Puheenvuoro

Terveysdatan ja tekoälyn eettiset haasteet: Kriittinen sosio-tekninen näkökulma

JOHDANTO

Terveyttä ja hyvinvointia koskevaa dataa tuotetaan tutkimuksissa, terveys- ja hoitopalveluiden arkisissa toimissa sekä etenevässä määrin myös verkkoalustoilla ja yksityishenkilöiden mobiililaitteilla. Kiinnostus terveysdatan tehokkaampaan hyödyntämiseen on kasvanut viimeisen vuosikymmenen aikana, eikä hidastumisen merkkejä ole nähtävissä. Esimerkiksi Euroopan Parlamentti on viime vuosina rakentanut nk. eurooppalaista terveysdata-avaruutta, joka mahdollistaa kansalaisten omaehtoisen terveystietojen hallinnoimisen sekä terveysdatan tietoturvallisen jakamisen ja käytön esimerkiksi tutkimustarkoituksiin (1). Tekoälyn ja koneoppimisen edistysaskeleet siivittävät mainittua kehityskulkua, mahdollistaen monimutkaisten mallien rakentamisen laajojen terveysdatamassojen pohjalta. Tekoäly onkin osoittautunut hyödylliseksi muun muassa lääketieteellisessä kuvantamisessa ja riskitekijöiden tunnistamisessa. Viimeisimmät rintasyövän tunnistamiseen kehitetyt ennuste- ja luokittelumallit ovat esimerkiksi johdonmukaisesti saavuttaneet erinomaisen osumatarkkuuden (2). Generatiiviset tekoälysovellukset, kuten OpenAI:n suosittu ChatGPT-kielimallisovellus, ovat myös astumassa osaksi terveyspalveluita ja tutkimusta. Ne tuottavat käyttäjän syöttämän kuvauksen pohjalta esimerkiksi tekstiä tai kuvia, jotka vastaavat tai muokkaavat käyttäjän syötettä. ChatGPT:stä povataankin apuvälinettä esimerkiksi terveys- ja potilastietojen keräämiseen, tiivistämiseen ja analysoimiseen, lääkeannosteluun ja haittavaikutusten kuvailemiseen sekä potilaiden terveydentilan seuraamiseen (3).

Terveysdatan ja tekoälyn lisääntyvä käyttö sosiaali- ja terveydenhuollon palveluissa, tutkimuksessa sekä esimerkiksi yksityishenkilöiden mobiililaitteissa herättää lukuisia juridisia ja eettisiä kysymyksiä. Lainmukaisuuden näkökulmasta terveys- ja hyvinvointidatan keräämisessä ja käytössä edellytetään sovellusalueella toimintaa (esim. tutkimuksen toteuttamista, palveluiden järjestämistä) sääntelevien lakien ja säädösten huomioimista. Lisäksi henkilötietojen käsittelyä ja tekoälyn kehittämistä ja hyödyntämistä koskevat asetukset, kuten yleinen tietosuoja-asetus eli GDPR (4) ja tuore tekoälyasetus (5), tulee huomioida soveltuvin osin. Lainmukaisuus on kuitenkin vain ensimmäinen askel kohti terveysdatan vastuullista ja yhteiskunnallisesti kestävää käyttöä. Kaikkiin eettisiin haasteisiin ja dilemmoihin, joita väistämättäkin kohdataan terveystietoaineistoja kerätessä, mallinnettaessa ja prosessoitaessa, ei välttämättä löydetä ratkaisuja sen enempää lain pykälistä kuin teknologisista apuvälineistäkään. Täten on tärkeää myös ymmärtää suurien datamassojen ja tekoälyn käyttämiseen liittyviä eettisiä reunaehtoja ja yleisiä riskejä. Käyn seuraavaksi läpi tutkimuskirjallisuudessa tunnistettuja keskeisiä (juridis-)eettisiä riskejä sekä tapoja, joilla niihin on vastattu.

