

Mitä toistettavuusongelmat tuovat tullessaan tekoälytutkimukseen?

Tilastollisten metodien käyttö on syössyt monia aloja toistettavuusongelmiin. Erityisen herkkiä tälle ongelmalle ovat olleet tekoälyyn liittyvät ja sitä hyödyntävät tutkimukset. Järjestelmien nopea kehitys on johtanut myös niiden nopeaan käyttöönottoon melkein päällä kuin alla. Tekoälyn toimintaperiaatteiden sekä niiden heikkouksien ja vahvuuksien ymmärtäminen vaatii tietenkin erityisosaamista, joka ei kuitenkaan ole kasvanut samassa tahdissa järjestelmien käyttöönoton kanssa. Toistettavuuteen liittyvät ongelmat ovat synnyttäneet vyyhdin, joka ei näyttäisi olevan aukeamassa ainakaan hetkeen.

Tammikuussa vuonna 2020 julkaistiin arvostetussa *Nature*-lehdessä artikkeli, joka esitteli koneoppivaa algoritmia¹ hyödyntävän järjestelmän rintasyöpäseulontaa varten². Tutkijajoukko oli ajanut kokeita isolla otoksella niin Yhdistyneissä kuningaskunnissa (UK) kuin Yhdysvalloissa³. He alleviivasivat, että kyseinen järjestelmä ylitti ihmisasiantuntijan kyvyt rintasyövän ennakoinnissa: koe antoi varhaista näyttöä siitä, että tekoälyn avulla potilaista kerättyä kuvadataa voidaan luokitella ja yleistää ihmissilmää tarkemmin.⁴ Tätä yleistettävyyttä he perustelivat sillä, että koneoppiva malli oli koulutettu käyttäen pelkkää UK:sta saatua dataa, minkä jälkeen järjestelmä oli koeajettu Yhdysvaltojen datalla. Tutkija esittivät, että järjestelmä oli ”suoritetun paremmin kuin radiologit”⁵. Kaiken kaikkiaan näytti siis siltä, että kehittäjät ja tutkijat – joista iso osa työskenteli muun muassa Alfabetin omistamilla Google Healthilla sekä DeepMindilla – olivat onnistuneet kehittämään tärkeän työkalun, joka parantaisi rintasyövän löydettävyyttä.

Tulokset olivat tietysti tervetulleita, onhan rintasyöpä yleinen niin maailmalla kuin esimerkiksi Suomessa, jossa se on yleisin naisten syöpä⁶. Tutkimusartikkelissa vielä muistutettiin, että mammografian ammattilaisista on pulaa, ja tekoäly voisi myös parantaa korkealaatuisen terveydenhuollon saavutettavuutta⁷. Tulokset eivät kuitenkaan vakuuttaneet kaikkia, ja jo samaisen vuoden lokakuussa *Nature*ssa ilmestyi huolestuneiden tutkijoiden vastine⁸. Heidän karu tuomionsa oli, että ”yksityiskohtien puute metodeissa ja algoritmin koodissa heikentää julkaisun tieteellistä arvoa”⁹. Muut eivät siis voineet todentaa artikkelissa esitettyjen tulosten ja päätelmien pitävyyttä. Vastineen tekijät viittasivat myös *Nature*en omiin julkaisuperiaatteisiin, joita artikkeli ei selvästi täyttänyt. He perään-

kuuluttivat tällaisen tekoälytutkimuksen läpinäkyvyyttä ja huomauttivat, ettei pelkkä ”tekstuaalinen kuvaus” syväoppivan mallin toimintaperiaatteista riittänyt, vaan tarvitaan myös yksityiskohtaista tietoa kokeiden metodeista sekä itse koodista. Lisäksi vastineessa esiteltiin toimia, joilla avoimuutta olisi voitu lisätä ilman, että se olisi vaarantanut koehenkilöiden yksityisyyden suojaa tai yrityssalaisuuksia. Kirjoittajat huomauttivat myös, että tällaisissa tapauksissa, joissa jopa ihmishenget saattavat olla vaakalaudalla, läpinäkyvyyden vaatimuksen olisi oltava entistä kovempi. He alleviivasivat, että artikkelin nykyiset heikkoudet eivät tee tutkimuksesta tiedettä, vaan lähinnä suljetun teknologian mainostusta.¹⁰

Samassa numerossa julkaistiin myös alkuperäistä tutkimusta tekemässä olleen ryhmän vastaus¹¹. Heidän kantansa oli, että he olivat julkaisseet riittävästi dataa itse algoritmista ja hyödynnetyistä otoksesta. He mainitsivat myös, että heidän esittelemänsä teknologia lasketaan yleensä ”lääketieteelliseksi teknologiaksi” (*medical device software*), jonka julkaiseminen ilman minkäänlaista sääntelyä tai valvontaa voisi johtaa sen väärinkäyttöön¹² ja keskeneräisen järjestelmän julkaiseminen voisi tuottaa myös ongelmia, joita ei ole tunnistettu. Ryhmän mukaan tekoälyn käyttöön terveydenhuollossa liittyvät vastuukysymykset ovat vieläkin ratkaisemattomia, joten olisi vastuutonta julkaista järjestelmä kokonaisuudessaan¹³. Lopuksi kirjoittajat myös lupasivat itse varmistaa, että heidän järjestelmänsä testataan laajamittaisesti ennen käyttöä kliinissä ympäristössä.

Replikaatio kriisissä tieteessä kuin tieteessä?

Edellä kuvattu tapaus on yksi esimerkki tekoälytutkimukseen ja -kehitykseen liittyvistä kokeiden toistetta-

”Replikaatio tai ’toistettavuus’ on ajateltu yhdeksi luonnontieteen peruskivistä. Tulokset ovat luotettavia ja ’tieteellisiä’, koska koetilanteet voidaan toistaa ja näin tuottaa samat tulokset uudelleen.”

vuusongelmista. Nämä ongelmat eivät kuitenkaan ole alalla uusi asia, eivätkä ne koske vain tietojenkäsittelyä tai sen tutkimushaaroja. Esimerkiksi 2000-luvulla on palattu erityisesti psykologian kohdalla niin sanottuun ”replikaatiokriisiin”, joka lyhykäisyydessään tarkoittaa ongelmia tieteellisten kokeiden ja tutkimustulosten toistettavuudessa¹⁴. Samaiset toistettavuusongelmat koskevat tietysti myös muita tieteen aloja ja alueita (esimerkiksi yhteiskunta-, käyttäytymis- ja biolääketieteet)¹⁵. Vaikka näiden ongelmien taustalla voi olla monia syitä, usein kuitenkin kvantitatiivisia metodeja hyödyntävät tieteet ja tutkimukset ovat erityisen alttiita ongelmille. Tällaiset ongelmat liittyvät yleensä datanhallintaan ja keräykseen, ohjelmistoihin, laitteistoihin sekä tutkijoiden monialaiseen asiantuntemukseen. Toisaalta erityisesti yhteiskuntatieteissä yleiset kvalitatiiviset tutkimustavat eivät kuitenkaan ole alttiita näille ongelmille, sillä ne perustuvat hieman erilaisiin tieteentekemisen tapoihin¹⁶.

On kuitenkin hyvä muistaa, että tutkijat yleensä ymmärtävät koetilanteen muodostavien tapahtumien raportoinnin merkityksen. Kaiken kaikkiaan koetilanteisiin liittyvät raportointikäytännöt ovat vanhempia kuin viimeisen parikymmenen vuoden kriisit tai tutkimusperiaatteiden penkomiset.¹⁷ Replikaatio tai ”toistettavuus” onkin ajateltu yhdeksi *luonnontieteen* peruskivistä. Tulokset ovat luotettavia ja ”tieteellisiä”, koska ne tuottaneet koetilanteet voidaan toistaa ja näin tuottaa samat tulokset uudelleen.¹⁸ Erityisesti tarkka toistettavuus voi kuitenkin olla vaikeaa

joillain aloilla, kuten sellaisilla, joissa tutkimuksissa käsitellään valtavaa joukkoa kliinisiä näytteitä tai historiallista dataa¹⁹.

