Tekoäly ihmisen kognitiivisena avustajana: Kysymys tiedollisista riskeistä

Anna-Mari Rusanen
FT, yliopistonlehtori
kognitiotiede, digitaalisten ihmistieteiden osasto, Helsingin Yliopisto

Otto Lappi
Dos, yliopistonlehtori
kognitiotiede, digitaalisten ihmistieteiden osasto, Helsingin Yliopisto

> *"Tieto antaa mahdollisuuksia, mutta kaikki tieto ei ole luotettavaa eikä tietämisen illuusio johda hyviin päätöksiin"*
>
> *"Tiedon käyttäjien pitäisi pystyä arvioimaan tietojen luotettavuutta, erottelemaan, mikä on perusteltua tietoa…"*
>
> *"Ilman määrätietoista tietopolitiikkaa, jolla vahvistetaan kriittistä tieto-osaamista ja luotettavan tiedon tuottamista, voi ainoaksi vaihtoehdoksi jäädä, että tiedon aika on todellakin ohi…"*
>
> - *VVM: "Suomi tarvitsee tietopolitiikkaa" - julkaisu, 2017*

1. Johdanto

Nykyinen tekoäly koostuu lähinnä erikoistuneisiin tiedonkäsittelytehtäviin tarkoitetuista ohjelmistoista. Kehitys näyttää jatkuvan myös tulevaisuudessa samansuuntaisena. Nämä ohjelmistot eivät ole itsenäisiä kognitiivisia toimijoita, vaan lähinnä ihmistoimijan välineitä tai kognitiivisia avustajia. Niitä voidaankin tarkastella jatkeena ihmiskognition muisti-, tiedonkäsittely- ja ajatteluprosessien rajallisuutta kiertävien instrumenttien kehitykselle[1]. Se alkaa jo piirtämisen ja kirjoittamisen kehityksestä: Niiden avulla muistettavat sisällöt on voitu tallentaa ulkoisiksi representaatioiksi, joita on voitu työstää ja tallentaa tehokkaasti.

Tekoäly kuitenkin eroaa aiemmista kognitiivisista apuvälineistä siten, että se on kognitiivisesti organisoituneempi, kompleksisempi, oppivampi, usein osittain itsenäisempi ja dynaamisempi. Siten tekoäly laajentaa erilaisten ulkoisten tiedonkäsittelyn apuvälineiden, kuten visualisointimenetelmien tai piirtureiden kirjoa kognitiivisesti uudella tavalla, sekä tarjoaa mahdollisuuden ylittää kognitiivisen järjestelmämme rajalliset laskenta-, analyysi- ja päättelykyvyt. Toisaalta nykyiset, pitkälti erikoistuneet tekoälysovellukset eivät kykene samanlaiseen laaja-alaiseen tiedon integraatioon kuin ihmistoimijat. Siten ne

[1] Rusanen & Ylikoski 2017.

usein pikemminkin kaventavat ja rajoittavat ihmisen tiedonkäsittelyä kuin monipuolistavat sitä.

Tekoälypohjaiset instrumentit eivät siis pelkästään “ylitä”, “laajenna”, ”nopeuta” tai ”tehosta” tiedonkäsittelyä tai ajattelua vaan myös rajoittavat, muokkaavat ja joissakin tapauksissa jopa hidastavat sitä[2]. Tästä huolimatta tekoälykehitykseen sisältyy monia lupaavia piirteitä juuri sen kognitiivisesti monimutkaisemman rakenteen vuoksi. Toisaalta siihen liittyy myös joukko tiedollisia eli episteemisiä haasteita. Tarkastelemme tässä tekstissä vain muutamia niistä. Analysoimme ensin ihmistoimijan ja tekoälyn yhteistyöstä, yhteistyön taustalla olevasta tiedollisesta työnjaosta ja sen perustana olevasta kognitiivisesta vuorovaikutuksesta syntyviä tiedollisia haasteita. Lopuksi pohdimme viime vuosina paljon keskustelua herättäneitä ns. kognitiivisten vinoumien aiheuttamia ongelmia.

2. Ihmisen ja koneen kognitiivinen työnjako

Erityisesti tekoälyaikakauden alussa ihmisen ja koneen yhteistoiminnan kognitiiviseen organisointiin voi liittyä monenlaisia ongelmia. Tästä on jo nyt useita kokemuksia. Esimerkiksi Tesla, Volvo ja Rolls Royce ovat raportoineet autonomisiin ajoneuvoihin liittyvistä onnettomuustapauksista[3], joissa on ollut osittain kyse siitä, että ihmistoimija on arvioinut väärin tekoälypohjaisen järjestelmän toiminnan. (Toki on korostettava, että onnettomuuksista ei voida syyttää yksinomaan ihmisosapuolta, sillä usein niiden taustalla on myös tekoälyjärjestelmän heikkouksia.) Samoin paljon kohua herättäneissä tekoälyohjelmien tekemissä kyseenalaisissa luottopäätöksissä osa ongelmaa on se, että tekoälyjärjestelmien toimintatapaa ei ole riittävästi ymmärretty (tai niitä ei ole haluttu tehdä loppukäyttäjälle ymmärrettäviksi) ja siten ei ole osattu ennakoida niiden käyttöön liittyviä riskejä.

