HENRIK RYDENFELT, JAAKKO KUORIKOSKI
JA SALLA-MAARIA LAAKSONEN

TEKOÄLY, TUTKIMUKSEN LUOTETTAVUUS JA TIEDOLLINEN VASTUU

KUVAN LÄHDE: IUNSPLASH

Tieteen, tutkimuksen ja tutkijoiden näkökulmasta tekoäly ei ole uusi aihe. Tekoälyä tai sen esiasteita on käytetty tutkimuksessa eri tavoin jo vuosikymmenten ajan. Suurten kielimallien nopea kehitys kuitenkin haastaa myös vakiintuneita tieteellisiä käytäntöjä.

Ennen vuoden 2022 viimeisiä päiviä sanalla ”tekoäly” viitattiin arkisessa käytössä tyypillisesti ihmismäiseen päättelyyn ja toimintaan kykenevään teknologiaan, joka voisi kenties olla käsillä joskus tulevaisuudessa. Viimeisen kahden vuoden aikana sana on vakiintunut viittaamaan erityisesti suuriin kielimalleihin, joihin pohjautuvat sovellukset ovat tulleet kaikkien käyttöön (englanniksi *large language models*, LLM).

Yhtäältä kyse on teknologioiden ja tuotteiden onnistuneesta markkinoinnista. Toisaalta tekoälyn kehitys näyttää tulleen murroskohtaan, jossa inhimillisestä näkökulmasta tekoälyjärjestelmien tuottamat sisällöt, kuten tekstit ja kuvat, eivät aina ole erotettavissa ihmisen tuotoksista. Vakiintuneita tieteellisiä käytäntöjä haastaa erityisesti niin kutsuttu syväoppiminen, joka hyödyntää monikerroksisia neuroverkkoja monimutkaisten ilmiöiden mallintamisessa. Syväoppiminen johtaa kysymyksiin tekoälyä hyödyntävän tutkimuksen ja sen tulosten luotettavuudesta sekä tiedollisesta vastuusta.

TEKOÄLYN EPISTEEMISET, EETTISET JA YHTEISKUNNALLISET HAASTEET

Tekoälyyn ja algoritmisiin teknologioihin liittyvistä eettisistä ongelmista on viime vuosina kehkeytynyt valtava tutkimuskirjallisuus, jossa erilaisia huolenaiheita on eritelty ja kategorisoitu (esimerkiksi Mittelstadt ym. 2016; Tsamados ym. 2021; Lindgren 2023). Osaa tekoälyyn kytketyistä ongelmista voidaan pitää ensi sijassa episteemisinä eli tiedollisina. Ne korostuvat varsinkin silloin,

Tekoälyjärjestelmien kehittäminen vaatii valtavia taloudellisia panostuksia, minkä vuoksi edistyneimpien järjestelmien hallinta näyttää ainakin toistaiseksi keskittyvän rajatulle joukolle kansainvälisiä suuryrityksiä.

kun tekoäly toimii välineenä informaation tuottamisessa.

Ensinnäkin suuria kielimalleja hyödyntävien generatiivisten tekoälysovellusten toiminta perustuu tilastollisiin malleihin, todennäköisyyksiin ja osin myös satunnaisuuteen, mikä tekee niiden toiminnasta perustavanlaatuisesti erilaista verrattuna ihmisen päättelyyn. Generatiivinen tekoäly ”tunnistaa” tilastollisia yhteyksiä ja kykenee lajittelemaan aineistoja, mutta se ei ”ymmärrä” sisältöä samalla tavalla kuin ihminen, eikä sen tuotoksia voida rinnastaa ihmisen päättelyn tai järkeilyn tuloksiin.

Toiseksi koneoppiminen perustuu dataan. Generatiivisen tekoälyn tuotosten laatu heijastelee datassa – ja sen keskeisenä lähteenä toimivissa ihmisissä ja yhteiskunnissa – esiintyviä virheitä, harhoja ja puutteita (esimerkiksi Vallor 2024). Jos koneoppimisessa hyödynnetty data on vinoutunutta, järjestelmien sovellettavuus ja luotettavuus heikkenee. Ongelma korostuu koneoppivien järjestelmien tutkimuskäytössä, jossa datan vinoutumat voivat vääristää mallien tuottamia tuloksia.