Aluksi on hyvä erotella kolme eri käsitettä, jotka yhdistetään replikaatiokriisiin²⁰. Näille käsitteille on erilaisia määritelmiä ja ne saattavat viitata joko samoihin tai toisistaan poikkeaviin asioihin ja toimiin²¹. Myös käännöksessä saattaa piillä ongelmia. Englannin *repeatability* voidaan kääntää ”toistettavuudeksi”, kun taas *reproducibility* on selvästi enemmänkin ”uusinnettavuutta”. Näiden lisäksi on hankala ja jo edellä esitelty termi *replicability* (tai *replication*), joka voidaan kääntää anglismilla ”replikaatio”. Termeillä saattaa olla suuria eroja myös eri tieteenalojen välillä. Esimerkiksi tietojenkäsittelyssä ”uusinnettavuus” viittaa usein komputaatioiden (laskentojen) toistettavuuteen, kun taas vaikkapa psykologiassa uusinnettavuus on *joko* laskentojen tai kokeiden uusintamista.²²

Tieteenfilosofi Fiona Fidler ja kognitiotieteilijä John Wilcox esittävät, että yleensä kun ihmiset sanovat jonkin tutkimuksen olevan ”replikoitavissa”, he saattavat tarkoittaa sillä kahta asiaa: Tutkimus voi olla ”yleisesti” (*in principle*) replikoitavissa niin, että se voidaan tehdä uudelleen, kun sen menet, proseduurit ja analyysit on esitelty läpinäkyvästi ja riittävän tarkasti. Toisaalta ihmiset voivat viitata replikoitavuudella siihen, että tutkimus on replikoitavissa niin, että se voidaan toistaa onnistuneesti, ja tämä toistaminen tuottaa samat tai riittävän samat tulokset kuin alkuperäinen.²³ Tieteenfilosofi Eduardo

”Tekoälytutkimuksen ja -kehityksen yhteydessä replikaation arvioiminen on hankalaa siksi, että ala saattaa yhdistellä monia tieteenaloja.”

Machery esittelee tieteellisen replikaatiokokeen rakentuvan alkuperäisen kokeen ”koekomponenteista” (*experimental components*), joita kohdellaan satunnaismuuttujina. Tällainen koekomponentti on siis eräänlainen kokeen osa, jota voidaan kuitenkin muokata tai muuntaa itsenäisesti. Esimerkiksi psykologiassa erotellaan tällaisiksi komponenteiksi mittaukset tai esimerkiksi kokeen puitteet.²⁴

Tietojenkäsittelytieteessä replikaatioon liittyviä ongelmia käsitteli jo 1990-luvun alussa esimerkiksi Stanfordin yliopiston geofyysikko Jon Claerbout²⁵. Alalla erotellaan usein ”suora” (*direct*) ja ”käsitteellinen” (*conceptual*) uusinnettavuus. Edeltävä viittaa tiettyyn tulokseen, joka saadaan samasta datajoukosta samalla koodilla ja ohjelmistolla. Käsitteellinen uusinnettavuus on nimensä mukaisesti abstraktimpi ja viittaakin vain saman raakadatajoukon käyttöön.²⁶ Yleisesti ottaen tietojenkäsittelyssä *uusinnettavuuden* ajatellaan olevan kokeen uusintamista vaikkapa toisessa laboratoriossa (kehitysympäristössä). *Toistettavuus* taas saattaa olla vain kokeen *toistamista*, joka lopulta näyttäytyy alkuperäisen tutkimuksen osana. Näyttäisikin siltä, että uusinnettavuus ja replikaatio ovat alalla suurelta osin synonyymejä.²⁷

Arvovaltainen Association for Computing Machinery (ACM) on esittänyt, että edellä mainitut kolme käsitettä voitaisiin tulkita seuraavalla tavalla²⁸. Toistettavuus (*repeatability*) vaatisi tässä määritelmässä saman tutkijaryhmän ja samat tutkimuspuitteet. Näin ollen ainakin tutkijat itse voisivat toistaa oman kokeensa. Uusinnet-

tavuus (*reproducibility*) taas viittaa tilanteeseen, jossa niin ryhmä kuin puitteet poikkeavat alkuperäisestä. Tietojenkäsittelyn tapauksessa itsenäinen tutkijaryhmä voi saavuttaa samat tulokset välineillä, jotka he ovat kehittäneet itse. Replikaatio (*replicability*) taas kattaa saman tutkimuspuitteen kuin alkuperäinen koe, mutta toisella tutkijaryhmällä. Tietojenkäsittelyn kohdalla tämä tarkoittaa, että toinen ryhmä voi saavuttaa samat tulokset alkuperäisen ryhmän välineillä (*artifacts*).

Tekoälytutkimuksen ja -kehityksen yhteydessä replikaation arvioiminen on hankalaa siksi, että ala saattaa yhdistellä monia tieteenaloja – informaatiotiedettä, tietojenkäsittelytiedettä, psykologiaa, kognitiotiedettä, filosofiaa ja biologiaa. Etenkin koneoppivia malleja myös hyödynnetään monilla aloilla, aina terveystieteestä biologiaan, fysiikkaan, kemiaan, sosiologiaan, historiaan ja kirjallisuustieteeseen. Nykypäivänä alaa kuitenkin hallitsee, erityisesti koneoppivien algoritmien osalta, tilastotiede ja sen menetelmät. Tilastotiedettä sovelletaan niin erityistieteiden käytänteissä kuin myös itse tietojenkäsittelyn metodeissa ja sovelluksissa. Suuri osa toistettavuuden ongelmista liittyy juuri näihin metodeihin, esimerkiksi tilastojen otoskokoon, käyttöön ja niiden tulkintaan. Toisaalta ongelmia voi syntyä esimerkiksi replikaatiokokeen paikan, kokeen suorittajien, laitteiden (ja niiden toimintojen) sekä hyödynnettyjen datajoukkojen (tai populaatioiden) kohdalla.²⁹ Viimeisenä muttei vähäisimpänä ongelmana on usein myös suljettu, yrityssalaisuuden alainen koodi.

Sivupolku: algoritmit ja koneoppiminen³⁰

On hyvä palauttaa lyhyesti mieleen, millaisia järjestelmiä nykypäivän koneoppivat sovellukset ovat. Koneoppimisessa korostuvat nykyään neuroverkot, vaikka koneoppimisen tehtäviin voidaan soveltaa myös monia perinteisiä ja hyvin tunnettuja tilastollisia malleja. Karkeasti ottaen neuroverkot voidaan jakaa kahden tyyppiin malleihin: niin sanottuihin syväoppiviin (*deep learning*) sekä ei-syväoppiviin (esimerkiksi *perceptron*). Tiivistäen: neuroverkoiksi näitä järjestelmiä kutsutaan siksi, että niiden toiminta perustuu jokseenkin samankaltaiseen prosessiin kuin ihmisaivojen neuronit. Yksinkertaiset verkot (esimerkiksi *perceptron*) sisältävät vain syöte- ja tulostetasot, mutta syväoppivat verkot taas rakentuvat useista piilotetuista neuronitasoista (*hidden layers*), joiden suhteiden ja näiden painotusten välisinä ketjuina algoritmin ”laskenta” tapahtuu. Monimutkaisimmissa järjestelmissä voi olla jopa satoja tasoja.

Hieman yksinkertaistaen voidaan sanoa koneoppivien mallien perustuvan siihen, että niitä ”opetetaan” tiettyjen parametrien avulla löytämään esimerkiksi suhteita tietyistä datajoukosta. Koulutuksessa tuotetaan ”malli”, joka sitten sisällytetään lopulliseen algoritmiseen järjestelmään – tavallaan ”käännetään” algoritmiseen muotoon. Koulutusta on monenlaista, mutta karkeasti voidaan todeta, että on ohjattua oppimista (*supervised learning*) sekä ohjaamatonta oppimista (*unsupervised learning*). Ohjatussa oppimisessa ihminen jatkuvasti avustaa mallia (ja on annotoinut opetusdatan), kun taas ohjaamaton perustuu suurelta osin mallin omiin tai ympäristön palautteisiin (ja opetusdata saattaa olla tuntematonta).

