkitystä (Kuorikoski ja Ylikoski 2015). Tieteellisten mallien tarkoitus ei aina edes ole ennustaminen tai kohteen tarkka kuvaaminen. Mallintajat voivat olla kiinnostuneita selventämään erilaisten teoreettisten oletusten seurauksia. Tällaisia malleja kutsutaan joskus käsitteellisiksi tai teoreettisiksi malleiksi, ja ne ovat yleisiä esimerkiksi talous- ja yhteiskuntatieteissä.

Eräs tunnettu esimerkki on sosiologi Thomas Schellingin (1971) segregaatiomalli. Schelling tutki Yhdysvaltojen kaupunkien segregaatiokehitystä eli sitä, miksi esimerkiksi etniset ryhmät eriytyivät omille asuinalueilleen. Tyypillinen selitys segregaatiolle oli vähemmistövastaiset asenteet – käytännössä, että valkoinen väestö muuttaa pois alueilta, joissa asuu etnisiä vähemmistöjä. Schelling halusi osoittaa, ettei tätä tai mitään muutakaan yksittäistä selitystä voi päätellä suoraan segregaatiosta. Hän asetteli ruutupaperille kahdenlaisia merkkejä ja jätti osan ruuduista tyhjiksi. Merkit edustivat eri sosiaalisiin ryhmiin kuuluvia perheitä. Hän kävi merkkejä yksi kerrallaan läpi, ja mikäli merkin ympäristössä oli vain vähän samaan ryhmään kuuluvia merkkejä, se muutti lähimpään tyhjään sijaintiin, jossa saman ryhmän edustajia oli tarpeeksi. Kun tätä jatkoi aikansa, merkit aina joko sekoittuivat satunnaisesti tai ne päätyivät kahdeksi erilliseksi ryhmäksi. Schelling huomasi, että hänen mallissaan täydellinen segregaatio syntyy jo silloin, kun muuttajat suosivat alueita, joissa on vähintään 1/3 heidän kaltaisiaan, eli vaikka perheet olisivat valmiita muutamaan alueille, joissa he itse ovat vähemmistöä.

Mallin tavoite ei ollut osoittaa, etteikö vähemmistövastaisuus voisi olla selitys segregaatiolle. Schelling ei myöskään väittänyt simuloivansa oikeaa kaupunkien muuttoliikettä tai sen taustalla olevia mekanismeja. Hänen pointtinsa oli, että samanlainen makrotason käyttäytyminen voi periaatteessa seurata erilaisista mikrotason mekanismeista. Malli osoitti lähinnä, että segregaatiokehitys ei välttämättä edellytä muuttajien halua päästä eroon vähemmistöistä, koska sama ilmiö voi periaatteessa syntyä myös, jos ihmiset suosivat alueita, joissa asuu edes jonkin verran heidän kaltaisiaan. Molemmat mekanismit vaikuttavat äkkiseltään uskottavilta, ja ilmiön todellisen luonteen selvittäminen vaatii erillistä empiiristä tutkimusta.

Schellingin malli oli simulaationa äärimmäisen yksinkertainen ja siksi myös havainnollinen. Joskus tieteelliset mallit ovat lähinnä havainnollistamisvälineitä, joiden tarkoitus on tutkia, minkälaiset selitykset ovat uskottavia ja minkälaisia seurauksia erilaisilla teoreettisilla oletuksilla voi olla. Tällainen tutkimus on yksi tapa ymmärtää myös tekoälymallien roolia mielenfilosofiassa tai muussa mielen teoreettisessa tutkimuksessa. Tekoälymalleilla on tosin katsottu olevan muunkinlaisia käyttötarkoituksia, jotka ovat lähempänä täsmällisiä simulaatiomalleja empiirisen tutkimuksen tukena.

3. Tekoälyjärjestelmät tieteellisinä malleina

Kognitiotieteen historiasta löytyy kaksi tekoälyn kehityslinjaa, joista toista kutsutaan perinteiseksi tekoälyksi ja toista neurolaskennaksi tai konnektionismiksi. Molempien juuret ovat 1950-luvulla, mutta perinteinen tekoäly, jota kutsutaan myös logiikkapohjaiseksi tai symboliseksi tekoälyksi, oli vallitseva lähestymistapa 1980-luvulle asti. Sen taustalla vaikutti useita ideoita. Ydinajatuksen voi tiivistää vaikka niin, että ihmiset ovat pääpiirteissään rationaalista ja logiikka kertoo, mitä rationaalisuus on, joten inhimillinen järjenkäyttö perustuu logiikalle. Toki tiedettiin, että ihmiset eivät aina toimi loogisesti, mutta parempaakaan yleistä teoriaa nimenomaan käsitteellisestä järjenkäytöstä ei ollut. Lähtökuopissaan ollut tietojenkäsittelytiede

oli myös vahvasti sidoksissa matemaattiseen logiikkaan, joten logiikka tarjosi luontevan lähtökohdan tietojenkäsittelytieteilijöille ja kognitiivisille psykologeille yhdistää voimansa.

