Otto Sahlgren

Oikeudenmukaisuutta ymmärtävä tekoäly – vain ykkösiä ja nollia?

Tekoäly päättelee ja ennustaa. Robottikuoreen upotettuna se myös kävelee ja puhuu. Yhä autonomisemmin ja sopeutuvammin toimivat tekoälyjärjestelmät ovat myös filosofisen ihmettelyn kohteena. Perinteiset filosofiset kysymykset tiedosta, ymmärryksestä ja oikeudenmukaisuudesta kietoutuvat yhteen oppivien koneiden ympärillä. Voiko ihminen luottaa koneen asiantuntijuuteen, vai onko tarvetta pitää robotti ruodussa?

Tammikuussa Tampereella järjestetyssä Suomen Filosofisen Yhdistyksen vuosikollokviossa tarkasteltiin laajaa kirjoa tekoälyn ja robotiikan kysymyksiä. Onko tekoälyn kehityksessä tapahtunut viimeisellä vuosikymmenellä jotakin olemuksellisesti uutta ja erityistä? Vai onko ainoa uutuus se, että tekoäly lävistää yhteiskunnan eri osa-alueet vahvemmin kuin koskaan, kuten Samuli Reijula, Jaakko Lehtinen ja Jaakko Kuorikoski paneelikeskustelussaan pohtivat? Merkittävänä uutuutena lienee ainakin tekoälyn suorittamien tehtävien kasvava monimutkaisuus. Mutta mikä erottaa esimerkiksi ymmärtämisen tai moraalisen toiminnan tehokkaasta, mekaanisesta laskennasta?

Miten konemieli toimii?

Kognitiivista ja episteemistä toimijuutta esitelmässään tarkastellut Anna-Mari Rusanen katsoi, että tekoälyä voidaan pitää kapeassa mielessä kognitiivisena toimijana. Se suorittavaa tehtäviä, kuten tunnistusta ja luokittelua, prosessoimalla suuria määriä informaatiota. Tekoäly ei kuitenkaan nykyisellään pysty samankaltaisiin oleellisuusarvioihin näissä tehtävissä kuin ihminen. Esimerkiksi kuvantunnistusalgoritmeilta puuttuu kyky arvioida, mitkä ympäristön ärsykkeiden esille nousevista piirteistä ovat relevantteja tehtävän suorittamisen kannalta. Niiden toiminta on sidottu datassa esiintyviin tilastollisiin yhteyksiin, kun taas ihmisaivojen korkeamman tason prosessit ohjaavat etsimään

”Robottiruumis linkkinä maailmaan mahdollistaa oppimisen ympäristöä tutkimalla ja muokkaamalla.”

havainnosta piirteitä, jotka lisäävät todennäköisyyttä suorittaa kognitiivisia tehtäviä.

Evoluution esivirittämä ihminen ymmärtää, mikä on tähdellistä, mutta tätä taitoa tekoäly ei vielä hallitse. Tietyissä tehtävissä tekoäly ei kenties tarvitsekaan tällaista kykyä. Syväoppivia neuroverkkoja hyödyntävät järjestelmät, kuten Google DeepMindin Go-peliä pelaava AlphaGo, suoriutuvat rajatuissa konteksteissa tehtävistään hämmentävän tehokkaasti ja tarkasti, paremmin kuin ihminen[1]. Tällaisten järjestelmien toimintaa voitaisiin Ilkka Niiniluodon mukaan tulkita dreyfusilaisen asiantuntijatiedon näkökulmasta: jopa miljoonien harjoitusesimerkkien avulla oppinut neuroverkko on ekspertti, joka toimii sääntöjen seuraamisen sijaan ”intuitiivisesti” ja rationaalisesti[2].

Syväoppivien neuroverkkojen kohdalla on tultu pitkälle symbolisesta, sääntöihin pohjautuvasta tekoälystä, johon John Searlen vaikutusvaltainen kritiikki aikanaan kohdistui[3]. Searlen ”kiinalaisen huoneen” ajatuskokeen ympärillä käytyä keskustelua tarkastellut Panu Raatikainen katsoi, että vaikka Searlen alkuperäinen argumentti voitaisiin hyväksyä, hänelle esitetyt vasta-argumentit (niin kutsutut järjestelmä- ja robottivastaus) saattaisivat yhdessä toimia puolustuksena mahdollisuudelle, että tekoäly ymmärtäisi kieltä[4].