Tällaisen mallin muodostaminen tapahtuu periaatteessa niin, että malli yritetään saada tietynlaiseen tasapainotilaan, joka vuorostaan määräytyy haluttujen (tai saavutettujen) painotusten mukaan. Erityisesti syväoppivien järjestelmien, neuroverkkojen, kohdalla tärkeiksi käsitteiksi ovat nousseet gradienttimenetelmä (*gradient descent*) sekä vastavirta-algoritmi (*backpropagation*). Syöte-, tuloste- ja mahdolliset piilotasot rakentuvat neuronikimpuista, joiden suhteilla on jokin painoarvo. Tällaisen verkon koulutuksessa pääosassa on oppimissääntö, jonka avulla voidaan muokata suhteiden painoja niin, että saavutetaan pysyvä rakenne – eli ”malli” –, jolla sitten saadaan aikaan haluttu kohdetuloste. Vastavirta-algoritmi viittaa siihen, että, sen sijaan, että järjestelmän prosessit kulkisivat syötteestä (mahdollisten piilotasojen kautta) tulosteeseen, tapahtuu myös ikään kuin käänteistä prosessia. Järjestelmässä saadun tuloksen ja halutun kohdetuloksen välisistä aste-eroista rakentuva tieto palautuu takaisin päin, jotta järjestelmä voi välittömästi asettua uudelleen vastaamaan entistä paremmin haluttua kohdetulostetta. Vastavirta-algoritmi siis varmistaa ”progressiivisen virheiden minimoinnin”. Tämä ”vastavirta” on samalla myös mallin oikeanlaisen painotuskokonaisuuden etsimistä – ja koko neuroverkon malli rakentuu tavallaan oikeanlaisten neuronisuhteiden painoarvojen kokonaisuudesta, jolla saavutetaan haluttu kohdetuloste. Tämän kokonaisuuden etsintää voidaan kutsua gradienttimenetelmäksi.³¹

Kaiken kaikkiaan suuressa osassa koneoppimista on kyse ennakoitavuusmallien rakentamisesta: esimerkiksi millä todennäköisyydellä kuvantunnistamisalgoritmien – jotka on koulutettu erottelemaan kuvajoukko kissakuviksi ja ei-kissakuviksi – avulla voidaan tunnistaa uudesta datajoukosta kissat ja ei-kissat. Koko algoritmin koulutuksen tärkein elementti onkin juuri data. Järjestelmä toimii oikeastaan vain niin hyvin kuin mihin koulutusdata antaa mahdollisuudet. Datan täytyykin olla mahdollisimman kattavaa ratkaisua vaativan ongelman kannalta. Näin on varsinkin silloin, jos halutaan, että algoritmin malli on yleistettävissä mahdollisimman moninaiisiin datajoukkoihin. Tietenkään algoritmi ei kykene käsittelemään mitä tahansa dataa, vaan uuden, opetusdatasta poikkeavan testidatan on jollain tavalla vastattava opetuksessa käytettyä dataa. Esimerkiksi kissakuvatunnistusalgoritmin uuden datan on oletettava sisältävän ainakin joitain kissoja, eikä esimerkiksi vain norsuja ja päästäisiä, jolloin mallin antamat tulokset eivät olisi hyödyllisiä.

Tiivistäen voidaan todeta, että toistettavuusongelmien näkökulmasta haasteena on ensinnäkin se, että *tällaisten algoritmien laskenta perustuu tilastollisiin metodeihin*. Toiseksi syväoppivien järjestelmien laskenta – ja erityisesti se, mitä tapahtuu piilotasoilla – on eräänlainen ”musta laatikko” (*black box*): usein kehittäjätkään eivät tiedä, miten järjestelmä päättyy lopulliseen malliinsa. Algoritmeihin yleisesti liittyy tietysti myös paljon muita ongelmia (esim. vinoumat ja adversariaalit), jotka kuitenkin kytkeytyvät tavalla tai toisella laskennassa käytettyihin metodeihin.³²

Tekoäly ja tietojenkäsittely erityissyyneissä?

Yksi syy, miksi syväoppiminen katsotaan erityisen alttiiksi näille ongelmille, on sen suhteellisen lyhyt historia. Kuten Meta AI:n tutkija ja McGillin yliopiston apulaisprofessori Joelle Pineau toteaa, alan koeluonteisuus on muotoutunut vasta viime vuosikymmenillä.³³ Vaikka tekoälytutkimus on jatkunut vuosikymmeniä, syväoppivat algoritmit ovat saaneet *nykymuotoaan* vasta 1980-luvun loppuvuosina.³⁴ Se, tulkitaanko tämä historia ”lyhyeksi” vai ”pitkäksi” näyttäisi riippuvan tulkitsijasta. Esimerkiksi tietojenkäsittelyn professori Arvind Narayanan ja jatko-opiskelija Sayash Kapoor esittävät *preprint*-artikkelissaan, että niin sanotun syväoppivan järjestelmän ”keskeiset innovaatiot” esiteltiin jo vuoden 1986 *Nature*-artikkelissa³⁵. He kutsuvat näitä innovaatioita ”muinaisiksi”. On kuitenkin tärkeää huomata, että laajemmin tekoälyn historian voi johtaa jopa antiikin Kreikkaan asti, ja neuroverkkojakin oli jo viimeistään 1950-luvulla³⁶. Viime vuosien laitteiden käytön ja tuotannon nopea kasvu ovat tietysti lisänneet mahdollisuuksia kehittää entistä parempia ja nopeampia järjestelmiä. Toisaalta järjestelmien ydinperiaatteiden kehitys on ollut hidasta. Taustalla saattaa myös olla se fakta, että järjestelmien pohjalla olevat tilastolliset mallit – ja laajempi teoreettinen, poikkeittieteellinen viitekehys – eivät ole kehittyneet yhtä nopeasti. Näin empiirisen kehitystyön ja teorian välille on saattanut syntyä kuilu.

”Ongelmat saattavatkin lävistää tekoälytutkimuksen lisäksi lukuisia muita aloja, joilla näitä teknologioita hyödynnetään.”

Tietysti myös muut alat ovat alkaneet hyödyntää tietokonesimulaatioita ja tekoälysovelluksia tutkimustulosten koonnissa sekä analysoinnissa. Näiden sovellusten käyttöä ovat lisänneet myös laajasti käytettävissä olevat avoimet tietokannat.³⁷ Näin on niin luonnontieteissä kuin ihmis- ja yhteiskuntatieteissäkin³⁸. Ongelmat saattavatkin lävistää tekoälytutkimuksen lisäksi lukuisia muita aloja, joilla näitä teknologioita hyödynnetään. Esimerkiksi eräässä tutkimuksessa yritettiin toistaa 306 laskennallisia metodeja hyödyntävän fysiikan alan tutkimusjulkaisun tulokset. Replikaation tekijät eivät kyenneet toistamaan yhdenkään tutkimusjulkaisun kaikkia tuloksia. Osa tuloksista onnistui, kun alkuperäisten tutkimusten tekijät tarjosivat lisämateriaalia.³⁹

Yhden esimerkin algoritmiaivusteisten tutkimusten toistettavuusongelmien laajuudesta tarjosi tutkijajoukko vuonna 2019: heidän mukaansa yli 20 000 tekoälyä lääketieteen kuvantamisessa hyödyntäneen tutkimuksen joukosta vain 5 prosenttia sisälsi riittävästi tietoa, jotta ne voitaisiin toistaa⁴⁰. Tänä vuonna jo edellä mainitut Kapoor ja Narayanan taas tarjosivat alustavia tutkimustuloksia koneoppivien mallia hyödyntävien tutkimusten toistettavuudesta⁴¹. Heidän katsaus- ja toistettavuuskoejulkaisussaan kartoitettiin isoa joukkoa tekoälyä hyödyntävien tutkimusalojen katsauksia. Kaiken kaikkiaan 329 tutkimusjulkaisun tuloksia ei voitu täysin toistaa.

Kapoorin ja Narayananin oma toistettavuuskoe koski politiikantutkimuksen aluetta, jossa hyödynnetään koneoppimista sisällissodan ennakoinnin tutkimuksessa. Tutkijat

eivät keskittyneet kaikkiin mahdollisiin toistettavuuden ongelmiin, vaan niin sanottuihin ”datavuotoihin” (*data leakage*). Nämä vuodot viittaavat ongelmiin, joissa data, josta koneoppiva malli ”oppi”, on sisällytetty osittain tai sen osia on sisällytetty siihen dataan, jonka avulla mallia ”arvioidaan”. Koulutuksessa siis opetusdata ja testidata ovat sekoittuneet. Käytännössä tämä tarkoittaa sitä, että ”malli on nähnyt vastaukset etukäteen” – ja näin ollen sen ennakointikykyä luullaan paremmaksi kuin se todellisuudessa on⁴². Kapoor ja Narayanan tunnistavat kahdeksan erityyppistä vuotoa, joihin he myös tarjoavat ratkaisuja – tai ainakin ohjeita kuinka minimoida näitä vuotoja.