TERVEYSDATAN KERÄÄMINEN JA KONEELLINEN MALLINTAMINEN: KESKEISIÄ JURIDIS-EETTISIÄ RISKEJÄ

Terveysdata on lähtökohtaisesti aina arkaluontoista ja sijoittuukin esimerkiksi GDPR:n erityisen henkilötiedon kategorian alle. Tutkittavien henkilöiden, potilaiden ja esimerkiksi verkko- ja mobiilisovellusten käyttäjien yksityisyyttä ja tietosuojaa koskevat haasteet ovatkin herättäneet huolta (6). Esimerkiksi tutkimuspuolella tunnistetaan sosiaalisen median alustojen keräämän ja tuottaman datan hyödyt, mutta alustojen puutteellinen suostumuksenhallinta tuottaa haasteita tutkimusetiikan näkökulmasta (7). Lisähaasteita tietosuojan näkökulmasta tuottavat laajoilla datamassoilla opetettujen mallien tietosuojahaavoittuvaisuudet ja niiden väärinkäytön mahdollisuudet. Koneoppimisella tuotetuista malleista voidaan esimerkiksi erilaisin hyökkäysmenetelmin kaivaa

esiin tunnistettavaa tietoa yksilöistä (8) ja niiden oppimat korrelaatiot voivat paljastaa myös yllättävää, arkaluontoista tietoa kokonaisista kohorteista. Uuden tiedon tuottaminen on toki keskeinen päämäärä malleja rakennettaessa erityisesti tutkimuskonteksteissa, mutta riskinä on, että mallien oppimat korrelaatiot voidaan valjastaa myös ongelmallisiin tarkoituksiin, kuten profilointiin, kohdennettuun mainontaan tai jopa syrjintään (9). Tietosuojariskejä tuottavat perinteiseen tapaan myös esimerkiksi käyttäjät ja sovellusten varomaton käyttö. Esimerkiksi Australiassa lääkärit tuottivat ChatGPT:llä potilaskertomuksia nähtävästi tietämättään, että OpenAI voi säilyttää sovellukseen syötettyä tietoa ja käyttää sitä kielimallin jatko-opettamiseen (10).

Tekoälysovellusten käyttäminen kliinisissä ja palvelukonteksteissa – esimerkiksi terveyttä koskevien riskiarvioiden tai hoitosuositusten tuottamiseen – voi olla myös riskialtista sovellusten läpinäkyvyyden puutteen vuoksi. Laajoilla tietoaineistoilla opetettujen ennustemallien koosta ja monimutkaisuudesta johtuen niillä tuotettujen ennusteiden ja luokitteluiden totuusperäisyydestä, saati näiden taustalla olevista perusteista, ei ole usein varmuutta (11). Tekoälysovelluksista puhutaankin tästä syystä usein ”mustina laatikoina”: tiedämme, mitä menee sisälle ja mitä tulee ulos, mutta emme, mitä tapahtuu välissä. Läpinäkyvyyden puute saattaa muodostua esteeksi tekoälyn turvalliselle käytölle joissain tapauksissa. Se nostaa esille kuitenkin myös perustavanlaatuisia kysymyksiä yhtäältä terveyttä koskevien päätösten ja vallankäytön legitimiteetistä ja toisaalta päätöksenteon kohteiden, kuten potilaiden ja terveyssovellusten loppukäyttäjien, oikeuksien toteutumisesta (esim. ”oikeus selitykseen”, ks. GDPR:n (4) artiklat 13–15).

Yhdenvertaisuuden näkökulmasta erityistä huolta herättävät datan ”vinoumat” ja tekoälyn mahdolliset eriarvoistavat vaikutukset. Koneoppimiseen käytetyt datajoukot voivat olla paitsi vaillinaisia ja osittain virheellisiä myös vinoutuneita, jolloin mallien ennusteet ja suositukset voivat olla keskimäärin epätarkempia joidenkin ihmisryhmien kohdalla tai muutoin vahingollisia. Neljän markkinoillakin olevan kielimallin (Bard, GPT-4, ChatGPT ja Claude) osoitettiin esimerkiksi tuottavan ”rotupohjaiseen lääketieteeseen” perustuvia hoitosuosituksia (12). Mallit saattavat myös heijastaa historian saatossa tuotettua rakenteellista eriarvoisuutta ihmisryhmien välillä. Tätä havainnollistaa Yhdysvalloissa noin 200 miljoonaan ihmiseen vuosittain käytetty riskinarviointialgoritmi, jolla monet terveydenhuoltojärjestelmät kohdentavat potilaita ”korkean riskin hoidon” ohjelmiin tarjotakseen erityistä hoitoa sitä tarvitseville henkilöille. Riskiarviointialgoritmin ennusteiden osoitettiin systemaattisesti aliarvioivan mustien potilaiden hoidontarvetta verrattuna valkoisiin potilaisiin (13). Syynä oli se, että sen ennusteet kuvasivat potilaiden arvioituja terveydenhoitokustannuksia (eivät sairautta sinänsä), uusintaen olemassa olevia eroja hoidontarpeen arvioinnissa ja hoidon saavutettavuudessa mustan ja valkoisen väestön välillä.