Syntaksista ei ehkä yksinään saada semantiikkaa, mutta järjestelmä, joka on kausaalisesti vuorovaikutuksessa ympäristönsä kanssa, havainnoi ja liikkuu, voisi mahdollisesti muodostaa semanttisia yhteyksiä manipuloitujen symbolien välille. Pii Telakivi ja Valtteri Arstila esittivätkin, että aktiivinen ja adaptiivinen, ruumiillinen vuorovaikutus ympäristön kanssa voidaan nähdä mahdollisena reunaehtona konemielen syntymiselle. Kognition riippuvaisuutta ruumiista on korostettu kehollisen mielen teoriasuuntauksen sisällä.[5] Tietoisuuden kehittyminen roboteillakin edellyttäisi täten järjestelmältä sensorimotorista kytkentää ympäristöön. Robottiruumis linkkinä maailmaan mahdollistaa oppimisen ympäristöä tutkimalla ja muokkaamalla. Kuten Rusanenkin huomautti, keskushermoston merkitys on suurelta osin kytköksissä organismin mahdollisuuteen liikkua. Ymmärrykselle, tietoisuudelle ja autonomiselle toimijuudelle näyttäisi olevan olennaista (inter)aktiivinen ja päämääräsuuntautunut toiminta ympäristössä.

Oikeintekevät robotit ja reilut algoritmit

Myös Antti Kauppisen esitelmässä ymmärryksen mahdollisuus oli suuressa roolissa. Kauppinen kysyi, millainen kone voisi olla moraalinen toimija ja miten se rakennettaisiin. Moraalisen toimijan erottaa ”pelkästä” oikein tekevästä toimijasta hänen mukaansa se, että vaikka molemmat toimisivat johdonmukaisesti oikein moraalisessa mielessä, vain moraalinen toimija ymmärtää, miksi näin toimitaan. Ymmärrys toiminnan perusteista, oikeasta ja väärästä, motivoi moraalista toimijaa toimimaan oikein. Moraalisen toimijuuden myötä myös moraalinen vastuu nousee keskeiseksi kysymykseksi – mutta voiko robotti kantaa vastuuta, kuten Pekka Mäkelä kysyi. Olennaista on, perustuuko robotin toiminta aidosti sen omille arvoille ja tavoitteille. Ottaen huomioon, että robotti on ihmisen rakentama järjestelmä, tämä ei näyttäisi toteutuvan.

Autonomiselle moraaliselle toiminnalle ei riitä vapaus valita ihmisen ennalta koodaamien vaihtoehtojen väliltä. Toimiminen datasta louhittujen säännönmukaisuuksien pohjalta ei vaikuttaisi sekään riittävän. Valinnan tulisi olla autenttisesti toimijan itsensä. Henrik Rydenfeltin mukaan eettinen ymmärrys ja arvojen autenttisuus voivat olla edellytys sille, että voimme todella luottaa tekoälyyn. Koneen asianmukainen toiminta rakentaa luottamusta sen ennakoitavuuteen ja johdonmukaisuuteen, mutta ihmisten välinen aito luottamus sisältää myös eettisen motivaation ulottuvuuden. Rydenfeltin mukaan niin kutsuttu froneettinen luottamus edellyttää luottamuksen kohteelta, eli koneelta, intentionaalisuutta. Froneettinen luottamus perustuu luottajan odotukseen siitä, että luottamuksen kohde myös itse uskoo toimivansa oikein toimiessaan luottajan odottamalla tavalla.

Kuten Kauppinen pohti, saattaa kuitenkin riittää, että koneet ovat oikeintekijöitä, sen sijaan että pyrkisimme rakentamaan ymmärtämiseen kykeneviä moraalisia toimijoita. Oikeintekijöidenkin rakentaminen on kuitenkin hankalaa. Esimerkiksi automatisoitu päätöksenteko datan tilastollisten yhteyksien perusteella on osoittautunut vaikeaksi tekoälyn eettisen suunnittelun ja käytön näkökulmasta. Tekoälyn on esimerkiksi huomattu syrjivän päätöksissään ihmisiä muun muassa etnisyyden ja sukupuolen perusteella[6]. Lisäksi päätöksentekoprosessia voi olla vaikea ymmärtää algoritmien

monimutkaisuuden vuoksi, mitä on kirjallisuudessa kutsuttu ”mustan laatikon ongelmaksi”[7].