Mustat laatikot ja tiedon puute

Toisaalta näihin ongelmiin liittyy usein ns. mustien laatikkojen ongelma. Usein joko tekoälyn käyttäjällä ei ole riittävästi tekoälyjärjestelmän toimintaan liittyvää osaamista tai tekoälyn käyttäjällä ei ole edes mahdollisuutta saada tekoälyjärjestelmän toiminnasta riittävää osaamista (esim. yritysten liikesalaisuuksien vuoksi). Tällöin kyse on saatavilla olevan tiedon puutteesta, ja

[2] Tekoäly ja erikoistuneet tiedonkäsittelyongelmat, ks. Lappi, Rusanen & Pekkanen 2018.
[3] https://www.theregister.co.uk/2017/06/20/tesla_death_crash_accident_report_ntsb/
http://www.dsf.my/2017/06/volvo-cars-explains-the-accident-video-circulating-on-social-media/
https://techcrunch.com/2018/02/01/we-were-in-an-accident-during-an-automated-driving-tech-demo/

se on ratkaistavissa esim. riittävällä koulutuksella ja luomalla yrityksille paine avata liikesalaisuuksia riittävällä tasolla.

Toisaalta on nostettu esiin myös tekoälytutkimuksen sisäisten mustien laatikkojen ongelmat. Esimerkiksi hiljattain Ali Rahimi ja Ben Recht[4] vertasivat tekoälyohjelmointia "alkemiaan". Rahimi ja Recht painottivat, kuinka tekoälyalgoritmien kehityksessä joudutaan toisinaan korjaamaan esimerkiksi DL- arkkitehtuurien algoritmeja tavoilla, joiden vaikutuksia edes mallintajat itse eivät pysty analysoimaan. Näissä tilanteissa tekoäly ei ole musta laatikko vain sen käyttäjille vaan myös sen kehittäjille. Ongelma ei kuitenkaan ole se, etteikö tietoa järjestelmästä olisi saatavissa, vaan - kuten Rahimi ja Recht painottavat- se, että koneoppimiseen ja algoritmien kehittämiseen liittyvien analyyttisten työkalujen puutteesta. On huomattava, että tämäntyyppisiä ongelmia ei ratkota pelkästään lisäämällä koulutusta, tai velvoittamalla kertomaan tekoälyjärjestelmien toiminnasta. Tällaisten ongelmien ratkaisu vaatii myös tutkimusta, tutkimuksen menetelmien tarkastelua ja ehkä jopa tekoälysovellusten kehittämiseen liittyvien eettisten ohjeistusten laatimista.

Tekoäly, ihmiset ja kognitiiviset vinoumat

Viime vuosina on keskusteltu paljon myös siitä, kuinka tekoäly ihmisen työparina monimutkaistaa kysymystä kognitiivisista vinoumista. Kognitiivinen vinouma tarkoittaa kognitiivisen järjestelmän systemaattista tapaa painottaa tiedonkäsittelyn kannalta tiettyjä, usein virheellisiä piirteitä[5]. Negatiivisesta nimestään huolimatta vinoumilla on myös hyötyarvoa. Kognitiivisesta näkökulmasta ne nimittäin toimivat peukalosääntöinä, jotka nopeuttavat ja helpottavat päätöksentekoa monimutkaisissa tilanteissa. Toisaalta, ne usein samalla vääristävät päätöksentekoa systemaattisella tavalla. Tämä usein johtaa virheellisiin tulkintoihin, vääriin päätöksiin ja irrationaalisiin valintoihin.

Tekoäly-ihminen- työparin osana tekoäly voi toimia siten, että se poistaa tai estää systemaattisia vinoumia kollektiivisten rakenteiden tasolla esimerkiksi toimimalla objektiivisesti ja johdonmukaisesti tilanteissa, joissa ihminen puolestaan on epäjohdonmukainen ja subjektiivinen. Toisaalta tekoäly saattaa vahvistaa jo olemassaolevia vinoumia, ja joissain tapauksessa myös tuottaa uusia vinoumien tyyppejä.