Rajatummassa mielessä eettiset ongelmat kytkeytyvät yhtäältä tekoälyn hyödyntämiseen päätöksenteossa. Tekoäly voi syrjäyttää ihmisen päätöksentekijänä tai tehdä ratkaisuja ihmisten puolesta, mikä on herättänyt huolia ihmisen autonomian tai toiminnan ja päätöksenteon vapauden toteutumista. Edelleen tekoälyn tekemät päätökset voivat olla sisällöltään tai vaikutuksiltaan eettisesti ongelmallisia, esimerkiksi epäoikeudenmukaisia tai syrjiviä.

Toiset eettiset ongelmakohdat liittyvät valtaviin datamääriin, joita algoritmiset teknologiat ja erityisesti suuret kielimallit hyödyntävät. Niiden toimintatavat ovat herättäneet uusia vaatimuksia tiedon, informaation ja datan yksityisyydestä. Yksilöiden toiminnasta ja laajemmin ihmiselämästä kerätty data mahdollistaa päätelmiä yksilöiden ja ryhmien ominaisuuksista, kuten haluista ja tarpeista.

Tutkimuksessa on esitetty tapoja, joilla algoritmiset järjestelmät ja tekoäly voivat käyttää hyväkseen ihmisten heikkouksia ja haavoittuvuuksia hyödyntäen niitä esimerkiksi kaupallisiin tai poliittisiin tarkoituksiin tai jopa vaarantaen yksilöiden kyvyn tehdä valintoja, jotka ovat perusteltuja ja itsenäisiä (esimerkiksi Eubanks 2018). Nämä ongelmat erottuvat erityisesti informaatioympäristössä ja median sisältöjen kulutuksessa, joita algoritmiset teknologiat ja koneoppiminen ovat jo pitkään muovanneet perusteellisesti (Rydenfelt ym. 2024; katso Rydenfelt ym. 2025).

Myös tekoälyteknologioiden laajat yhteiskunnalliset ja poliittiset vaikutukset ovat herättäneet kriittistä keskustelua. Tekoälyjärjestelmien kehittäminen vaatii valtavia taloudellisia panostuksia, minkä vuoksi edistyneimpien järjestelmien hallinta näyttää ainakin toistaiseksi keskittyvän rajatulle joukolle kansainvälisiä suuryrityksiä (van der Vlist ym. 2024; Kak ym. 2023). Tekoälyteknologioiden tuottama taloudellinen hyöty voi jakautua hyvin epätasaisesti.

Hallitsemansa teknologian avulla nämä yritykset voivat käyttää entistä suurempaa valtaa vaikuttamalla kaikkiin inhimillisen elämän aloihin, mukaan lukien poliittisiin ja demokraattisiin prosesseihin. Tekoäly-yritysten ja -sovellusten ympäristövaikutukset ovat myös huomattavia (Crawford 2024; Bender ym. 2021; Dhar 2020). Suurten kielimallien kouluttaminen vaatii massiivisesti laskentatehoa, ja myös niiden käyttö kuluttaa huomattavia määriä energiaa.

Monet edellä eriteltyihin tekoälyn episteemisiin ja eettisiin ongelmiin tarjotut ratkaisuehdotukset pohjautuvat läpinäkyvyyden ja tilivelvollisuuden käsitteille (englanniksi *transparency* ja *accountability*) (esimerkiksi Felzmann ym. 2020; Powell 2021; Novelli ym. 2024). Lähestymistavan lähtökohdat ovat helposti ymmärrettäviä.

Kielimallien toiminnan läpinäkyvyydellä voisi olla vähintäänkin välinearvoa: ne voisivat edistää episteemisten ja eettisten ongelmien ehkäisyä ja ratkaisua muun muassa kasvattamalla tekoälyjärjestelmien kontrolloitavuutta ja auttamalla havaitsemaan virheitä. Tekoälyn läpinäkyvyydestä ei kuitenkaan ole esitetty järin selkeää määritelmää, ja sen käytännön toteutukseen ei ole kehitetty selkeitä malleja (Rydenfelt ym. 2021; Powell 2021).

Koneoppimiseen perustuvat tekoälyjärjestelmät ovat usein niin monimutkaisia, että niiden toiminnan selittäminen vaikuttaa jopa mahdottomalta. Pahimmillaan keskittyminen läpinäkyvyyteen vie huomiota pois varsinaisista episteemisistä ja eettisistä ongelmista, joita sen avulla oli tarkoitus välttää.

Euroopan unionin dataa ja tekoälyä koskevassa sääntelyssä keskeiseksi noussut tilivelvollisuuden käsite on sekin monin tavoin jäänyt sisällöltään epäselväksi (Novelli ym. 2024). Kenelle ja millä tavoin tekoälyjärjestelmien kehittäjät ja käyttäjät tekisivät tekemisistään tiliä?