Taustalla vaikutti myös analyyttisestä filosofiasta tuttuja käsityksiä, joiden mukaan kieli tai käsitteellinen ajattelu voidaan periaatteessa kuvata logiikan tai jonkun vastaavan säännönmukaisen formaalisen kielen avulla. Monien mielestä tämä oli itse asiassa välttämätöntä. Esimerkiksi Jerry Fodorin (1975) mukaan oletus aivoissa tapahtuvasta symbolirakenteiden kombinatorisesta käsittelystä on ainoa hypoteesi, joka voi selittää, kuinka ihmiset kykenevät ymmärtämään periaatteessa mielivaltaisen määrän erilaisia monimutkaisia lauseita. Vastaavanlaisia ideoita oli aiemmin esittänyt kielitieteilijä Noam Chomsky (1957), jonka vaikutus kognitio- ja tietojenkäsittelytieteiden kehitykseen oli huomattava.

Perinteisen tekoälyn toinen ajatus oli, että mielen toimintaa koskevat teoriat voitaisiin muotoilla täsmällisesti tietokoneohjelmina. Ohjelmia voidaan suorittaa tietokoneilla, ja niiden tulosteita verrataan ihmisten toimintaan, kun he esimerkiksi ratkovat erilaisia päättelytehtäviä (esim. Newell ja Simon 1972; Boden 2006, luvut 6 ja 7). Tätä tarkoitusta varten näiden mallien tarkoitus oli siis olla tarkkoja kuvauksia ja simulaatioita ihmismielen toiminnasta, vaikkakin ne yleensä oli rajattu kuvaamaan vain esimerkiksi päättelyä tietynlaisissa tehtävissä.

Alkuinnostuksen jälkeen osoittautui, ettei formaalinen logiikka kuitenkaan ole toimiva malli ihmisjärjestä. Rajatuissa tehtävissä logiikka tai muut vastaavat eksplisiittisiin sääntöihin perustuvat järkeilymallit yleensä toimivat, mutta kun siirryttiin rajatuista tehtävistä monimutkaiseen inhimilliseen maailmaan, seurasi ongelmia. Perinteinen logiikka toimii, kun tieto on varmaa, muuttumatonta ja ristiriidatonta, mutta tämä ei kuvaa alkuunkaan arkista todellisuuttamme.

Oletetaan esimerkiksi, että ystäväsi sanoo tulevansa illalla baariin, jos hän saa työnsä tehtyä. Myöhemmin hän lähettää viestin, että hän sai työnsä tehtyä, mutta jäi auton alle ja on nyt sairaalassa. Tuleeko hän baariin vai ei? Logiikka sanoo ”kyllä”

mutta terve järki ”ei”. Logiikkapohjaiseen järjestelmään voidaan toki lisätä sääntö, että jos joku on sairaalassa, hän ei ole tulossa baariin. Toisaalta taas tällöin hän on tulossa baariin, koska hän sai työnsä tehtyä, mutta toisaalta ei ole tulossa baariin, koska hän on sairaalassa. Seuraa siis ristiriita, joka on jotenkin ratkaistava. Lisätään taas sääntö, että sairaalassa oleminen kumoaa baariin tulemisen, mutta on aika ilmeistä, että tällaisia poikkeuksia ja ristiriitatilanteita on ihmiselämässä lähes äärettömästi. Jos ne pitäisi kaikki kirjata ylös, tehtävä olisi mahdoton. Poikkeusten tähdellisyys riippuu usein myös monista asiayhteyteen vaikuttavista tekijöistä (ystävä voi olla sairaalassa hakemassa sukulaistaan), ja yksinkertaisissakin arkipäivän asioissa vaadittavien päättelyketjujen määrä räjähtää äkkiä tähtitieteellisen suureksi. Jos koneeseen koetetaan saada älyä tällaisten sääntöjen avulla, ei tietenkään voida olettaa, että sillä jo olisi jonkinlaista järkeä arvioida eri seikkojen tähdellisyyttä.