Eri vinoumilla on erilaisia vaikutuksia, ja myös niiden analysointi vaatii usein erilaisia käsitteitä. Siksi onkin syytä erotella tekoälyyn mahdollisesti liittyvien

[4] http://www.argmin.net/2017/12/05/kitchen-sinks/

[5] Vinoumia on useita, ehkä jopa kymmeniä. Ks. esim. Gilovich, Griffin & Kahneman 2002: Heuristics and biases: The psychology of intuitive judgment. Cambridge, UK: Cambridge University Press.

kognitiivisten vinoumien tyypit esimerkiksi seuraavalla tavalla (Rusanen ja Koskinen, valmisteilla):

(i) datan vinoumat
(ii) algoritmien aiheuttamat vinoumat
(iii) tekoälyarkkitehtuurien rakenteelliset kognitiiviset vinoumat
(iv) tutkimuksen aiheuttamat vinoumat

Julkisuudessa on eniten keskusteltu lähinnä (i) dataan liittyvistä vinoumista. Niillä viitataan usein joko tekoälysovelluksen harjoittamisessa käytetyn datan vinoumien siirtymiseen tai siihen, kuinka tekoälysovellus ei tunnista datasta vinoutuneita tai painottuneita rakenteita. Nämä vinoumat eivät kuitenkaan varsinaisesti ole tekoälysovelluksen, vaan siihen syötetyn datan aiheuttamia.

Sen sijaan (ii) tekoälyalgoritmit, tai (iii) tekoälyjärjestelmä kokonaisuudessa ("kognitiivinen arkkitehtuuri") saattavat tuottaa rakenteellisia kognitiivisia vääristymiä. Tutkimusta aiheesta ei juurikaan ole, mutta jo nyt epäillään, että tietyt koneoppimisalgoritmit tuottavat vahvistusvinouman kaltaisia tiedonkäsittelyn vääristymiä[6]. Samoin on mahdollista, että tekoälyalgoritmien korjailuun käytettävät satunnaistamismenetelmät saattavat vaikuttaa ennakoimattomilla tavoilla algoritmien toimintaan. Toisaalta on alettu myös miettiä sitä, missä määrin monimutkaiset tekoälyarkkitehtuurit tuottavat uudenlaisia kognitiivisia vinoumia. Viime aikoina on alettu myös keskustella siitä, missä määrin (iv) itse koneoppimisalgoritmeja tai tekoälyarkkitehtuureja tulkitaan kognitiivisesti vinoutuneesti[7].

3. Lopuksi

Tekoäly ihmisen työparina avaa uusia mahdollisuuksia niin tieteelliselle tutkimukselle, teknologiselle kehitykselle, yhteiskunnalliselle hyvinvoinnille ja ekologisesti kestävämmälle kehitykselle. On kuitenkin huomattava, että näiden mahdollisuuksien lisäksi tekoälyn käyttöön sisältyy useita tiedollisia eli episteemisiä haasteita. Näitä haasteita ovat mm. ihmisen ja koneen kognitiivisen vuorovaikutuksen haasteet, tekoälyn käyttöön ja sen kehittämiseen liittyvät erilaiset mustat laatikot, sekä ihmis-kone-hybridin kognitiiviset vinoumat.

Näihin haasteisiin reagoiminen ennaltaehkäisevällä tavalla vaatii sitä, että tekoälysovelluksia tarkastellaan myös ihmis-kone- vuorovaikutuksen

[6] Vahvistusvinouma (engl. "confirmation bias") on eräs tyypillisimmistä kognitiivisista vääristymistä. Se syntyy kognitiivisen järjestelmän taipumuksesta painottaa jo olemassaolevia uskomuksia tukevaa todistusaineistoa siten ("etsi vain sellaista todistusaineistoa, joka osoittaa, että olet oikeassa").
[7] Kliegr, Bahnik & Fürnkranz, 2018. https://arxiv.org/pdf/1804.02969.pdf

näkökulmasta. Ilman monipuolista käyttäytymistieteellistä tutkimusta ihmisen ja koneen kognitiivisesta työnjaosta, sen periaatteista tai reunaehdoista ei ole mahdollista ratkoa tällaisia episteemisiä haasteita, eikä myöskään ole mahdollista kehittää ihmisen kognitiota tukevaa tekoälyä ilman, että ymmärretään riittävästi ihmisen ja koneen välistä vuorovaikutusta.

Lisäksi tulevaisuudessa on edelleen kehitettävä tarkoituksenmukaisia auditointimenetelmiä eri tyyppisille tekoälysovelluksille. Tekoälyä on useaa erilaista, ja välttämättä ei ole yhtä oikeaa auditointimenetelmää, joka sopisi kaikille eri algoritmien tai arkkitehtuurien perheille. Myös tekoäly-yhteisöä tulisi kannustaa käymään laajempaa keskustelua Rahimin ja Rechtin tavalla tekoälysovelluksien kehittämiseen käytettävistä menetelmistä sekä tekoälysovellusten kehittämisen eettisestä ohjeistuksesta.