Arto Laitinen eritteli esitelmässään samaa tematiikkaa päätöksen selityksen ja oikeutuksen näkökulmasta. EU:n GDPR-tietosuoja-asetusta tarkastellut Sandra Wachter tutkimusryhmineen on esittänyt, että GDPR ei sisällä vahvaa yksilön lainmukaista oikeutta ymmärrettävään selitykseen algoritmisen päätöksenteon kontekstissa[8]. Ryhmän mukaan sellainen kuitenkin tarvitaan. Ymmärrettäväksi selitykseksi riittäisi, että asianomaiselle yksilölle kerrotaan, minkä tekijöiden olisi tarvinnut olla toisin, jotta algoritmi olisi tuottanut erilaisen lopputuleman hänen kohdallaan. Tällaiset kontrafaktuaaliset selitykset eivät edellytä algoritmin ”konepellin alle katsomista”.

Mutta tulisiko selityksen lisäksi myös päätöksen oikeutusta tarkastella kontrafaktuaalisesti? Jos esimerkiksi päätöksen kohteena olevan yksilön sukupuoli vaikuttaa päätösprosessiin, mutta on epäolennainen päätöksen kannalta, voitaisiin päätöksen katsoa olevan kontrafaktuaalisesti epäreilu[9]. Laitinen huomautti kuitenkin, että tämä ei yksinään riitä tilanteessa, jossa kaikkia kohdellaan jo lähtökohtaisesti epäoikeudenmukaisella tavalla – päätös voi olla kontrafaktuaalisesti reilu, mutta epäoikeudenmukainen yhtä kaikki.

Lauri Lahikaisen mukaan eettisistä haasteista huolimatta kannuste rakentaa älyä, joka korvaa ihmistyötä joko kokonaisten työnkuvien tai yksittäisten tehtävien muodossa, on suuri. Jos korvaaminen on mahdollista tai taloudellisesti kannattavaa, sitä ainakin yritetään. Lahikainen näkee taloudellisten muutosten lisäksi muitakin seurauksia, mikäli tekoäly vie työt. Esimerkiksi yksilön mahdollisuudet rakentaa omaa identiteettiään työn piirissä ja saada sosiaalista tunnustusta heikkenevät.

Tehokkuuteen tähtäävä uusliberaali talousajattelu ja sen mahdolliset epäkohdat näkyvät Juho Rantalan tulkitsemana myös uudenlaisten talousinfrastruktuurien, desentralisoitujen autonomisten organisaatioiden (DAO) ja lohkoketjun toiminnassa. Tällaiset keskushallinnottomat ja hajautetusti toimivat järjestelmät on nähty tehokkuuden lisäämisen ohella keinoina tasa-arvoistaa ja valtaistaa yksilöitä, sillä ne mahdollistavat sopimuksien solmimisen vertaisten välillä automaattisesti ja anonyymisti. Käytännössä ne sisältävät kuitenkin epädemokraattisia piirteitä: DAO:ita ohjaavat säännöt eivät ole usein neuvoteltavissa, ja ne hyödyttävät käyttäjiä ja organisaatioita, joilla on ennestään valtaa joko järjestelmän teknisen ymmärryksen tai pääoman muodossa.

Tekoälyn uusi tuleminen on herätellyt perinteisiä filosofisia kysymyksiä toimijuudesta, tiedosta ja oikeudenmukaisuudesta. Muuttunut teknologinen konteksti kutsuu filosofeja sekä kriittiseen ajatteluun että vastaamaan uuden aikakauden sille asettamiin vaatimuksiin.

Viitteet

1 Ks. Silver ym. 2016. Syväoppimista hyödyntämällä AlphaGo-tekoälyjärjestelmä opetettiin pelaamaan Go-peliä, jossa AlphaGo voitti 99,8 % peleistä muita Go-peliä pelaavia järjestelmiä vastaan. AlphaGo päihitti myös 18-kertaisen Gon maailmanmestarin Lee Sedolin neljässä viidestä ottelusta.