Tekoälyn hyödyntämiseen, kuten kaikkeen inhimilliseen toimintaan, liittyy eettinen vastuu. Ajatus (erillisestä) tilinteon mekanismista voi pahimmillaan pyrkiä korvaamaan tämän

vastuun. Läpinäkyvyyteen ja vastuuseen liittyvät kysymykset ovat tästä huolimatta keskeisiä myös tekoälyn tieteellisessä käytössä.

Kuten seuraavassa esitämme, monet tekoälyavusteisen tutkimuksen ja sen tulosten luotettavuuden arvioinnin ongelmat liittyvät läpinäkyvyyden puutteeseen. Myös tiedollista vastuuta koskevat käytännöt on jäsennettävä uudestaan tilanteessa, jossa koneet suorittavat yhä suuremman osan tutkijoiden tieteellisestä työstä.

GENERATIIVINEN TEKOÄLY JA TUTKIMUKSEN LUOTETTAVUUS

Tekoälypohjaisia menetelmiä on määritelmästä riippuen hyödynnetty tutkimusaineistojen analysoinnissa jo pitkään. Erilaiset koneoppimismenetelmät ovat olleet keskeisessä roolissa niin numeerisen aineiston kuin laajojen tekstimassojenkin analyysissä.

Perinteiset tilastolliset menetelmät, kuten regressioanalyysi tai dimensionaalisuuden analysointiin kehitetyt menetelmät, voidaan nähdä koneoppimisen läheisinä sukulaisina. Aiemmin käytössä olleet työvälineet ovat kuitenkin tyypillisesti edustaneet niin sanottua kapeaa tekoälyä, joka on suunniteltu suorittamaan tehtäviä, jotka ovat tarkasti rajattuja ja ennalta määriteltyjä (katso Sajja 2021).

Uudemmat tekoälysovellukset, erityisesti generatiivinen tekoäly, lähestyvät niin kutsuttua yleistä tai laajaa tekoälyä ainakin käyttökokemuksensa puolesta. Yleinen tekoäly kykenee toimimaan erilaisissa konteksteissa ja ratkaisemaan monenlaisia ongelmia (Sajja 2021).

Uusien tekoälysovellusten kehityksen haittapuolena on, että ymmärrämme merkittävästi rajallisemmin sitä, miten sovellukset toimivat ja hyödyntävät dataa. Näiden järjestelmien hyödyntäminen vaikeuttaa olennaisesti tutkimuksen luotettavuuden arviointia.

Tilastollisen eli määrällisen tutkimuksen luotettavuutta arvioidaan tyypillisesti reliabiliteetin ja validiteetin käsitteiden avulla. Reliabiliteetilla viitataan mittauksen johdonmukaisuuteen. Sen arviointi perustuu usein tulosten pysyvyyteen eli stabiliteettiin, jota arvioidaan mittaustulosten toistettavuuden kautta. Tällöin tarkastellaan esimerkiksi sitä, kuinka samankaltaisia tuloksia saadaan toistettaessa samaa mittausta eri ajankohtana tai eri tutkijoiden toteuttamana. Toiseksi reliabiliteetin ulottuvuudeksi kuvataan tyypillisesti sisäinen johdonmukaisuus eli konsistenssi, jota voidaan arvioida esimerkiksi tarkastelemalla, mittaavatko tutkimusvälineen, mittarin tai kokeen eri osat samaa asiaa.

Validiteetti eli pätevyys puolestaan viittaa siihen, kuinka hyvin tutkimuksessa käytetty mittausmenetelmä mittaa juuri sitä ilmiötä tai ominaisuutta, jota tutkimuksessa pyritään selvittämään. Ulkoinen validiteetti koskee tutkimustulosten yleistettävyyttä (tai väljemmin soveltuvuutta) muihin vastaaviin tilanteisiin. Sisällöllisen validiteetin termein tarkastellaan sitä, kuinka tarkasti mittari kattaa tutkimuksen kohteena olevan ilmiön ja sen osa-alueet. Konstruktiovaliditeetin tarkastelussa taas pyritään arvioimaan, kuinka hyvin mittaus vastaa teoreettisesti määriteltyä ilmiötä ja kuinka sopivia käytetyt käsitteet ovat tutkimuksen kohteeseen ja kontekstiin.

Suurten kielimallien toiminta perustuu osin satunnaisuuteen. Vaihtelu voi johtua myös järjestelmien päivityksistä ja vikatiloista.