”Vuotoja” ovat esimerkiksi koulutus- ja testausdatan erottelemattomuus sekä mallissa käytetyt epäoikeudet ominaisuudet. Lisäksi ”otantavinouma” (*sampling bias*) voi aiheuttaa datavuotoa: esimerkiksi testausdata valitaan maantieteellisestä sijainnista A, mutta sen pohjalta tehdään lopulta päätelmiä maantieteellisestä paikasta B⁴³. Kapoor ja Narayanan esittelevät lyhyesti myös muita, datavuotoihin liittymättömiä ongelmia, joita käsitellään myöhemmin. Datavuotojen käsittelylle he kuitenkin tarjoavat joukon ohjeita, jotka ovat hyvin suoraviivaisia, kuten ”erota selvästi ja tarkasti koulutus- ja testidatat”, ja ”käytä vain ominaisuuksia, jotka voidaan katsoa mallin viitekehityksessä ’laillisiksi’ (*legitimate*)”. Kirjoittajat tarjoavat myös laajempia ohjeita, joilla koneoppivia metodeja hyödyntäviä tutkimuksia voidaan parantaa. Palaamme näihin ohjeistuksiin sekä ongelmiin – joita myös muut tutkijat ovat käsitelleet – myöhemmin.

”Järjestelmien harjoittamat tarkat ennakoinnit ohjaavat kehittäjiä ja tutkijoita usein oletamaan, että nämä teknologiat kykenisivät käsittelemään ’ongelmien todellisia rakenteita ihmisten tapaan’.”

Kapoorin ja Narayananin esimerkkinä käyttämän sisällissodan ennakoinnin kohdalla useat julkaisut ovat päätyneet korostamaan kuinka syväoppivat mallit toimivat paremmin kuin esimerkiksi klassiset regressiomallit. Heidän tutkimuksensa mukaan kaikki tarkastelun alaiset julkaisut olivat kontaminoituneet datavuodoista, ja kun nämä ongelmat pyrittiin korjaamaan, tulokset eivät osoittaneet, että syväoppivat mallit olisivat parempia tai tehokkaampia kuin näissä tutkimuksissa perinteisesti käytössä olevat regressiomallit. Kapoor ja Narayanan tiivistävät, että syväoppivia malleja hyödyntävän tieteen toistettavuuskriisi johtuu kahdesta syystä:

”Ensinnäkin tuloksemme osoittavat, että toistettavuuden ongelmat [...] ovat systemaattisia. Lähes jokaisella tieteenalalla, joilla on suoritettu järjestelmällisiä tutkimuksia toistettavuuden ongelmista, julkaisut ovat täynnä yleisiä sudenkuoppia. [...] Näin ollen näyttäisi myös siltä, että samanlaisia ongelmia esiintyy lukemattomilla aloilla, jotka ovat omaksuneet koneoppivia metodeja. [...] Toisekseen, ei ole olemassa systemaattisia ratkaisuja, joita olisi otettu käyttöön näille heikkouksille. Tieteelliset yhteisöt kohtaavat samoja ongelmia alasta toiseen, mutta eivät ole vielä päässeet yhteisymmärrykseen parhaista käytänteistä, joilla näitä toistettavuuden ongelmia voisi välttää.”⁴⁴

Järjestelmien harjoittamat tarkat ennakoinnit ohjaavat kehittäjiä ja tutkijoita usein oletamaan, että nämä tek-

nologiat kykenisivät käsittelemään ”ongelmien todellisia rakenteita ihmisten tapaan”⁴⁵. Tämä on erityisen huolestuttavaa juuri silloin kun taustalla on vuodon aiheuttama lähes tautologinen itseään toistavuus⁴⁶. Tekoälytutkimuksessa on tietysti myös muita ongelmia, joita erityisesti kvantitatiivisiin metodeihin keskittyvissä tutkimuksissa ja aloilla esiintyy. Esimerkiksi P-hakkerointi, *HARKing* (*Hypothesising After Results are Known*), julkaisemiseen liittyvät vinoumat sekä tieteen palkitsemismekanismit. Näistä esimerkiksi P-hakkerointi on suoraan kytköksissä tilastoihin, ja se tarkoittaa ”sitä, että aineistosta kalastellaan tuloksia vaihtelemalla otoskokoa, muuttujia, analyysitapoja tai jopa hypoteesia”⁴⁷. *HARKing* viittaa siihen, että lähtöodotuksia ja -oletuksia muokataan vastaamaan esimerkiksi yllättäviä tutkimustuloksia – eli esitetään, että ne olisivat olleet juuri odotettuja tuloksia⁴⁸. Julkaisemiseen sekä palkitsemismekanismeihin liittyvät vinoumat ovat tietysti monisyisiä. Ne pitävät sisällään esimerkiksi sen, että vain menestyneitä tai uusia tuloksia julkaistaan, tai että tutkijat usein piilottavat ”pöytälaatikkoon” esimerkiksi tilastollisesti merkityksettömät tutkimustulokset (*file-drawer problem*). Lisäksi akateemisen työn luonne – ”julkaise tai tuhoudu” – painostaa helposti taivuttamaan tilastoja haluttua tulosta kohti.⁴⁹

Tietojenkäsittelyssä ja tekoälytutkimuksessa on tietysti omia erityispiirteitä. Jon Claerhout onkin todennut, että tietojenkäsittelyssä toistettavuus voidaan toteuttaa jopa niin yksinkertaisesti, ettei siihen tarvita ”asiantuntijaa”⁵⁰. Näin on varmasti esimerkiksi ei-sy-

väoppivien järjestelmien kohdalla. Joka tapauksessa suuri osa koevälineistä, -mittareista, -metodeista ja -puitteista on oltava tallennettuna koodina, ja jokaisesta toimesta voidaan tallentaa merkintä lokiin⁵¹. Tämä tietysti vaatii ensinnäkin sen, että koodi on avoimesti ja läpinäkyvästi saatavilla. Koodi ei kuitenkaan usein ole kokonaisuudessaan saatavilla, vaan siitä saattaa olla vain otteita tai algoritmeista saattaa olla pelkkä tekstuaalinen kuvaus tai vuokaavio. Saattaa myös olla niin, että jos kehitysympäristöä ei esimerkiksi jäädytetä johonkin palveluun tai alustalle, vuosien kuluessa koodin hyödyntämät kirjastot tai vaikka koko ohjelmointikieli saattavat vanhentua. Toisaalta nämä voivat päivittyä myös niin, ettei vanhoja käskyjä enää tunnusteta.

Koneoppivissa malleissa, joissa tarkoituksena on kehittää ”malli”, jota sitten opetetaan koulutusdatalla, tärkeää on myös itse koulutusprosessi ja siinä tapahtuva parametrien säätäminen. Lisäksi alkuperäistä koulutusdataa ei välttämättä ole tarjolla. Lopulta pääsemme klassiseen musta laatikko -ongelmaan: järjestelmän kehittäjät eivät välttämättä tiedä, miten monimutkainen koneoppiva neuroverkko päättyy syötteeseensä – ”piilotetut” neuroverkon tasot ovat todellakin piilossa. Ohjelmistopuolen lisäksi ongelmia saattaa syntyä myös laitteistopuolella: vanhat laitteet eivät välttämättä ole yhteensopivia uusien kanssa tai ne ovat saattaneet esimerkiksi vaurioitua.⁵² Näiden lisäksi vielä kuvankäsittelyn yhteydessä mukaan voi tulla tahattomia ja tahallisia ”artefakteja”, pikseleitä, jotka saattavat ohjata kuvien tulkintaa suuntaan jos toiseenkin⁵³.