Tekoälytuotteiden ja -palveluiden turvallisuus on myös herättänyt keskustelua. Koneoppimisella tuotetut mallit voivat käyttäytyä yllättävillä tavoilla ja niitä voidaan myös käyttää vahingollisiin tarkoitusperiin. Äärimmäisenä esimerkkinä väärinkäytön riskeistä voidaan mainita lääkekehitystä varten opetettujen mallien käyttö biokemiallisten aseiden suunnittelussa (14). Useimmiten turvallisuusriskejä luo kuitenkin yksinkertaisesti teknologian toimimattomuus. Paraatiesimerkkinä tästä toimivat sadat mallit, joita kehitettiin koronapandemian aikana COVID-19 -infektion tunnistamista ja prognoosia varten. Katsaukset näitä malleja esitteleviin tutkimuksiin kertovat hälyttävää tarinaa. Yhdessä systemaattisessa katsauksessa esimerkiksi todettiin, että suurin osa tarkastelluista tutkimuksista kärsi metodologisista ongelmista ja puutteellisesta dokumentaatiosta – yhdenkään esitellyistä malleista ei katsottu soveltuvan kliiniseen käyttöön (15). Toisessa katsauksessa löydökset olivat samansuuntaisia: tutkimuksissa esitetyt arviot mallien tarkkuudesta olivat kautta linjan liioiteltuja (16).

Monet edellä mainituista ongelmista ovat oireita syvemmistä haasteita, jotka ovat johtaneet koneoppimismenetelmiä hyödyntävän tieteen ”toistettavuuskriisiin” (17–18). Sadat tutkimukset eri tieteenaloilta, joissa koneoppimista on hyödynnetty kvantitatiiviseen analyysiin, sisältävät metodologisia ongelmia ja liioiteltuja väitteitä esiteltyjen mallien tarkkuudesta (19). Nämä ongelmat voivat murentaa tieteen uskottavuutta ja luotettavuutta ja kantautuvat väistämättä myös teknologiamarkkinoille asti. ”Tekoälyhypen” kyllästämillä teknologiamarkkinoilla potentiaalisen kuluttajan tai loppukäyttäjän, kuten tutkijan tai

terveyspalvelun tarjoajan, tuleekin pystyä erottamaan laadukas, tutkitusti toimiva teknologia käärmeöljyyn rinnastettavista työkaluista.