Laadullisessa tutkimuksessa tilastollisia menetelmiä ei yleensä hyödynnetä, ja validiteetin ja reliabiliteetin käsitteitä ei sellaisenaan sovelleta tutkimuksen laadun tai luotettavuuden arviointiin. Laadullisen tutkimuksen luotettavuus perustuu paljolti tutkijan suorittaman tulkinnan subjektiivisen elementin hallintaan, jotta tulokset eivät heijastele tutkijan omia näkemyksiä ja ennakko-oletuksia. Voidaan esimerkiksi tarkastella, päätyisivätkö muut tutkijat vastaaviin tuloksiin käyttäen samoja menetelmiä tai soveltaen valittua lähestymistapaa muihin samankaltaisiin aineistoihin.

Validiteettia muistuttavana tekijänä laadullisessa tutkimuksessa voidaan tarkastella sitä, miten käytetyt käsitteet ja menetelmät sopivat tutkittavaan ilmiöön ja miten käytetyt aineistot soveltuvat vastaamaan tutkimuskysymyksiin. Toisaalta laadullisen tutkimuksen luotettavuutta kasvattaa tutkijan kyky tunnistaa, selkeästi ilmaista ja reflektoida tutkimuksen ja sen tulosten subjektiivisia elementtejä.

TEKOÄLY JA RELIABILITEETTI

Tekoälysovellusten tuottamien tulosten reliabiliteetin arvioinnissa yksi keskeinen ongelma liittyy tulosten toistettavuuteen. Tällä hetkellä generatiiviset tekoälysovellukset ovat yhä useammin tutkijoiden käytettävissä helppokäyttöisinä selainsovelluksina tai teknologiayritysten tarjoamien tuotteiden lisäpalveluina. Samasta syystä järjestelmät päivittyvät ja muuttuvat tavoilla ja hetkillä, joita tutkijat eivät kontrolloi.

Suurten kielimallien tuotokset voivat vaihdella ennalta-arvaamattomasti eri mallien vä-

Tiede ei ainoastaan tuota ja ennusta uusia havaintoja tai kerro, millainen maailma on – tieteen pitäisi myös kertoa, miksi maailma on sellainen kuin se on.

lillä. Myös hyödyntämällä samaa mallia voi saada eri tuloksia, jos sitä käyttää eri tavoin tai eri kerralla. Suurten kielimallien toiminta perustuu osin satunnaisuuteen. Vaihtelu voi johtua myös järjestelmien päivityksistä ja vikatiloista.

Toinen keskeinen haaste liittyy kielimallien tuottamien tuotosten laatuun. Suurten kielimallien toimintakyky perustuu laajaan koulutusaineistoon, jonka avulla ne käsittelevät uutta tietoa muuntamalla sen matemaattisiksi esityksiksi. Kielimallin kouluttamiseen käytetyn aineiston sisältö vaikuttaa ratkaisevasti mallin tuottamiin lopputuloksiin (esimerkiksi Bender ym. 2021).

Mallin tuottaman analyysin laatu riippuu pitkälti koulutusaineistosta ja sen edustavuudesta. Tekoälyn avustama tai tuottama analyysi ei ole automaattisesti ihmisen tekemää objektiivisempaa. Esimerkiksi opetusaineiston sisältämät vinoumat voivat toistua tuloksissa vääristävinä stereotypioina.

Täysin tasapuolisen ja vinoumista vapaan opetusaineiston kokoaminen on puolestaan käytännössä mahdotonta. Kyse on valtavista, tällä hetkellä terabittien kokoisista tekstimassoista, jotka ovat usein puutteellisesti dokumentoituja (Bender ym. 2021). Vinoumien vähentämiseksi teknologiayritykset ovat kouluttaneet kielimallejaan ihmiskouluttajien avulla. Tämäkään koulutusprosessi ei kuitenkaan ole arvovapaa: se heijastelee sekä ihmiskouluttajien näkemyksiä että koulutuksessa hyödynnettyjä ohjeita ja arvoja, joista teknologioita kehittävät yritykset linjaavat.

Näiden näkökohtien pohjalta on mahdollista myös luonnehtia tapoja, joilla generatiivisen tekoälyn avulla tuotettujen tulosten reliabiliteettia voidaan tukea ja edistää. Ulkoisten palveluntarjoajien tuottamien sovellusten tarjoamien tulosten reliabiliteettia voi tällä hetkellä pyrkiä kasvattamaan lähinnä siten, että suunnittelee ja muotoilee huolellisestitekoälylle annetut kehotteet (englanniksi *prompt*). Lisäksi tulosten arvioinnissa voidaan hyödyntää ihmiskoodauksen samanmielisyyttä varten kehitettyjä mittareita (esimerkiksi Krippendorf 2011). Pienten aineistojen sekä monen laadullisen tutkimuksen kohdalla tutkijan on myös mahdollista rajata tekoälyn käyttö aineistoa koskevaan ”keskusteluun” tekoälyn kanssa analyysin helpottamiseksi.