Tällaiset ongelmat ajoivat perinteisen tekoälyn umpikujaan 1980-luvulla. Myös kognitiivisen psykologian puolella alkoi kasautua empiirisiä tuloksia, joiden mukaan inhimillinen järkeily ei yleensä noudata logiikkaa edes yksinkertaisissa tehtävissä (ks. Evans ym. 1993). Logiikka ei siis tarjonnut täsmällistä tai tuskin kovin suurpiirteistäkään mallia ihmisälystä. Lähestymistavan hylkäämiseen vaikutti myös noihin aikoihin perinteisen tekoälyn vaihtoehdoksi vahvasti kehittyvä neuroverkkoteoria.

Neuroverkkoteoria syntyi 1940-luvulla (McCulloch ja Pitts 1943), mutta lähestymistapa sai tuulta purjeisiin vasta Frank Rosenblattin (1958, 1962) tutkimusten myötä. Hänen ajatuksensa ei ollut mallintaa abstraktia ajattelua vaan hermosoluryhmien tietojenkäsittelyä ja assosiatiivista oppimista. Tiedettiin, että hermosoluverkostojen toiminta perustuu toimintapotentiaaliksi kutsuttujen signaalien välittämiseen soluilta ja soluryhmiltä toisille. Esimerkiksi verkkokalvolle osuva valo aktivoi aistinsoluja, joiden lähettämä signaali kulkee verkkokalvon alla sijaitseville hermosoluille. Kun nämä solut aktivoituvat, signaali etenee edelleen takaraivon näköaivokuorelle ja sieltä muualle

aivoihin. Oli ilmeistä, että hermosolut itse eivät suorita monimutkaista tietojenkäsittelyä, joten aivojen tietojenkäsittelyn täytyy tapahtua, kun signaali etenee soluryhmiltä toisille. Tiedettiin myös, että solujen väliset kytkennät muuttuvat solujen aktivaation seurauksena. Yhdessä aktivoitujen solujen yhteydet vahvistuvat, mutta jos solujen aktivaatio ei liity toisiinsa, niiden väliset kytkennät heikkenevät.

Näiden periaatteiden nojalla Rosenblatt ryhtyi mallintamaan hermosoluryhmien toimintaa. Hänen mallinsa eivät varsinaisesti kuvanneet aivobiologiaa, vaan ne pyrkivät karkeasti mallintamaan yllä mainitut periaatteet, joita pidettiin oleellisena hermoston tietojenkäsittelylle. Ero logiikkapohjaisiin malleihin oli selvä. Nämä järjestelmät poimivat tilastollisia yhteyksiä hermosoluryhmien aktivaatiokuvioissa ja liittivät niitä toisiinsa. Esimerkiksi jos kaksi havaintoa liittyivät säännönmukaisesti yhteen, neuroverkko kykeni ainakin periaatteessa yhdistämään ne toisiinsa ja ennustamaan, että tietystä ärsykkeestä seuraa toinen. Alun perin Rosenblattin *perseptroniksi* kutsumia verkkoja käytettiin lähinnä visuaaliseen hahmontunnistukseen, mutta perusajatus muistuttaa assosiatiivisen psykologian periaatteita yleisemmin.

Ongelma oli, että yksinkertaiset verkot eivät kyenneet kovin monimutkaisiin temppuihin ja monimutkaisempia verkkoja ei osattu opettaa. Ongelma ei ollut soluryhmien koko vaan signaaliketjujen pituus. Mielivaltaisen suuria soluryhmiä voitiin kyllä kytkeä yhteen, mutta jos kytkentäkerroksia oli useampi kuin yksi, verkkoa ei osattu opettaa. Neuroverkkotutkimus hyytyi, kun kävi ilmeiseksi, että yhden kytkentäkerroksen verkot ovat periaatteessakin kykenemättömiä suoriutumaan monista yksinkertaisista tehtävistä (Minsky ja Papert 1969). Tilanne muuttui 1980-luvulla, jolloin keksittiin nykyisten syväoppimisjärjestelmienkin taustalla oleva menetelmä opettaa verkkoja, joissa kytkentäkerroksia ja soluryhmiä voi olla mielivaltaisen paljon (Rumelhart ym. 1986).