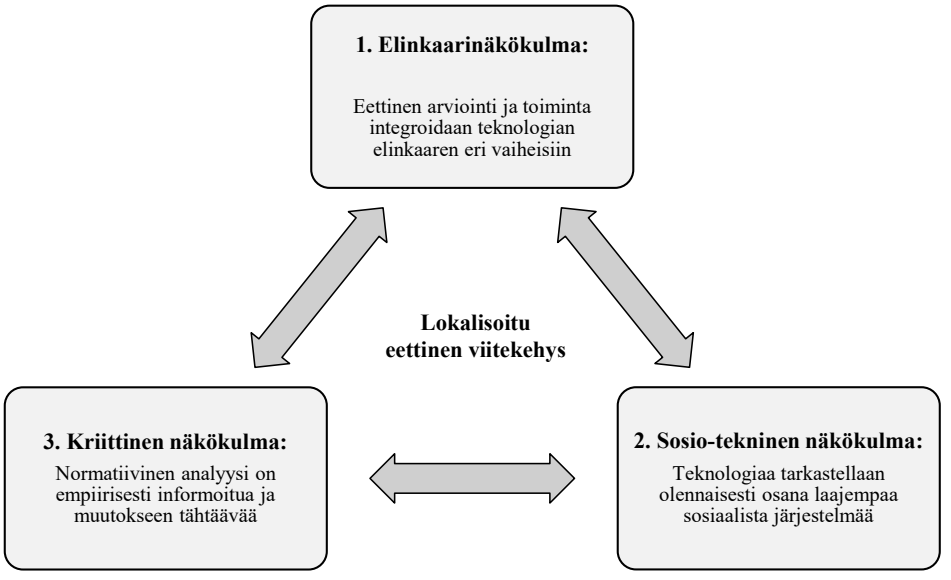
HAASTEITA VASTUULLISELLE TERVEYSDATAN JA TEKOÄLYN HYÖDYNTÄMISELLE

Edellä käsiteltyihin juridis-eettisiin ongelmakohtiin on vastattu eri keinoin. Lainsäädännöllisten keinojen, kuten tekoälysäädöksen (5), lisäksi mainitsemisen arvoisia ovat lukuisat eettiset ohjeistukset sekä eettisen vaikutustenarvioinnin ja riskienhallinnan menetelmät ja työkalut. Sadat eri toimijat teknologiayrityksistä kansallisiin asiantuntijaryhmiin ovat esittäneet eettisiä periaatteita data- ja tekoälyratkaisujen suunnittelulle, kehittämiselle ja käytölle. Useimmissa ohjeistuksissa esiintyy periaatteet, kuten läpinäkyvyys, oikeudenmukaisuus, vahingonteon välttäminen, vastuullisuus ja yksityisyyden ja autonomian kunnioittaminen (20), jotka yhtenevät terveyden ja hyvinvoinnin alueilta tuttujen bioetiikan periaatteiden kanssa (21). Tutkimusyhteisö on kehittänyt myös lukuisia arviointityökaluja ja menetelmiä, joilla nämä periaatteet voitaisiin viedä ”teoriasta käytäntöön”, samalla vastaten kentällä tunnistettuihin haasteisiin. Esimerkiksi ”mustan laatikon ongelmaan” on pyritty vastaamaan nk. selitysmenetelmillä, joilla kuvataan syitä tekoälyn tuottamille ennusteille katsomalla algoritmien ”konepellin alle” (11). Samaten lukuisia menetelmiä on tuotettu syrjivien vinoumien kvantifioimiseksi ja niiden poistamiseksi esimerkiksi opetusdataa, oppimisalgoritmia tai mallin tulosteita muokkaamalla (22). Etiikan tehokas jalkauttaminen osaksi käytäntöä on kuitenkin osoittautunut haasteelliseksi. Kuvaan seuraavaksi muutamia keskeisiä haasteita ja pohdintaluvussa hahmottelen näkemykseni siitä, mitä näihin haasteisiin vastaaminen edellyttää.

Ensimmäinen haaste koskee etiikan operationalisointia paikallisissa tutkimus-, palvelu-, tuotekehitys- ja käyttökonteksteissa. Tekoälyä koskevat eettiset ohjeistukset ovat usein paitsi normatiivisilta sisällöiltään kiistanalaisia myös turhan abstrakteja ja monitulkintaisia toimiakseen konkreettisina ohjenuorina toimijoille esimerkiksi datankeruussa tai tekoälysovellusten suunnittelussa (23–24). Teknologian välineellinen ja monikäyttöinen luonne luo omat haasteensa tässä suhteessa. Tietoaineistoja ja tekoälyä koskevat yleiset riskit ilmenevät nimittäin eri tavoin riippuen niiden käyttötavoista ja -konteksteista (vrt. tutkimus, sosiaali- ja terveyspalvelut, hyvinvointia tukevat mobiilisovellukset). Tämän takia riskejä ja vastuullisen käytön eettisiä reunaehtoja ei useimmiten voidakaan määritellä vaadittavalla tarkkuudella ilman, että spesifioidaan miten, missä ja mihin tarkoitukseen niitä tosiasiallisesti hyödynnetään.