Avointen kielimallien kohdalla reliabiliteetin kontrollointi voi entistä enemmän perustua siihen, että käytetty kielimalli on vakaa ja tutkijoiden itsensä hallittavissa, jolloin vikatilat, päivitykset ja muutokset eivät heikennä tuloksia tai tuota yllätyksiä. Tietyt ja rajatut tekoälypohjaiset työkalut voivat vakiintua tutkijoiden käyttöön siinä määrin, että niiden tuottamien tulosten reliabiliteetista – sekä sen mahdollisista rajoista – alkaa muodostua selkeää näyttöä ja vakiintuneita tutkimusprotokollia.

TEKOÄLY JA VALIDITEETTI

Tekoälyn hyödyntäminen herättää vielä merkittävämpiä kysymyksiä validiteetin arvioinnin kohdalla. Pienten aineistojen analyysin kohdalla tutkijan on mahdollista hahmottaa ja arvioida, miten hyvin tulokset tavoittavat tutkimuksen kohteena olevan ilmiön. Samoin monen kvalitatiivisen tutkimuksen kohdalla tutkija kykenee usein arvioimaan, miten hyvin tulokset suhteutuvat aineistoon ja teoriaan sekä teorian tarjoamiin käsitteisiin. Suuriin aineistoihin ja niihin perustuvaan koneoppimiseen pohjaavien tulosten validiteetin arviointi samaan tapaan on kuitenkin vaikeaa tai mahdotonta.

Esimerkiksi syväoppivat hermoverkkomallit kykenevät usein löytämään moniulotteisista ja laajoista havaintoaineistoista hämmästyttävän ennustekykyisiä muuttujien yhdistelmiä. Käyttäjä tai rakentaja ei kuitenkaan pysty täysin hahmottamaan tai ymmärtämään, millä tavoin säännönmukaisuuksiin on päädytty, saati tarkastamaan niitä mittavasta aineistosta. Vaikka algoritmiset periaatteet voivat olla pohjimmiltaan yksinkertaisia, käyttäjä ei voi jäljittää tarkasti, miten malli on tuottanut tietyn tuloksen. Tämä vaikeuttaa tulosten tulkintaa ja yleistettävyyden arviointia.

Tätäkin vaikeammaksi – suorastaan mahdottomaksi – voi muodostua sisällöllisen validiteetin ja konstruktiovaliditeetin arvioiminen. Tutkija ei useinkaan voi hahmottaa, kattavatko tulokset tutkimuksen kohteena olevan ilmiön eri osa-alueet. Syväoppivan mallin ei voi olettaa vastaavan sisällöllisesti tutkimuksen kohteena olevaa ilmiötä. Tällainen malli ei perustu ilmiötä koskevaan teoriaan, eikä sen sisäinen rakenne näin itsessään kuvaa mallinnettavan kohteen rakennetta.

Teoriapohjaisilla malleilla on tutkimuksessa keskeinen rooli juuri siksi, että niiden rakenteen ajatellaan heijastavan mallinnettavan ilmiön mekanismeja. Tällöin ilmiön ymmärrys nojaa siihen, että mallin sisäiset toimintaperiaatteet vastaavat ilmiön sisäistä logiikkaa. Koneoppimisen seurauksena syn-

tyneet mallit eivät kuitenkaan ”toimi” samalla tavalla kuin mallinnettavat ilmiöt.

Koneoppiminen ratkaisee ennuste- ja luokitustehtäviä, eikä se pyri löytämään havaintoaineistosta syy-seuraussuhteita. Merkittävä seuraus tästä on, että syväoppimiseen perustuvat mallit eivät ole selittäviä. Yleisesti hyväksytyn näkemyksen mukaan suuri osa tieteellisestä selittämisestä perustuu syy-seuraussuhteiden paljastamiseen ja havaittavia säännönmukaisuuksia tuottavien mekanismien tunnistamiseen (esimerkiksi Woodward 2003).

Selittäminen taas on mielletty tieteen perustehtäväksi. Tiede ei ainoastaan tuota ja ennusta uusia havaintoja tai kerro, millainen maailma on – tieteen pitäisi myös kertoa, miksi maailma on sellainen kuin se on. Ratkaisuksi ongelmaan on esitetty selitettävää tekoälyä (*explainable AI*, XAI) tai siihen tähtääviä teknologioita (esim Zednik ja Boelsen 2022).