Termi ”syväoppiminen” tulee juuri siitä, että neuroverkon kerroksilla voi olla ”syvyyttä” niin paljon kuin laskentateho sallii. Termi tosin lanseerattiin vasta 2000-luvun puolella. Viime

vuosituhannen lopulla laskentateho ei nimittäin paljoa sallinut. Verkoissa oli yleensä vain muutamia kerroksia verrattain pieniä soluryhmiä. Tällöinkin niiden opettaminen oli työlästä. Neuroverkkoja arvosteltiin siitä, että ne olivat liian karkeita malleja ollakseen uskottavia kuvauksia hermoston toiminnasta, ja toisaalta niiden suorituskykykään ei vastannut ihmisten psykologisia kykyjä. Oli teoreettisia syitä olettaa, että neuroverkoissa olisi mahdollisuuksia vaikka mihin, mutta vastaavia perusteluja löytyi myös logiikkaan pohjautuville malleille. Näytöt kuitenkin jäivät verrattain vaatimattomiksi. Neuroverkot kohtasivat oleellisesti saman ongelman kuin logiikkapohjaiset mallit: kun ongelmat menivät monimutkaisiksi, järjestelmät eivät skaalautuneet, ja suorituskyky yksinkertaisesti sakkasi. Ihmismielen malleina, tai oikeammin mallijoukkoina, molemmat jäivät teoreettisten tarkastelujen asteelle. Tässä mielessä ne muistuttivat lopulta ehkä enemmän Schellingin segregaatiomallia kuin tarkkoja tai edes kovin karkeita simulaatiomalleja mielen toiminnasta.

4. Älykkäät välineet mielen malleina

Vuosituhannen taitteessa neuroverkot jäivät hieman muiden koneoppimismenetelmien varjoon. Niiden kehitys ei tosin lakannut kuten ei logiikkamallienkaan. Lopulta hermoverkot kuitenkin löivät läpi varsinaisten syväoppimisverkkojen muodossa, jotka voivat sisältää miljoonia soluja ja satoja kerroksia. Artikkelin alussa mainitsin jo syitä, jotka johtivat näiden järjestelmien läpimurtoon. Palaan alun teemoihin muutaman huomion kanssa.

Syväoppimisjärjestelmiä ei ole kehitetty aiempaa paremmiksi ihmismielen malleiksi. Ne ovat syntyneet tietoteknisen perustutkimuksen seurauksena, ja uudet tekoälysovellukset, kuten ChatGPT, on laadittu lähinnä tuotantotaloudelliseksi teknologiaksi. Tästä huolimatta ne ovat ainakin osin lunastaneet niitä odotuksia, joita tekoälylle on aiemmin asetettu. Nämä järjestelmät voivat olla hämmästyttävän älykkäitä, joustavia ja suorituskyvyltään jopa ihmismäisiä aiempiin verrattuna. Ei silti

ole mitään syytä olettaa, että niiden toiminta vastaisi ihmismielen toimintaa ainakaan mekanismien tasolla.

Tämä ei kuitenkaan tarkoita, etteikö uusista tekoälymalleista voisi oppia jotain mielenfilosofian kannalta mielenkiintoista. Tieteelliset mallit eivät yleensäkään ole kohdejärjestelmänsä tarkkoja kuvauksia, joten mikään ei estä tarkastelemasta myös uusia tekoälymalleja sellaisina. Tieteessä mallit toimivat usein heuristisina ja argumentatiivisina välineinä muun teoreettisen päättelyn ohessa (ks. esim. Kuorikoski ja Ylikoski 2015). Mallin täsmällisillä mekanismeilla ei ole välttämättä suurta merkitystä. Teoreettisten mallien tapauksessa edes mallin vastaavuus kohdejärjestelmän kanssa ei aina ole ratkaisevan tärkeää, jos jotkut mallin keskeisistä ominaisuuksista kuitenkin voivat kertoa jotain mielenkiintoista sen kohteesta. Esimerkiksi logiikkamallit eivät kuvaa mielen toimintaa, mutta ne eivät ole tieteellisesti hyödyttömiä. Vastaavasti uusien tekoälyjärjestelmien kyvyt voivat tukea joitakin teoreettisia käsityksiä ihmismielestä – täsmällisemmin sanottuna esimerkiksi kielellisestä ajattelusta suurten kielimallien tapauksessa.