Toisinaan tutkijoilla saattaa olla vaikeuksia ymmärtää ennustettavuuden ja ennustamisen prosessien rajoja⁵⁴. Abstraktimmin ajateltuna kehitysprosessissa on hyvä huomioida se, mitä voidaan lopulta kääntää algoritmeiksi tai ylipäätään digitalisoida ja mitä ei⁵⁵. Kun ilmiöt ja asiat käännetään digitaaliseen muotoon, jotain jää aina ulkopuolelle. Kehitysprosessin aikana on tasapainoitettava tarkasteltavien ilmiöiden ja ratkaisuja vaativan, ongelmia koskevan asiantuntijuuden sekä teknisen asiantuntijuuden välillä. Valtava laskentateho tai datamäärä eivät korvaa laadullisia puutteita, joita mittaamisen tai ilmiötä numeeriseksi kääntämisen prosesseissa on jo saattanut syntyä. Havahduttavat laskennan aikaan saamat tai tekoälysovelluksen sylkemät upealta näyttävät ”tulokset” on palautettava ilmiömaailmaan ja tulkittava uudelleen.⁵⁶

Oikeanlainen asiantuntijuus onkin tärkeää. Esimerkiksi pääasiassa ohjelmistokehittäjistä ja -tutkijoista koostuva ryhmä ei välttämättä kykene huomioimaan kaikkia osa-alueita, joita sovellusympäristö vaatisi. Tämä voi tietysti todellistua myös kääntäen: ”humanisteilla” ei välttämättä ole riittävää osaamista ymmärtää käyttämiensä koneoppimisalgoritmiensa nyansseja ja rajoja. On myös mahdollista, että ongelmat kasautuvat, kun tietojenkäsittelyn tutkimuksessa ja siitä kumpuavassa kehitystyössä käytetään ongelmallisia metodeja, joita sitten käyttävät myös luodut koneoppivat sovellukset sekä lopulta erityistieteilijät, jotka hyödyntävät järjestelmiä omalla osaamisalueellaan⁵⁷.

Näiden lisäksi useissa tapauksissa – kuten edelläkin on jo todettu – on kyse yksityisen yrityksen omistamasta tuotteesta, koodista. On tietysti ymmärrettävää, että tuotteita valmistavat tahot pyrkivät säilyttämään jonkinasteisen yrityssalaisuuden. Tästä huolimatta tieteellinen tutkimus vaatii metodien ja tulosten avoimuutta niin, että ne on mahdollista toistaa ja varmistaa. Muuten muu tiedeyhteisö on vain kehittäjien ja alkuperäisten tutkijoiden sanan varassa. Tällaisissa tapauksissa myös tieteellisillä julkaisuilla, kuten vaikkapa arvovaltaisella *Nature*lla, on oma vastuunsa⁵⁸. Joka tapauksessa on noussut monia yritysajattelisiä, jotka pyrkivät tuomaan nopeasti markkinoille monenlaisia sovelluksia. Näin monia tuotteita siirretään nopeasti arkielämän käyttöön, myös sellaisiin ympäristöihin, joissa ongelmat saattavat olla jopa hengenvaarallisia (kuten terveydenhuoltoon)⁵⁹.

Teknologiajättiläisten ja tutkimuksen välille on noussut myös monia esteitä. Yksi syy siihen, että monet tutkijat ovat siirtyneet Alfabetin, OpenAI:n tai Facebookin kaltaisten yritysten palkkalistoille, on näiden toimijoiden suuret resurssit. Teknojättiläisillä on mahdollisuuksia hankkia niin laitteistoja kuin dataa – toisin kuin monilla puhtaasti akateemisilla tutkimusprojekteilla. Monet tärkeät järjestelmät maksavat valtavia summia. Esimerkiksi viime vuosina paljon esillä ollut tekstintunnistus ja -tuottamisalgoritmi GPT-3:n kouluttaminen on maksanut arviolta 10–12 miljoonaa dollaria – ja tällöin kyse on vain viimeisestä mallista, ei esimerkiksi sitä edeltävästä kehitystyöstä, joka on sisältänyt prototyyppisiä ja näiden koulutusta⁶⁰. Myös monet ilmastotieteelliset mallinnukset vaativat valtavaa konetehoa, jota kaikilta tutkimusprojekteilta ei löydy⁶¹.

Ratkaisuja ja suuntaviivoja?

Alalla kuin alalla on tärkeää alleviivata, että replikaation epäonnistuminen ei tietenkään automaattisesti tarkoita tutkimuksen epäonnistumista tai epätieteellisyyttä eli ”huonoa tiedettä”. Monet tärkeät löydökset ja tulokset ovat vaikuttaneet löytöhetkellä hyvin epätodennäköisiltä, ja toisaalta selkeiden ja ideaalisten kokeiden ja koetilanteiden suunnittelu ja toteutus voivat olla vaikeita tai jopa mahdottomia. Lisäksi epäonnistuneet replikaatiot saattavat itsessään olla ”väärä negatiivisia” tuloksia (*false negative*).⁶²

Kuten jo aikaisemmin todettiin, replikaatio-ongelmat eivät ole uusia oikeastaan millään alalla, eivät myöskään tietojenkäsittelyssä. Viime vuosina on kuitenkin ryhdytty toden teolla laajamittaisempiin uusintamisyhteyksiin. Toisinaan näitä ovat toteuttaneet tutkijat ja kehittäjät itse omille vanhoille koodeilleen⁶³. Esimerkiksi psykologiassa on käynnistetty *The Reproducibility Project: Psychology*, jota koordinoidaan Center for Open Sciencen taholta. Myös niin sanottu *meta-science*- tai *meta-research*-tutkimussuunnat ovat pyrkineet tarttumaan replikaatioon muiden tieteen ongelmien ohella.⁶⁴

Vastaiskuna teknologiajättiläisten koneteholle on syntynyt monia projekteja, joissa tarjotaan myös esimerkiksi yli-

”Valtava laskentateho tai datamäärä eivät korvaa laadullisia puutteita, joita mittaamisen tai ilmiötä numeeriseksi kääntämisen prosesseissa on jo saattanut syntyä.”

opistojen käyttöön suuria laskentatehoja. Hieman samantyyppisiä projekteja on ollut jo aiemminkin, kuten esimerkiksi fysiikassa ”suuri hadronitörmäytin” (*Large Hadron Collider*), jonka käyttöä on allokoitu erilaisille tahoille. Suomeen perustettiin jo 70-luvulla CSC – Tieteen tietotekniikan keskus, joka on tarjonnut vuosikymmenet Suomen yliopistoille laskentatehoa. CSC kunnostautunut myös kvanttietokoneiden tutkimuksessa, ja se on esimerkiksi Suomeen juuri rakennetun ja käyttöön otetun LUMI-supertietokoneen taustalla. Tutkimusprojektien on tarkoitus päästä käyttämään tätä konetta koko Euroopan laajuisesti vuoden 2022 aikana.⁶⁵ Tällaisia projekteja on kehitetty myös muualla⁶⁶.

Samalla monet julkaisut ovat alkaneet vaatia, että tutkimusprosessin eri vaiheet, kaikki data ja esimerkiksi koodi olisivat läpinäkyviä ja ne olisivat tavalla tai toisella ylöskirjattu. Lisäksi joissain julkaisuissa voi pyytää erikseen ”replikaatioarviointia”. Julkaisujen käytännössä on kuitenkin paljon eroja. Kaikissa ei esimerkiksi pyydetä kaikkea materiaalia jaettavaksi, mutta saatetaan vaatia, että tarkempaa dataa on jaettava pyydettyä. Joissain julkaisuissa on otettu käyttöön – tai ainakin suunniteltu – myös uudenlaisten tunnusten käyttöä niille tutkimusartikkelille, jotka ovat käyneet läpi replikaation. Vielä ei kuitenkaan ole syntynyt mitään laajoja ja yhtenäisiä toimintaperiaatteita – edes alojen sisällä.⁶⁷ Yhtenäiset periaatteet vaatisivat myös yhtenäisiä alustoja, joilla jakaa koodia ja dataa. Joitain tällaisia tietysti jo on (esim. GitHub, Zenodo), mutta ne eivät ole välttämättä tulleet kaikilta

osin standardiksi. Lisäksi CodeOceanin kaltaiset palvelut tarjoavat jopa ”pilvilaskentaa” (*cloud computing*), jolla voi luoda toistettavuuskoetta varten alkuperäisen tutkimuksen kaltaiset ”laskennalliset” olosuhteet⁶⁸.

Avoimen tieteen käytännöt ovat saaneet tuulta alleen myös replikaatiokriisien johdosta⁶⁹. Tämä on saattanut jo jonkin verran muuttaa tieteen ”palkitsemiskäytäntöjä”: esimerkiksi Utrechtiin yliopisto on suunnannut palkkaamiskriteerit huippujulkaisujen painotuksesta kohti avoimen tieteen käytäntöjen painotusta. Myös Suomen Akatemia vaatii nykyään, että sen rahoituksen piirissä tuotetut tulokset ovat avointa tiedettä.⁷⁰ Replikaatiota voitaisiin tuottaa myös monien tutkijoiden toimesta eri vaiheissa ja eri rajapinnoilla⁷¹. Toistokokeiden tekemisestä ei tietystikään sinänsä palkita, vaan se toimii hieman samaan tapaan kuin vertaisarviointi yleensä – tai tarkkaan ottaen on osa vertaisarvioijien työkalupakkia. Vertaisarviointi ja replikoijat voivat myös olla ylikuormittuneita jo nyt, mikä entisestään lisää riskiä julkaista heikkoa tiedettä⁷².