Näissä menetelmissä koneoppimismallin toimintaa pyritään kuvaamaan uudella, yksinkertaisemmalla ja näin (ihmiselle) ymmärrettävämmällä mallilla. Tällainen ”selittävä” malli palauttaisi algoritmien tuottamat tulokset inhimillisesti ymmärrettävään muotoon, jotta voisimme hahmottaa, miten ja miksi algoritmi tuottaa ennusteensa havaintoaineiston perusteella. Ei kuitenkaan ole takeita siitä, että selityksellinen tekoäly ratkaisisi ongelman ainakaan täysin ja kaikkien tulosten kohdalla.

Edelleen voidaan kysyä, tulisiko yksinkertaisempia mutta mahdollisesti vähemmän tehokkaita menetelmiä suosia tilanteissa, joissa läpinäkymättömät mutta tehokkaammat menetelmät ovat käytettävissä. Toinen ratkaisu on lisätä koneoppiviin malleihin kausaalisen päättelyn algoritmisia periaatteita (Buljsman 2023). Tämä lähestymistapa voisi yhdistää koneoppimisen ennustekyvyn ja kausaalisuuden tarjoaman selityksellisyyden, jolloin syvempää ymmärrystä voitaisiin saavuttaa heikentämättä menetelmien tehokkuutta.

LÄPINÄKYVYYS JA TIETEELLINEN YMMÄRRYS

Tekoälyyn perustuvat järjestelmät tuottavat yhä enemmän tutkimuksen tuloksia tavoilla, jotka ovat niiden käyttäjille vaikeita tai mahdottomia käsittää. Tämä tiedollinen läpinäkymättömyys asettaa haasteita koneoppimisella tuotetun tiedon ymmärrettävyydelle, mikä edellä kuvatulla tavalla puolestaan vaikeuttaa tulosten luotettavuuden arvioimista.

Läpinäkymättömyyden haastetta on kuitenkin helppo myös liioitella. Tiedollisen läpinäkymättömyyden ja tieteellisen ymmärryksen välistä suhdetta arvioitaessa on olennaista erottaa subjektiivinen ymmärryksen tunne varsinaisesta tieteellisestä ymmärryksestä.

Merkittävä osa keskustelusta tiedollisesta läpinäkymättömyydestä tieteessä ja tekoälyn filosofiassa perustuu ongelmalliseen taustaoletukseen: sen mukaan yksittäisen tutkijan psykologiset tilat, kuten ymmärryksen tunne, olisivat olennaisia tieteellisen ymmärryksen kasvulle tai tulosten arvioinnille.

Kokemuksellinen ahaa-elämys tai introspektiivinen arvio ymmärryksen syvyydestä ei välttämättä kuvasta todellista ymmärtämistä. Tieteen resurssien ei myöskään ole tarkoitus palvella pelkästään tutkijoiden henkilökohtaisen onnistumisen kokemuksia, vaan niiden hyödyntämisen tavoitteena on tieteellisen tiedon ja ymmärryksen kollektiivinen ja kumulatiivinen lisääntyminen.

Filosofisessa keskustelussa tieteellisen ymmärryksen luonteesta ei olla saavutettu yksimielisyyttä. Yleisesti kuitenkin katsotaan, että tieteellinen ymmärtäminen perustuu selitykselliseen tietoon ja kykyyn soveltaa tätä tietoa monipuolisesti (esimerkiksi Kuorikoski 2022). Tämä ei tarkoita pelkästään faktojen muistamista vaan kykyä käyttää tietoa uusien ongelmien ratkaisemiseen.

Tieteellinen ymmärrys voidaan näin mieltää sosiaalisesti jaettuna kykynä hyödyntää tietoa, joka on järjestelmällisesti tuotettua ja julkista. Tällainen ymmärrys kasvaa selityksellisen tiedon ja sen mahdollistamien kykyjen lisääntyessä. Yksittäisten tutkijoiden psykologiset tilat ovat tässä toissijaisia.

TIEDOLLINEN VASTUU JA SEN JAKAUTUMINEN

Tekoälyavusteinen tiede tulee vääjäämättä yleistymään. Tämä johtaa muutoksiin tieteellisen tiedon tuottamisessa ja sen organisoinnissa. Kun yhä suurempi osa tiedosta tuotetaan tekoälyn ja tutkijoiden yhteistyönä, tieteellisen tiedon ja ymmärryksen todelliseksi subjektiksi voidaan kuvata ihmisten ja algoritmien muodostama hybriditoimija (Kuorikoski ja Ylikoski 2015). Kehitys tuo mukanaan kysymyksen tällaisten toimijoiden tiedollisesta vastuusta ja sen jakautumisesta.