Mikä suurten kielimallien, kuten ChatGPT, opetus mielenfilosofialle tässä mielessä voisi olla? Suuren koon, teknisten yksityiskohtien ja laskentakapasiteetin sijaan huomion voi suunnata verkkojen opettamiseen ja oppimateriaaliin. Esimerkiksi ChatGPT:n taustalla oleva kielimalli on opetettu ennustamaan, miten ihmisten tuottamat tekstit ja keskustelut etenevät, ja se oppii palautteesta, kun se koettaa uusintaa oppimaansa. Valtavan opetusmateriaalin ja suuren parametriavaruuden ansiosta malli ei vain toistele kirjaimellisesti oppimaansa vaan se kykenee myös joustavasti tuottamaan ja tulkitsemaan tekstiä, jota sille ei ole opetettu (ks. Manning 2020). Voisi kai sanoa, että tällaiset järjestelmät kykenevät implisiittisesti sisäistämään mielekästä kielenkäyttöä säätelevät normit, vaikka ne eivät niitä eksplisiittisesti tiedä. Tämä lienee pääpiirteissään totta myös ihmisten tyypillisestä kielikyvystä.

Muun muassa ruumiillisen ja laajennetun kognition sekä kognitiivisen ekologian tutkimusten myötä moni nykyinen

kognition filosofi on omaksunut ajatuksen, että inhimillinen rationaalisuus ei pohjaudu sisäsyntyiseen järkeen vaan se kehittyy vuorovaikutuksessa ympäristön kanssa. Kognitio ei ole deduktio- vaan induktiojärjestelmä, joka jäljittää toimintamme kannalta tärkeitä ympäristön säännönmukaisuuksia. Ihmisten tapauksessa tähän kuuluu myös sosiaalinen ympäristö, mukaan lukien säännönmukaisuudet toisten ihmisten kielellisessä käyttäytymisessä. Opimme muilta ihmisiltä, miten järkeä käytetään, samaan tapaan kuin – tai samalla kun – opimme, miten kieltä käytetään. Näin ollen järjenkäyttö ja käsitteellinen ajattelu pohjautuu ympäristömme ja rutiiniemme toisteisuuteen sekä toisilta ihmisiltä saamiimme esimerkkeihin, opastukseen ja palautteeseen.

Tämä on kiistanalainen, mutta ei ole kovin yllättävä ajatus. Näin opimme taitoja ylipäätään. Uutinen on lähinnä, että kielellisenä kykynä näkyvä järjenkäyttö ei ole erotettavissa kielellisistä sisällöistä tai kielen oppimisen taustalla olevista rutiineista, eikä kielellisen järjenkäytön taustalla ole mitään yleisestä proseduraalisesta ja tilastollisesta oppimisesta erillistä kognitiivista kykyä.

Samaan tapaan suurissa kielimalleissa kieltä ja maailmaa koskeva tieto eivät ole erillisiä kuten eivät myöskään syntaksi, semantiikka, pragmatiikka tai kyky kielelliseen järkeilyyn. Oikeastaan kielimallit eivät erityisemmin järkeile. Ne vain käyttävät kieltä tilastollisen oppimisen pohjalta ilman erillistä loogista tai muuta deduktiota. Kielimallit eivät tietenkään osoita, että sama pätisi ihmisten kielellisestä kognitiosta. Kysymyksen ratkaiseminen vaatii muutakin teoreettista tarkastelua ja lopulta empiiristä näyttöä. Tällaista tutkimusta tosin on olemassa, eikä tässä luonnehtimani käsitys kielellisestä kognitiosta varsinaisesti pohjaudu syväoppismalleihin. Samanlaisia teorioita on esitetty aiemmin kognitiivisen kielitieteen niin sanotun kielen käyttöteorian yhteydessä (esim. Tomasello 2003). Kielifilosofian puolella nämä ajatukset ovat sukua kielelliselle pragmatismille, jonka mukaan kielikyky perustuu sosiaalisten rutiinien implisiittiselle oppimiselle (ks. esim. Haugeland 1990).