Yksi tapa vastata tähän haasteeseen on ”lokalisoida” dataa ja tekoälyä koskevat yleiset eettiset periaatteet eli sovittaa ne tietoaineiston tai sovelluksen tosiasialliseen käyttökontekstiin ja sille ominaiseen arvojärjestelmään. Etiikkaa ei siis tule keksiä uusiksi, vaan soveltuvien lakien, säännösten, standardien sekä ammattieettisten säännöstöjen tulisi ohjata (terveys)dataa ja tekoälyä koskevien eettisten reunaehtojen tulkintaa ja operationalisointia. Esimerkiksi lääketieteen etiikka on teknologiasuunnitteluun ja -kehitykseen verrattuna laajalti institutionalisoitunut ja sillä on myös pitkä ammatillinen historia, vakiintunut normisto sekä joukko toimivaksi todettuja menetelmiä, joilla eettiset koodistot käännetään käytäntöön (24). Eri sovellusalueiden, kuten lääketieteen ja hoivatyön, vakiintuneet normit ja käytännöt (esim. suostumuksenhallintaa tai ihmisarvoista hoivaa koskien) antavat toimijoille välineitä reflektoida ja arvioida, toteutuvatko eettiset vaatimukset esimerkiksi, kun terveysdataa kerätään sosiaalisen median alustoilta, kun koneoppimista hyödynnetään tuottamaan tietoa riskiryhmistä tai kun generatiivisia malleja käytetään hoidon tai hoivatyön apuna.

Saatavilla olevien etiikkatyökalujen käyttö riskien tunnistamisessa ja hallinnoimisessa on osoittautunut myös haastavaksi erityisesti niiden kapean sovellettavuuden takia. Esimerkiksi tekoälyn vinoumien tunnistamista ja ehkäisemisestä varten kehitetyt teoreettiset viitekehykset ja ohjelmistotyökalut eivät ole vastanneet teknologiakehittäjien käytännön tarpeisiin (ks. (23)). Niiden varomaton käyttö voi myös johtaa negatiivisiin oheisvaikutuksiin, esimerkiksi ennustemallien kokonaistarkkuuden heikentymiseen tai uudenlaisten epätoivottavien vinoumien syntymiseen (25). Tarjolla olevat työkalut eivät myöskään ole sellaisenaan riittäviä varmistamaan yhdenvertaisuuden toteutumista tekoälyjärjestelmien tosiasiallisessa käytössä: monia yhdenvertaisuutta koskevia riskejä ei yksinkertaisesti pystytä määrittelemään, tunnistamaan, saati minimoimaan



Kuva 1. Kolme näkökulmaa datan ja tekoälyn eettisten riskien ja vaikutusten arviointiin ja hallintaan.

asianmukaisesti yksinomaan teknologian tasolla, sillä ne ovat riippuvaisia datan ja tekoälysovellusten käyttökontekstista (25–26). Sama pätee moniin tekoälysovellusten turvallisuusriskeihin: väärinkäytölle immuuni kielimalli esimerkiksi on ajatuksena yhtä mahdoton kuin tietokone, jota ei voitaisi käyttää vahingollisiin tarkoitusperiin (27). Pyrkimys tunnistaa ja hallita eettisiä riskejä yksinomaan teknologian tasolla ja kapein teknologisin keinoin voi pahimmillaan luoda lisää ongelmia. Parhaimmillaankin se tuottaa lähinnä laastarimaisia ratkaisuja, jotka eivät välttämättä poista taustalla piileviä, syvempiä ongelmia, jotka voivat liittyä esimerkiksi sosiaali- ja terveyspalveluiden resurssipulaan tai hoidon saavutettavuuseroihin alueiden ja ihmisryhmien välillä.

POHDINTA: KOLME TIETÄ ETEENPÄIN

Esitellyt haasteet eivät luonnollisestikaan tarkoita, etteikö terveys- ja hyvinvointidatan laajempi ja tehokkaampi hyödyntäminen olisi monessa tapauksessa toivottavaa tai etteivätkö tarjolla olevat etiikkatyökalut olisi vastuullisesti käytettynä hyödyllisiä rajallisiin tarkoituksiin. Ne puhuvat kuitenkin tarpeesta kokonaisvaltaisemmalle lähestymistavalle vastuulliseen terveysdatan hyödyntämiseen sekä eettiseen riskien arviointiin ja hallitsemiseen. Katson, että tällainen lähestymistapa edellyttää lokalisoidun eettisen viitekehyksen lisäksi kolmea toisiaan tukevaa näkökulmaa eettiseen arviointiin ja toimintaan (Kuva 1).