2 Dreyfus & Dreyfus (1986) katsovat, että asiantuntijatieto on intuitiivista ja kontekstualisoitua sekä ilmenee lähes reaktionomaisena toimintana. Toisin kuin tietokoneet ja niiden suorittamat laskentaprosessit, asiantuntijatietoa ja -taitoa käyttävä henkilö ei toiminnassaan nojaa eksplisiittisiin sääntöihin ja niiden soveltamiseen, vaan tunnistaa sopivan lähestymistavan ongelmaan intuitiivisesti.

3 Searle (1980) esittää kiinalaisen huoneen ajatuskokeen avulla argumentin, jonka mukaan puhtaasti symboleita manipuloiva kone ei pysty ymmärtämään kieltä. Kritiikki kohdistuu erityisesti ”vahvan tekoälyn”, eli tietoisuuteen ja ymmärrykseen kykenevän koneälyn mahdollisuutta vastaan. Ajatuskokeessa Searle pyytää lukijaa kuvittelemaan huoneen, jonka sisällä on mies, joka ei ymmärrä kiinaa. Huoneeseen syötetään sisälle kiinankielisiä merkkejä, joille miehen tehtävänä on tuottaa sopivia kiinankielisiä vastineita hyödyntämällä hänelle englanniksi annettuja ohjeita. Huoneen ulkopuoliselle tarkkailijalle näyttäisi siltä, että huone antaa ymmärrettäviä vastauksia sille syötettyihin kiinankielisiin kysymyksiin. Searlen keskeinen argumentti on, että vaikka huone – vertauskuvana tietokoneelle – käyttäytyisi täydellisen ymmärrettävästi ja asianmukaisesti, ei se ymmärrä kieltä. Operoitujen symbolien merkityksiä ei voida johtaa niiden syntaksista.

4 Järjestelmä- ja robottivastaukset edustavat pikemminkin eri teoreetikoiden esittämien vasta-argumenttien joukkoja kuin yksittäisiä vastauksia. Yleisesti ottaen järjestelmävastausten keskeisenä ideana on, että Searlen kuvaama kiinalainen huone – kokonaisena järjestelmänä – ymmärtää kieltä, vaikka huoneessa sisällä oleva mies ei. Robottivastauksen eri muotoja esittäneiden teoreetikoiden mukaan Searlen pääargumentti voi pitää paikkansa. Mikäli järjestelmällä kuitenkin olisi kyky liikkua, havainnoida ja manipuloida ympäristöään, se voisi oppia ymmärtämään operoimiensa symbolien merkityksen. Argumenteista tarkemmin, ks. esim. Cole 2019, osiot 4.1. ja 4.2.

5 Ks. esim. Wilson & Foglia 2017.

6 Ks. esim. Barocas & Selbst 2016.

7 Mittelstadt ym. 2016, 6–7.

8 Wachter ym. 2016.

9 Kusner ym. 2017.

Kirjallisuus

Barocas, Solon & Selbst, Andrew D, Big Data's Disparate Impact. *California Law Review*. Vol. 104, 2016, 671.

Cole, David, The Chinese Room Argument. *The Stanford Encyclopedia of Philosophy* (2019). Toim. Edward N. Zalta. Verkossa: plato.stanford.edu/archives/spr2019/entries/chinese-room/

Dreyfus, Hubert L. & Dreyfus, Stuart, *Mind Over Machine. The Power of Human Intuition in the Era of the Computer*. Free Press, New York 1986.

Kusner, Matt, Loftus, Joshua, Russell, Chris & Silva, Ricardo, Counterfactual Fairness. *Advances in Neural Information Processing Systems*, 2017, 4066–4076.

Mittelstadt, Brent D. ym., The Ethics of Algorithms. Mapping the Debate. *Big Data & Society*. Vol. 3, No. 2, 2016, 1–21.

Searle, John, Minds, Brains, and Programs. *Behavioral and Brain Sciences*. Vol 3, No. 3, 1980, 417–424.

Silver, David ym., Mastering the Game of Go With Deep Neural Networks and Tree Search. *Nature.* Vol. 529, No. 7587, 2016, 484–489.

Wachter, Sandra, Mittelstadt, Brent D. & Russell, Chris, Counterfactual Explanations Without Opening the Black Box. Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*. Vol. 31, No. 2, 2018, 841–887.

Wilson, Robert & Foglia, Lucia, Embodied Cognition. *The Stanford Encyclopedia of Philosophy* (2017). Toim. Edward N. Zalta. Verkossa: plato.stanford.edu/archives/spr2017/entries/embodied-cognition/