Johtopäätökseni saattavat näyttää lähinnä assosiatiiviselta ajattelulta, jossa ensin todetaan, että uudet kielimallit ovat yllättävän hyviä, ja sitten todetaan jotain laadullisia samankaltaisuuksia niiden sekä tiettyjen kognitiivisten kieliteorioiden välillä. Voidaan väittää, että suuret kielimallit ja muut uudet tekoälyt ovat ihmisjärjen ymmärtämisen kannalta irrelevantteja, koska:

(1) ne eivät lopulta kuvaa ihmismielen psykologisia prosesseja eivätkä siis kerro niistä mitään;

(2) ne oppivat kieltä eri tavalla kuin ihmiset;

(3) ne ovat psykologisilta ja behavioraalisilta kyvyiltään rajoittuneita verrattuna ihmisiin; ja

(4) tarpeeksi tehokas oppimisjärjestelmä oppii uusintamaan mitä tahansa säännönmukaisuuksia opetusdatassa; näin ollen on triviaalia, että sellaiset oppivat matkimaan myös kielenkäyttöä.

Toistan vielä, että syväoppimiseen perustuvat mallinnustekniikat ovat kuitenkin mahdollistaneet laadullisen hyppäyksen luonnollisen kielen prosessoinnissa ja arkisen järjenkäytön matkimisessa.

Kohtaan 4. voidaan todeta, että ei ole mitenkään triviaali seikka, että syväoppimisjärjestelmät kykenevät oppimaan joustavaa kielenkäyttöä. Kognitiotieteen historiasta löytyy tunnettuja argumentteja, joiden mukaan sen pitäisi olla käytännössä mahdotonta assosiatiivisille oppimisjärjestelmille. Uusien tekoälyjärjestelmien suorituskyky kuitenkin puhuu tätä vastaan. Samat laskentatehoon liittyvät mahdollisuudet ovat joka tapauksessa avoinna muillekin lähestymistavoille, mutta näytöt ovat toistaiseksi syväoppimisjärjestelmien puolella. Asia voi toki muuttua jatkossa.

Argumenttini kannalta on epäoleellista, etteivät suuret kielimallit kuvaa ihmisten psykologisia prosesseja tai kielen oppimista täsmälleen oikein. Mikäli näitä malleja tarkastellaan heuristisina välineinä, jotka auttavat meitä ymmärtämään jotakin ihmiskognitiosta, ei tarvitse olettaa, että mallin mekanismit

vastaisivat psykologisia mekanismeja tai että mallit selittäisivät kaikkea ihmisjärjestä tai edes kielikyvystä. Tarkasteluun riittää poimia jotain yleisiä mutta teoreettisesti mielenkiintoisia samankaltaisuuksia mallien ja ihmismieltä koskevien teorioiden välillä. Mielestäni oleellinen piirre on juuri opetusmateriaali ja opetusmateriaalissa piilevien rakenteiden tilastollinen oppiminen. Suurten kielimallien perusteella vaikuttaa selvältä, että ihmisten kielelliseen ulosantiin sisältyvien monimutkaisten toisteisuuksien sisäistäminen riittää hyvin pitkälle selittämään joustavaa kielellistä ymmärrystä ja järkeilyä. Vaikka suuret kielimallit eivät suoraan todista ihmiskognitiosta mitään, ne antavat argumentatiivista tukea kognition teorioille, joiden mukaan arkinen järkeilykyky perustuu tilastollisten säännönmukaisuuksien oppimiseen ihmisten kielellisestä vuorovaikutuksesta.

Myös logiikkapohjaiset tekoälymallit olivat merkittävä tieteellinen hanke, koska niiden umpikuja antaa argumentatiivista tukea sille, ettei arkinen järkeily nojaa logiikkaan. Kun syväoppimiseen perustuvat kielimallit toimivat varsin hyvin ja logiikkaan perustuvat mallit huonosti, voidaan malleja vertailemalla todeta konkreettisten näyttöjen tukevan tilastolliseen induktioon perustuvaa käsitystä kielikyvystä. Uudet tekoälymallit toimivat toki vain käsitteellisinä malleina, mutta erityisesti ne osoittavat vääräksi sen syvään pinttyneen käsityksen, että luova ja joustava kielen- tai järjenkäyttö edellyttää rekursiivisiin sääntöihin perustuvaa symbolirakenteiden käsittelyä.

Tekoälymallien puutteet voivat siis olla tieteellisesti yhtä mielenkiintoisia kuin niiden vahvuudet. Tarkastellaanpa lopuksi esimerkkinä ChatGPT:n heikkouksista seuraavaa kysymystä: Liisa katsoo Jarkkoa; Jarkko katsoo Merviä. Liisa on naimisissa; Mervi ei. Voidaanko päätellä, katsooko joku naimisissa oleva jotakuta, joka ei ole naimisissa?