Ensimmäinen näistä on *elinkaarinäkökulma*, joka ulottaa eettisen harkinnan, vaikutuksenarvioinnin ja toiminnan vaatimuksen tarkasteltavan teknologian elinkaaren eri vaiheisiin (28). Elinkaarinäkökulman perusolettamus on, että eettistä harkintaa edellytetään tietoaineistojen ja teknologioiden elinkaarien jokaisessa vaiheessa – mukaan lukien useissa päätöksentekopisteissä, jotka tuppaavat jäämään dataa ja tekoälyä koskevan etiikkakeskustelun laitamille. Esimerkkinä voidaan mainita tekoälytuotteen elinkaaren aikaisessa vaiheessa tapahtuva laskennallisen ongelman matemaattinen määrittely (engl. *problem formulation*), jolla voi olla eettisesti merkittäviä seurauksia. Tuore tutkimus, joka tarkasteli vakavan sydänlihassairauden koneellista tunnistamista, osoitti esimerkiksi, että koneoppimismallin osumatarkkuus voi vaihdella huomattavasti riippuen siitä, määritelläänkö tunnistustehtävä regressio- vaiko luokittelutehtävänä (29). Yhtä lailla eettiselle vaikutuksenarvioinnille on kuitenkin tarvetta myöhäisemmissäkin vaiheissa elinkaarta – esimerkiksi päätöksissä koskien kerätyn terveys- tai potilastietoaineiston tuhoamista tai aiemmin tarjotun digitaalisen terveyspalvelun lopettamista.

Elinkaarinäkökulmaa täydentää nk. *sosio-tek-*

ninen näkökulma (30), jossa teknologiaa tarkastellaan osana laajempaa sosiaalista järjestelmää (esim. tietoaineistoa osana lääketieteellistä tutkimusprojektia, digitaalista järjestelmää osana sosiaali- tai terveyspalvelua). Teknologian ominaispiirteiden tarkastelun lisäksi tämä näkökulma korostaa tarvetta huomioida muun muassa teknologian sosio-materiaaliset ennakkoehdot, kuten teknologian asianmukaisen käytön edellyttämät muutokset toimintaympäristössä ja -kulttuurissa, sekä teknologian vuorovaikutus käyttäjien ja käyttöympäristön kanssa. Yllä mainittu tapausesimerkki australialaisista lääkäreistä osoitti, että kielimallien kontekstisidonnaisten riskien tunnistaminen edellyttää ymmärrystä paitsi kyseisen teknologian ominaisuuksista myös esimerkiksi loppukäyttäjien uskomuksista, toimintatavoista ja näitä sääntelevistä normeista.

Edellä mainitut näkökulmat kulkevat luontevasti yhdessä kolmannen, kriittisen näkökulman kanssa. Kriittisyys ymmärretään tässä yhteydessä kriittisen yhteiskuntateorian perinteen mukaisesti muutokseen tähtäävänä, empiirisesti informoituna normatiivisena analyysinä (31). Terveyden ja hyvinvoinnin alueilla tapahtuvan datafikaatio- ja automaatiokehityksen kontekstissa kriittisen analyysin tehtävä on tehdä sen epäkohtia näkyväksi. Eettisen arvioinnin polttopiste ei ole tällöin pelkästään tekoälyratkaisun (tosiasiallisissa ja kuvitelluissa) hyödyissä vaan myös sen rajoitteissa ja vaikutuksissa esimerkiksi haavoittuvaisiin ryhmiin. Teknologiaa ei lisäksi tarkastella yksinomaan eettisten ideaalien valossa vaan pyritään valottamaan arvokonflikteja ja muita jännitteitä, joita väistämättä syntyy, kun uusia tietoaineistoja kerätään tai uudenlaisia tietojärjestelmiä otetaan käyttöön (32). Kriittinen analyysi myös tarkastelee ja pyrkii aktiivisesti purkamaan eriarvoistavia rakenteita, kuten vallan epätasaista jakautumista sidosryhmien (esim. palveluntarjoajien, loppukäyttäjien ja potilasryhmien) välillä. Yhtäältä esimerkiksi uuden teknologian käyttöönottoa ajavien toimijoiden (esim. palveluntarjoajien, poliittisten päättäjien) taustaintressit pyritään tekemään näkyviksi (33). Toisaalta vastustetaan myös vallalla olevia diskursseja, joissa ”datavetoisuus” ja automaatio kehystetään usein yksinkertaisena ratkaisuna monisyisiin ongelmiin, kuten sosiaali- ja terveyspalveluiden tehokkuusvajeisiin tai kehnoon saavutettavuuteen.