Kun yhä suurempi osa tiedosta tuotetaan tekoälyn ja tutkijoiden yhteistyönä, tieteellisen tiedon ja ymmärryksen todelliseksi subjektiksi voidaan kuvata ihmisten ja algoritmien muodostama hybriditoimija.

Nykyiset tieteen sosiaaliset käytännöt on rakennettu sille oletukselle, että tiedollinen vastuu kuuluu viime kädessä ihmisille. Vastuun tulosten luotettavuudesta on mielletty kuuluvan tutkijoille, joiden nimet löytyvät julkaistujen artikkeleiden otsikoiden alta. Koneoppimismenetelmien tiedollinen läpinäkymättömyys kuitenkin haastaa tämän periaatteen. Jos tutkija ei voi täysin ymmärtää, miksi käytetty menetelmä tuotti tietyt tulokset, voidaanko häntä pitää täysin vastuullisena tulosten esittämisestä ja niiden luotettavuudesta?

Tiedollisen vastuun sälyttäminen koneille ei liene mahdollista. Emme pidä koneita eettisesti vastuullisina toimijoina. On vaikea kuvitella, miten konetta voitaisiin esimerkiksi rankaista sen tekemistä virheistä. Myös tulosten merkityksen hahmottaminen inhimillisten käytäntöjen kannalta kuuluu ainakin toistaiseksi lähinnä hybriditoimijan ihmiskomponentille. Päävastuu tutkimuksesta ja tieteen tuloksista säilynee näin jatkossakin ihmisillä.

Tiedollinen vastuu ei kuitenkaan lepää vain yksittäisten tutkijoiden harteilla. Pikemminkin hybriditoimijuus korostaa tutkimuksen ja tieteen yhteisöllisyyttä. Tieteellinen työ on aina perustunut kollektiiviseen toimintaan, jossa yksittäinen tutkija nojaa muiden saavutuksiin ja rakentaa niille uutta tietoa. Suurin osa nykyisestä tieteestä tehdään tutkimusryhmissä, joissa kukaan yksittäinen jäsen ei ymmärrä täydellisesti kaikkia tutkimusprosessin vaiheita.

Edelleen tutkimuksessa on jo pitkään hyödynnetty monimutkaisia laskennallisia malleja ja simulaatioita, jotka ovat käyttäjilleen eli tutkijoille usein tiedollisesti läpinäkymättömiä. Vaikka näitä malleja on kehitetty luotettaviksi ihmisten valvonnassa, niiden toiminta perustuu osittain optimointiin yrityksen ja erehdyksen kautta. Syväoppimismallit lähinnä jatkavat tätä kehitystä: nekin ovat ihmisten kehittämiä välineitä, joiden tarkoitus on tuottaa tieteellistä ymmärrystä kasvattavia tuloksia.

Yksilön sijasta vastuu tuloksista ja niiden luotettavuudesta on tiedeyhteisöllä, joka tekee ratkaisut uusien teknologioiden käyttöönotosta ja niiden hyödyntämisen rajoista. Näiden ratkaisujen onnistuminen puolestaan edellyttää ihmisen tiedollisten kykyjen rajojen jatkuvaa tunnistamista ja tunnustamista.

Kirjoittajat kiittävät dosentti Jyrki Konkkaa kommenteista.

Henrik Rydenfelt on käytännöllisen filosofian ja viestinnän dosentti ja yliopistotutkija Helsingin yliopistossa.

Jaakko Kuorikoski on käytännöllisen filosofian professori Helsingin yliopistossa.

Salla-Maaria Laaksonen on viestinnän dosentti ja yliopistotutkija Helsingin yliopiston Kuluttajatutkimuskeskuksessa.

KIRJALLISUUS

Bender, Emily M., Gebru, Timnit, McMillan-Major, Angelina, ja Shmitchell, Shmargaret 2021. On the dangers of stochastic parrots. Can language models be too big? In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 610–623.

Buijsman, Stefan 2023. Causal Scientific Explanations from Machine Learning. Synthese 202(6). https://doi.org/10.1007/s11229-023-04429-3

Crawford, Kate 2024. Generative AI's environmental costs are soaring – and mostly secret. Nature 626 (8000), 693. https://doi.org/10.1038/D41586-024-00478-X

Dhar, Payal 2020. The carbon impact of artificial intelligence. Nature Machine Intelligence 2(8), 423–425. https://doi.org/10.1038/s42256-020-0219-9

Eubanks, Virginia 2018. Automating Inequality. How High-Tech Tools Profile, Police, and Punish the Poor. New York: MacMillan.