Suurin osa ihmisistä vastaa kysymykseen spontaanisti, että ei voi päätellä. Emmehän tiedä, onko Jarkko naimisissa. Oikea vastaus on kuitenkin kyllä. Jarkko nimittäin joko on naimisissa tai hän ei ole. Jos hän on, niin joku naimisissa oleva (Jarkko) katsoo jotakuta, joka ei ole naimisissa (Mervi). Jos Jarkko ei ole

naimisissa, niin joku naimisissa oleva (Liisa) katsoo jotakuta, joka ei ole naimisissa.

Myös ChatGPT vastasi kysymykseen väärin. Joissakin päättelytehtävissä se tekee samankaltaisia virheitä kuin ihmiset. Tässä tapauksessa sen vastaus oli kuitenkin epätyypillisen järjetön, eikä se yksinkertaisesti kyennyt ymmärtämään oikean vastauksen ajatusta, vaikka sitä kuinka koetin opastaa. Useimmat ihmiset taas tajuavat pointin melko vaivatta, kun sen heille kertoo. Ehkä tämä kyky perustuu mentaaliseen logiikkaan, mentaalisiin malleihin tai johonkin muuhun. Joka tapauksessa herää kysymys, mitä monimutkaisen assosiatiivisen oppimisen lisäksi tarvitaan, jos pelkästään sillä pääsee todella pitkälle, mutta se hyytyy näin yksinkertaisissa asioissa. Ehkä kielimallia pitää laadullisesti muuttaa. Ehkä riittää tehdä siitä entistä suurempi. Olipa miten hyvänsä, tämä looginen tai vastaava oivalluskyky ei näytä olevan inhimillisenkään järjenkäytön keskiössä, koska tämä yksinkertainen päättelytehtävä on myös useimmille ihmisille vaikea spontaanisti ymmärtää.

Tampereen yliopisto

Kirjallisuus

Boden, Margaret (2006). *Mind as Machine: A History of Cognitive Science.* Oxford: Oxford University Press.

Bowers, Jeffrey S. (2017). ”Parallel Distributed Processing Theory in the Age of Deep Networks”, *Trends in Cognitive Sciences* 21(12), 950–961.

Chomsky, Noam (1957). *Syntactic Structures.* Haag: Mouton.

Evans, Jonathan St. B. T., Stephen E. Newstead ja Ruth M. J. Byrne (1993). *Human Reasoning: The Psychology of Deduction.* Hove: Psychology Press.

Fodor, Jerry A. (1975). *The Language of Thought.* Cambridge, MA: Harvard University Press.

Haugeland, John (1990). ”The Intentionality All-Stars”, *Philosophical Perspectives* 4, 383–427.

ICPP (2014). *AR5 Synthesis Report: Climate Change 2014.* https://www.ipcc.ch/site/assets/uploads/2018/02/SYR_AR5_FINAL_full.pdf (16.09.2023).

Kuorikoski, Jaakko ja Petri Ylikoski (2015). ”External Representations and Scientific Understanding”, *Synthese* 192(12), 3817–3837.

Manning, Christopher D. (2022). ”Human Language Understanding & Reasoning”, *Dædalus* 151(2), 127–138.

McCulloch, Warren S. ja Walter Pitts (1943). ”A Logical Calculus of the Ideas Immanent in Nervous Activity”, *Bulletin of Mathematical Biophysics* 5(4), 115–133.

Minsky, Marvin L. ja Seymour A. Papert (1969). *Perceptrons: An Introduction to Computational Geometry.* Cambridge, MA: The MIT Press.

Parker, Wendy S. (2020). ”Model Evaluation: An Adequacy-for-Purpose View”, *Philosophy of Science* 87(3), 457–477.

Potochnik, Angela (2015). ”The Diverse Aims of Science”, *Studies in History and Philosophy of Science* 53, 71–80.

Rosenblatt, Frank (1958). ”The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain”, *Psychological Review* 65(6), 386–408.

Rosenblatt, Frank (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms.* Washington D.C.: Spartan Books.

Rumelhart, David E., James L. McClelland ja PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* Cambridge, MA: The MIT Press.

Schelling, Thomas C. (1971). ”Dynamic Models of Segregation”, *Journal of Mathematical Sociology* 1(2), 143–186.

Tomasello, Michael (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition.* Cambridge, MA: Harvard University Press.

Turing, Alan M. (1950). ”Computing Machinery and Intelligence”, *Mind* 59(236), 433–460.