Replikaatiokriisi näyttäytyy helposti vain tieteen sisäisenä ongelmana, vaikka sillä voi olla myös laajoja yhteiskunnallisia vaikutuksia varsinkin hauraassa asemassa olevien kansalaisten kannalta. Nykyisin monet tekoälyä hyödyntävät, kansalaisille suunnatut tai heitä koskevat sote-palvelut kehitetään yhteistyössä julkisten organisaatioiden ja yritysten välillä.⁷³ Näihin projekteihin sisältyy monia jännitteitä, jotka liittyvät erityisesti avoimuutta korostavan tieteellisen tutkimuksen ja liiketoiminnallisen

tuotekehityksen välille. Kun tutkimus- ja kehityshankkeita koskevia raportteja ei ole avoimesti saatavilla yrityssalaisuuden alaisten koodien takia, tutkimustulosten seuranta ei ole mahdollista esimerkiksi replikaatiotutkimuksen avulla. Kansalaisille keskeisten palvelujen kehitystyön perusteet jäävät helposti pimentoon sekä päättäjiltä että kansalaisilta. Läpinäkyvyyden puuttumisella voi olla merkittäviä yhteiskunnallisia seurauksia varsinkin silloin, kun palveluiden käyttöönotto aiheuttaa kansalaisille terveydellisiä tai taloudellisia ongelmia.⁷⁴

Erityisesti tietojenkäsittelyn ja tekoälykehityksen projekteissa jännite kehittyy usein juuri jaetun tieteellisen tutkimuksen ja liiketoiminnallisen tuotekehityksen välillä. Onkin ehdotettu, että *Nature*n kaltaiset joulunäytelmät voisivat tämän tyyppisissä tapauksissa ottaa käyttöön kaksi julkaisulinjaa. Yhdessä julkaistaisiin puhtaasti avoimia tieteellisiä tutkimuksia ja toisessa niin sanottuja *tech showcases* -tapauksia, joissa voitaisiin esitellä kehitteillä olevia uusia teknologioita⁷⁵. Tämä voisi helpottaa rahoittajien asettamien vaatimusten ja tutkijoiden tieteellisten kriteerien välisiä jännitteitä. Rahoittajat voivat vaatia ”tuotoksia” (*deliverables*), jotka ovat jotain muuta kuin tutkijoiden tuottamia tieteellisiä vertaisarvioituja artikkeleita. *Tech showcase*n käyttöönotto tosin saattaisi ajaa tutkijat noudattamaan yhä enemmän rahoittajien toiveita tieteellisen vertaisarvioinnin sijasta⁷⁶.

Myös poikkitieteellisyys voisi lisätä ja helpottaa ongelmien tunnistamista, jolloin myös kehitettäisiin parempia järjestelmiä. Toisaalta ”monitieteisten tutkimusasetelmien toteuttaminen voi kilpistyä jo lähtöasetelmissään menettelytapojen eroihin tai taustafilosofioiden ristiriitaan⁷⁷. Lähestymistapojen erot voivat kärjistyä jopa ylimielisyydeksi: tekoälykehittäjät saattavat nähdä projektiin osallistuvat muiden alojen asiantuntijat lähinnä dataa

keräävinä työläisinä, ei vertaisinaan asiantuntijoina⁷⁸. Ylimielisyyttä voi tietenkin ilmetä myös toisin päin: humanistit voivat kokea tietojenkäsittelyn asiantuntijat vain ohjelmien tuottajina, eivät tutkijoina. Monitieteisyys ja erilaisten nyanssien huomioiminen varmasti hidastavat omalla tavallaan kehitysprosessia, mutta tällainen ilmiöiden moniulotteinen kartoitus tuskin on lopulta huono asia. Tietysti myös humanistisemmin orientoituneiden alojen edustajat voisivat uskaltautua kiinnostumaan ”uudenlaisista numeerista dataa hyödyntävistä teknologioista ja päätöksenteon tavoista”⁷⁹. Tämä voisi vuorostaan rohkaista esimerkiksi tietojenkäsittelytieteilijöitä kiinnostumaan ilmiöiden sosiaalisista ulottuvuuksista⁸⁰.

Kootusti voidaan todeta, että replikaatiokriisien kohdalla – alalla kuin alalla – uudistusta ja kehitystä kaipaivat niin tutkijakoulutus, tutkimusmenetelmät, raportointi, levitys, vertaisarvioprosessi kuin palkkiojärjestelmätkin. Monet alleviivaavatkin yhteistyön ja avoimuuden periaatteita. Replikaation kohdalla metodien toistettavuutta voidaan parantaa tarjoamalla riittävän yksityiskohtaiset tiedot käytänteistä ja datasta. Vaikka ”riittävä” voidaan tässä tietysti ymmärtää monin tavoin, pitäisi tietojen avulla ainakin pystyä toistamaan sama koe samoin toimin. Näin myös tulokset ja johtopäätökset voisivat olla samat tai lähes samat.⁸¹

Replikaation mukana on tieteeseen siirtynyt myös oletus tieteen itsekorjaavuudesta⁸². Itsekorjaavuus ei kuitenkaan tarkoita, että tieteessä ei voisi olla käytänteitä, ohjeistuksia ja standardeja – nämähän myös ovat syntyneet usein tieteen sisältä käsin. Felipe Romeron mukaan saattaa olla, että replikaatiokriisi (tai -kriisit) paljastavat kuilun tiedekäsityksemme ja todellisuuden välillä⁸³. Jotkut ovat jopa vielä radikaalimmalla kannalla: replikaatiota ei pitäisi käyttää onnistuneen tieteen mittana lainkaan⁸⁴.

Viitteet

1 Tekstissä käytetään koneoppivia algoritmeja, algoritmisia järjestelmiä ja tekoälyä pääasiassa synonyymeinä. Tästä huolimatta näillä käsitteillä on tietysti myös eroja, ja esimerkiksi tekoäly viittaa usein niin käytännöllisiin (algoritmeja hyödyntäviin) teknologioihin, mutta myös näiden tutkimukseen ja kehitykseen sekä esimerkiksi kognitiiviseen, psykologian ja filosofian osa-alueisiin. Käytämme ter-

mistä *computer science* yksinkertaisuuden vuoksi suomennosta ”tietojenkäsittely” tai ”tietojenkäsittelyoppi”, vaikka termi saatetaan kääntää myös toisin (esim. tietotekniikaksi).

2 McKinney ym. 2020a. Ks. myös Heaven 2020.

3 Esimerkiksi UK:n otos kattoi mammografioita 25856 naiselta, ja nämä oli otettu vuosina 2012–2015. Yhdysvaltojen osalta otos kattoi vuosina 2001–2018 otettuja kuvia 3097 naiselta.