Felzmann, Heike, Fosch-Villaronga, Eduard, Lutz, Christoph ja Tamò-Larrieux, Aurelia 2020. Towards Transparency by Design for Artificial Intelligence. Science and Engineering Ethics 26(6), 3333–61. https://doi.org/10.1007/s11948-020-00276-4

Kak, Amba, Myers West, Sarah ja Whittaker, Meredith 2023. Make no mistake – AI is owned by Big Tech. MIT Technology Review 5.12.2023. https://www.technologyreview.com/2023/12/05/1084393/make-no-mistake-ai-is-owned-by-big-tech/. Viitattu 18.12.2024

Krippendorff, Klaus 2011. Agreement and information in the reliability of coding. Communication methods and measures 5(2), 93–112.

Kuorikoski, Jaakko 2022. Factivity, pluralism, and the inferential account of scientific understanding. Teoksessa Scientific understanding and representation. Toim. Insa Lawler, Kareem Khalifa ja Elay Shech. New York: Routledge, 217–233.

Kuorikoski, Jaakko ja Ylikoski, Petri 2015. External Representations and Scientific Understanding. Synthese 192(12), 3817–37. https://doi.org/10.1007/s11229-014-0591-2

Lindgren, S. (toim.) 2023. Handbook of critical studies of artificial intelligence. Cheltenham: Edward Elgar Publishing.

Mittelstadt, Brent Daniel, Allo, Patrick, Taddeo, Mariarosaria, Wachter, Sandra ja Floridi, Luciano 2016. The Ethics of Algorithms. Mapping the Debate. Big Data & Society 3(2). https://doi.org/10.1177/2053951716679679

Novelli, Claudio, Taddeo, Mariarosaria ja Floridi, Luciano 2024. Accountability in Artificial Intelligence. What It Is and How It Works. AI & Society 39(4), 1871–82. https://doi.org/10.1007/s00146-023-01635-y

Powell, Alison 2021. Explanations as governance? Investigating practices of explanation in algorithmic system design. European Journal of Communication 36(4), 362–375.

Rydenfelt, Henrik, Haapanen, Lauri ja Lehtiniemi, Tuukka 2021. Dataa näkyvissä. Läpinäkyvyys algoritmien ja datan journalistisessa hyödyntämisessä. Media & Viestintä 44(2), 1–22. https://doi.org/10.23983/mv.109857

Rydenfelt, Henrik, Haapanen, Lauri, Haapoja, Jesse ja Lehtiniemi, Tuukka 2024. Personalisation in Journalism. Ethical Insights and Blindspots in Finnish Legacy Media. Journalism 25(2), 313–33. https://doi.org/10.1177/14648849221138424

Rydenfelt, Henrik, Lehtiniemi, Tuukka, Haapoja, Jesse ja Haapanen, Lauri 2025. Autonomy and Algorithms. Tracing the Significance of Content Personalisation. International Journal of Communication, tulossa.

Sajja, Priti 2021. Illustrated Computational Intelligence. Examples and Applications. Singapore: Springer. https://doi.org/10.1007/978-981-15-9589-9

Strathern, Marilyn 2017. The tyranny of transparency. The Anthropology of Organisations 26(3), 485–497. https://doi.org/10.4324/9781315241371-38

Tsamados, Andreas, Aggarwal, Nikita, Cowls, Josh, Morley, Jessica, Roberts, Huw, Taddeo, Mariarosaria ja Floridi, Luciano 2022. The Ethics of Algorithms. Key Problems and Solutions. AI & Society 37(1), 215–30. https://doi.org/10.1007/s00146-021-01154-8

Vallor, Shannon 2024. The AI mirror. How to reclaim our humanity in an age of machine thinking. Oxford: Oxford University Press.

van der Vlist, Fernando, Helmond, Anne ja Ferrari, Fabio 2024. Big AI. Cloud infrastructure dependence and the industrialisation of artificial intelligence. Big Data and Society 11(1). https://doi.org/https://doi.org/10.1177/20539517241232630

Woodward, James 2003. Making things happen. A theory of causal explanation. Oxford: Oxford University Press.

Zednik, Carlos ja Boelsen, Hannes 2022. Scientific Exploration and Explainable Artificial Intelligence. Minds and Machines 32(1), 219–39. https://doi.org/10.1007/s11023-021-09583-6

Zuboff, Shoshana 2019. The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power. PublicAffairs.