4 McKinney ym. 2020a, 89.

5 Sama, 93.

6 Esim. Pitkäniemi ym. (toim.) 2022, 5.

7 McKinney ym. 2020a, 89; Ks. automatisaation puolustamisesta terveydenhuollossa myös esim. Parviainen & Rantala 2022.

8 Haibe-Kains ym. 2020.

9 Sama, E14. Kaikki alkuperäistekstien käännökset ovat kirjoittajien.

10 Sama. Ks. myös Heaven 2020.

11 McKinney ym. 2020b.

- 12 Sama, E17.
- 13 Sama.
- 14 Ks. tästä ja tapausesimerkeistä Silfverberg 2021; myös Koskinen 2021 sekä Machery 2020; Romero 2019; Plesser 2018.
- 15 Esim. Romero 2019; Fidler & Wilcox 2021.
- 16 Ks. tästä ja laajemmin tekstin teeman avaamista tieteenfilosofisista kysymyksistä ja ongelmista esim. Koskinen 2016.
- 17 Machery 2020, 550; myös esim. Fidler & Wilcox 2021.
- 18 Romero 2019, 1–2. Ks. joistain tähän liittyvistä ongelmista lyhyesti esim. Jensen 2019.
- 19 Smaldino & McElreath 2016, 14.
- 20 Tietysti myös itse replikaatiokriisin ajatus on hyvin moniulotteinen. Esimerkiksi Fidler & Wilcox (2021) esittävät viisi erilaista näkökulmaa tai lähestymistapaa, joiden kautta tällainen kriisi yleensä ymmärretään tai jolla sen olemassaoloon viitataan: a) replikaatiotutkimuksia ei ole oikeastaan ollenkaan tieteellisessä kirjallisuudessa (tietyllä alalla); b) laajoja ongelmia julkaistujen tutkimusten replikaatioissa; c) todisteita julkaisuvuonoumasta (*publication bias*); d) kyseenalaisten tutkimuskäytänteiden laaja levinneisyys (esim. jollain alalla); e) puutteita metodien, datan ja analyysien läpinäkyvyydessä.
- 21 Ks. esim. Fidler & Wilcox 2021; Plesser 2018; Kapoor & Narayanan 2022.
- 22 Fidler & Wilcox 2021.
- 23 Sama.
- 24 Machery 2020, 547, 550.
- 25 Plesser 2018, 1.
- 26 Ks. Machery 2020, 545–546; Romero 2019, 5.
- 27 Plesser 2018; Fidler & Wilcox 2021. Esimerkiksi Jon Claerbout on tiivistänyt toistettavuuden tarkoittavan saman ohjelman ajamista samalla datasyötteellä ja saavuttamalla saman tuloksen (Plesser 2018, 1).
- 28 Ks. myös ACM 2020. ACM on vuonna 1947 perustettu järjestö, jonka tarkoituksena on edistää tietojenkäsittelyn tutkimus- ja opetustyötä.
- 29 Fidler & Wilcox 2021; Koskinen 2021.
- 30 Tämä tiivis esittelykappale perustuu esimerkiksi teksteihin Alpaydin 2021; Buckner 2019; DeLanda 2015; Kelleher 2019; Louridas 2020.
- 31 DeLanda 2015, 88–90.
- 32 Ks. adversariaaleista esim. Rusanen 2019. Yleisesittely algoritmien ja tekoälyn ongelmista esim. Ojanen ym. 2022, 51–.
- 33 Heaven 2020.
- 34 Ks. aikaisemmasta keskustelusta esim. Dreyfus 1992; Weizenbaum 1976.
- 35 Narayanan & Kapoor 2022; alkuperäinen, Rumelhart ym. 1986.
- 36 Ks. esim. Nivala 2019.
- 37 Esim. Peng 2011, 1226; Coveney ym. 2021, 2. Siirtyminen symbolisesta tekoälystä kohti koulutettavia neuroverkkoja on lisännyt tilastojen ja tilastollisten metodien merkitystä tekoälytutkimuksessa.
- 38 Ks. esim. Nivala 2018.
- 39 Ks. Coveney ym. 2021, 2; Krafczyk ym. 2021; Stodden ym. 2018.
- 40 Liu ym. 2019; lyhyesti myös Gibney 2022.
- 41 Ks. arXiv-sivustolla julkaistu vertaisarviota edeltävä artikkeli Kapoor & Narayanan 2022; ks. tiivistys ja kirjoittajien kommentteja, Gibney 2022.
- 42 Gibney 2022, 250; Kapoor & Narayanan 2022. Ks. “datavuodoista” myös esim. Kaufman ym. 2011.
- 43 Kapoor & Narayanan 2022. Toinen konkreettinen esimerkki, jonka kirjoittajat tarjoavat, on tutkimukset, joissa koneoppivilla malleilla pyritään ennakoimaan autismia, mutta joissa autismin rajatapaukset suljetaan ulos, jolloin data ei enää edusta kattavasti yleistä populaatiota.
- 44 Kapoor & Narayanan 2022, 4–5.
- 45 Gibney 2022, 251.
- 46 Ks. tästä tautologisuudesta Kaufman ym. 2011.
- 47 Silfverberg 2021, 6.
- 48 Tietysti tutkimustuloksia ja dataa voidaan muokata vastaamaan haluttuja tuloksia (ks. esimerkkitapauksesta Piller 2022).
- 49 Romero 2019, 4; Fidler & Wilcox 2021; Plesser 2018, 1. Ks. myös “vehkeilyrenkaista” (*collusion ring*) esim. Littman 2021.
- 50 Plesser 2018, 1.
- 51 Peng 2011, 1226.
- 52 Plesser 2018; Peng 2011; Heaven 2020.
- 53 Piller 2022. Erityisesti kuvantunnistamisessa ongelmaksi nousevat edellä mainitut adversariaalit, ks. näistä Rusanen 2019.
- 54 Esim. Kapoor & Narayanan 2022.
- 55 Ks. esim. Rantala 2018.
- 56 Alanen 2022.
- 57 Kiinnostavaan on, kun tähän lisätään vielä uusi taso koneoppivan sovelluksen yrittäessä kirjoittaa ”oma tieteellinen artikkelinsa” (ks. Thunström 2022).
- 58 Ks. esim. Peng 2011, 1226.
- 59 Ks. myös Parviainen & Rantala 2022.
- 60 Heaven 2020; Ks. GPT-3:sta, esim. Floridi & Chiriatti 2020. Myös muita malleja on, kuten läpinäkyvyyttä korostava BLOOM (Heikkilä 2022).
- 61 Peng 2011, 1226.
- 62 Smaldino & McElreath 2016, 14.
- 63 Ks. Perkel 2020.
- 64 Fidler & Wilcox 2021.
- 65 Keränen & al. 2022. Ks. lyhyesti LUMI:n toimintaperiaatteista ja tehosta esim. Markomanolis 2022. Ks. myös CSC:n verkkosivut: csc.fi/tietoa-meista.
- 66 Esim. Heaven 2020.
- 67 Silfverberg 2021, 23; Peng 2011, 1226–1227.
- 68 Kapoor & Narayanan 2022.
- 69 Fidler & Wilcox 2021.
- 70 Silfverberg 2021, 23. Suomen Akatemian osalta, ks. verkkosivu: aka.fi/tutkimusrahoitus/vastuullinen-tiede/avoin-tiede/akatemian-linjaukset-avoimesta-tieteesta/.
- 71 Heaven 2020.
- 72 Smaldino & McElreath 2016, 14.
- 73 Alastalo, Parviainen & Choroszewicz 2022.
- 74 Ks. Parviainen 2022.
- 75 Heaven 2020.
- 76 Smaldino & McElreath 2016, 14.
- 77 Alanen 2022, 105.
- 78 Narayanan & Kapoor 2022; ks. alkuperäinen tutkimus, Sambasivan & Veeraraghavan 2022.
- 79 Alanen 2022, 105.
- 80 Vrt. Sama.
- 81 Fidler & Wilcox 2021; Plesser 2018, 3, 8–9.
- 82 Romero 2019, 2.
- 83 Sama, 5.
- 84 Jensen 2019.

Kirjallisuus

ACM, Artifact Review and Badging – Current. *acm.org* 24.8.2020. Verkossa: acm.org/publications/policies/artifact-review-and-badging-current.

Alanen, Paula, *Koulutusjärjestelmän ohjauskeinit ja numeerinen data yhteiskunnallisen kommunikaation muokkaajina. Tietämi-*

sen ja tietämättömyyden seuraukset aikuis-koulutuksen päätöksenteossa. Tampereen yliopisto, Tampere 2022. Verkossa: urn.fi/URN:ISBN:978-952-03-2413-1

Alastalo, Marja, Parviainen, Jaana & Choroszewicz, Marta, Tekoälytekniologian kotoistaminen julkisiin palveluihin: Tapaus Espoon tekoälykokeilu. *Yhteiskuntapolitiikka* 3/2022, 185–196. Verkossa:

urn.fi/URN:NBN:fi-fe2022060844676

Alpaydin, Ethem, *Machine Learning*. The MIT Press, Cambridge 2021.

Buckner, Cameron, *Deep Learning: A Philosophical Introduction. Philosophy Compass*. Vol. 14, No. 10, 2019, e12625.

Coveney, Peter V., Groen, Derek & Hoekstra, Alfons G., *Reliability and Reprodu-*

- cibility in Computational Science: Implementing Validation, Verification and Uncertainty Quantification *in silico*. *Philosophical Transactions A*. Vol. 379, 2021.
- DeLanda, Manuel, *Philosophy and Simulation. The Emergence of Synthetic Reason*. Bloomsbury, London 2015.
- Dreyfus, Hubert L., *What Computers Still Can't Do. A Critique of Artificial Reason*. The MIT Press, Cambridge 1992.
- Fidler, Fiona & Wilcox, John, Reproducibility of Scientific Results. *Stanford Encyclopedia of Philosophy*. Toim. Edward N. Zalta. 2021. Verkossa: plato.stanford.edu/archives/sum2021/entries/scientific-reproducibility
- Floridi, Luciano & Chiriatti, Massimo, GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, Vol. 30, 2020, 681–694.
- Gibney, Elizabeth, Is AI Fueling a Reproducibility Crisis in Science? *Nature*. Vol. 608, 2022.
- Haibe-Kains, Benjamin, ym., Transparency and Reproducibility in Artificial Intelligence. *Nature*. Vol. 586, 2020, E14–E16.
- Heaven, Will Douglas, AI is Wrestling with a Replication Crisis. *MIT Technology Review* 12.11.2020.
- Heikkilä, Melissa, Inside a radical new project to democratize AI. *MIT Technology Review* 12.7.2022. Verkossa: technologyreview.com/2022/07/12/1055817/inside-a-radical-new-project-to-democratize-ai
- Jensen, Alex, Replication as Success and Unsuccessful Replication. *University of Minnesota* 7.5.2019. Verkossa: cla.umn.edu/philosophy/story/replication-success-and-unsuccessful-replication
- Kapoor, Sayash & Narayanan, Arvind, Leakage and the Reproducibility Crisis in ML-based Science. *arXiv* 2022. Verkossa: arxiv.org/abs/2207.07048
- Kaufman, Shachar, Rosset, Sharon, Perlich, Claudia, Leakage in Data Mining: Formulation, Detection, and Avoidance. *Transactions on Knowledge Discovery from Data: Proceedings of the 17th ACM SIGKDD International Conference*. ACM, New York 2011, 556–563.
- Kelleher, John D., *Deep Learning*. The MIT Press, Cambridge 2019.
- Keränen, Timo, Takalo, Joni & Uusitalo, Kaisa, Suomessa otettiin käyttöön maailman kolmanneksi tehokkain tietokone. *Yle.fi* 13.6.2022. Verkossa: yle.fi/uutiset/3-12490372
- Koskinen, Inkeri, Objektiiivisuus humanistisissa tieteissä. *niin & näin* 4/2016, 35–42. Verkossa: netn.fi/artikkeli/objektiiivisuus-humanistisissa-tieteissa
- Koskinen, Inkeri, Millaisia aloja toistettavuuskriisi koskee? *Eufemia*-blogi 22.9.2021. Verkossa: blogs.tuni.fi/eufemia/teema1/millaisia-aloja-toistettavuuskriisi-koskee
- Krafczyk, Matthew S., Shi, A., Bhaskar, Adhithya, Marinov, D. & Stodden, Victoria, Learning from Reproducing Computational Results: Introducing Three Principles and the *Reproduction Package*. *Philosophical Transactions A*. Vol. 379, 2021.
- Littman, Michael L., Collusion Rings Threaten the Integrity of Computer Science Research. *Communications of the ACM*. Vol. 64, No. 6, 2021, 43–44. Verkossa: cacm.acm.org/magazines/2021/6/252840-collusion-rings-threaten-the-integrity-of-computer-science-research/fulltext#FNA
- Liu, Xiaoxuan, ym., A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digital Health*. Vol. 1, 2019, e271–e297.
- Louridas, Panos, *Algorithms*. The MIT Press, Cambridge 2020.
- Machery, Edouard, What is a Replication? *Philosophy of Science*. Vol. 87, 2020, 545–567.
- Markomanolis, George S., Utilizing AMD GPUs: Tuning, programming models, and roadmap. FOSDEM'22 HPC. 6.1.2022. Verkossa: fosdem.org/2022/schedule/event/utilizing_amd_gpus/attachments/slides/5163/export/events/attachments/utilizing_amd_gpus/slides/5163/fosdem22_hpc_amd_gpus_markomanolis.pdf
- McKinney, Scott Mayer ym., International evaluation of an AI system for breast cancer screening. *Nature*. Vol. 577, 2020a, 89–113.
- McKinney, Scott Mayer ym., Reply to: Transparency and Reproducibility in Artificial Intelligence. *Nature*. Vol. 586, 2020b, E17–E18.
- Narayanan, Arvind & Kapoor, Sayash, Why are deep learning technologists so overconfident? *AI Snake Oil* 31.8.2022. Verkossa: ainsakeoil.substack.com/pl/why-are-deep-learning-technologists?
- Nivala, Asko, Mitä on digitaalinen humanismi? *niin & näin* 1/2018, 63–69. Verkossa: netn.fi/node/7668
- Nivala, Asko, Onko tekoälyä olemassa? *niin & näin*, 3/2019, 21–22. Verkossa: netn.fi/node/7421
- Ojanen, Atte, Sahlgren, Otto, Vaiste, Juho, Björk, Anna, Mikkonen, Johannes, Kimppa, Kai, Laitinen, Arto & Oljakka, Nea, *Algoritminen syrjintä ja yhdenvertaisuuden edistäminen – Arviointikehikko syrjimättömälle tekoälylle*. Valtioneuvoston kanslia, Helsinki 2022. Verkossa: urn.fi/URN:ISBN:978-952-383-404-0
- Parviainen, Jaana, Läpinäkyvyys puuttuu julkisen sektorin tekoälyn kehitystyöstä: Voiko Alankomaiden tosalenaffaire toistua Suomessa? *Alusta!* 31.8.2022. Verkossa: tuni.fi/alustalehti/2022/08/31/lapinakyvyys-puuttuu-julkisen-sektorin-tekoalyn-kehitystyosta
- Parviainen, Jaana & Rantala, Juho, Chatbot breakthrough in the 2020s? An ethical reflection on the trend of automated consultations in health care. *Medicine, Health Care and Philosophy*. Vol. 25, 2022, 61–71.
- Peng, Roger D., Reproducible Research in Computational Science. *Science*. Vol. 334, 2011, 1226–1227.
- Perkel, Jeffrey M., The Digital Archaeologists. *Nature*. Vol. 584, 2020.
- Piller, Charles, Blots On a Field? *Science* 21.7.2022. Verkossa: science.org/content/article/potential-fabrication-research-images-threatens-key-theory-alzheimers-disease
- Pitkäniemi, Janne, Mailila, Nea, Tanskanen, Tomas, Degerlund, Henna, Heikkinen, Sanna & Seppä, Karri (toim.), *Syöpä 2020. Tilastoraportti Suomen syöpätalanteesta*. Suomen Syöpäyhdistys, Helsinki 2022.
- Plesser, Hans E., Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in Neuroinformatics*. Vol. 11, 2018.
- Rantala, Juho, Koodin ja ohjelmiston filosofia. *niin & näin* 1/2018, 101–108. Verkossa: netn.fi/node/7674
- Romero, Felipe, Philosophy of Science and The Replication Crisis. *Philosophy Compass*. Vol. 14, 2019.
- Rumelhart, David E., Hinton, Geoffrey E. & Williams, Ronald J., Learning Representations by Back-propagating Errors. *Nature*. Vol. 33, 1986, 533–536.
- Rusanen, Anna-Mari, Pikseleitä, kohinaa ja haurautta. *niin & näin* 3/2019, 47–53. Verkossa: https://netn.fi/node/7424
- Sambasivan, Nithya & Veeraraghavan, Rajesh, The Deskilling of Domain Expertise in AI Development. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM, New York 2022.
- Silverberg, Anu, Niin totta kuin osaamme. *Long Play*, 104, 2021.
- Smaldino, Paul E. & McElreath, Richard, The Natural Selection of Bad Science. *Royal Society Open Science*. Vol. 3, 2016.
- Stodden, Victoria, Krafczyk, Matthew S. & Bhaskar, Adhithya, Enabling the Verification of Computational Results. An Empirical Evaluation of Computational Reproducibility. *P-RECS'18: First International Workshop on Practical Reproducible Evaluation of Computer Systems*. ACM, New York 2018.
- Thunström, Almira Osmanovic, We Asked GPT-3 to Write an Academic Paper About Itself... *Scientific American* 30.6.2022. Verkossa: scientificamerican.com/article/we-asked-gpt-3-to-write-an-academic-paper-about-itself-mdash-then-we-tried-to-get-it-published
- Weizenbaum, Joseph, *Computer and Human Reason. From Judgment to Calculation*. W.H. Freeman & Company, New York 1976.