KIRJALLISUUS

(1) Euroopan komissio. Eurooppalainen terveysdata-avaruus. Luettu 21.3.2024. https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_fi

(2) Nassif AB, Talib MA, Nasir Q, ym. Breast cancer detection using artificial intelligence techniques: A systematic literature review. Artif Intell Med 2022;127:102276. https://doi.org/10.1016/j.artmed.2022.102276

(3) Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595. doi: 10.3389/frai.2023.1169595

(4) Yleinen tietosuoja-asetus 27.4.2016/679. https://eur-lex.europa.eu/legal-content/FI/TXT/?uri=celex%3A32016R0679

(5) Ehdotus Euroopan parlamentin ja neuvoston asetus tekoälyä koskevista yhdenmukaistetuista säännöistä (tekoälysäädös) ja tiettyjen unionin säädösten muuttamisesta 2021/0106(COD), toimielinten välinen tiedosto 26.1.2024. https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf

(6) Price WN, Cohen IG. Privacy in the age of medical big data. Nat Med 2019; 25:37–43. https://doi.org/10.1038/s41591-018-0272-7

(7) Mittelstadt BD, Floridi L. The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. Sci Eng Ethics. 2016;22(2):303-41. doi: 10.1007/s11948-015-9652-2

(8) Liu B, Ding M, Shaham S, ym. When machine learning meets privacy: A survey and outlook. ACM Computing Surveys (CSUR) 2021;54(2):1-36. https://doi.org/10.1145/3436755

(9) Wachter S, Mittelstadt B. A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. Colum Bus L Rev 2019;2. https://ssrn.com/abstract=3248829

(10) The Guardian. AMA calls for stronger AI regulations after doctors use ChatGPT to write medical notes 27.7.2023, luettu 22.2.2024. https://www.theguardian.com/technology/2023/jul/27/chatgpt-health-industry-hospitals-ai-regulations-ama

(11) Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy 2021;23(1):18. https://doi.org/10.3390/e23010018

(12) Omiye JA, Lester JC, Spichak S, ym. Large language models propagate race-based medicine. npj Digit Med 2023;6:195. https://doi.org/10.1038/s41746-023-00939-z

(13) Obermeyer Z, Powers B, Vogeli C, ym. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2016;366(6464):447-453. doi: 10.1126/science.aax2342

(14) Urbina F, Lentzos F, Invernizzi C, ym. Dual use of artificial-intelligence-powered drug discovery. Nat Mach Intell, 2022;4(3):189-191. https://doi.org/10.1038/s42256-022-00465-9

(15) Roberts M, Driggs D, Thorpe M, ym. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell 2021;3(3):199-217. https://doi.org/10.1038/s42256-021-00307-0

(16) Wynants L, Van Calster B, Collins GS, ym. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020:369;m1328. https://doi.org/10.1136/bmj.m1328

(17) Hutson M. Artificial intelligence faces reproducibility crisis. Science 2018;359(6377):725–726. https://doi.org/10.1126/science.359.6377.725

(18) Rantala J, Alanen P, Parviainen J. Mitä toistettavuusongelmat tuovat tullessaan tekoälytutkimukseen?. *Niin & Näin* 2022;4:47–58.

(19) Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns 2023;4(9). https://doi.org/10.1016/j.patter.2023.100804

(20) Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell 2019;1(9):389-399. https://doi.org/10.1038/s42256-019-0088-2

(21) Floridi L, Cowls J. A unified framework of five principles for AI in society. Harvard Data Science Review 2019;1(1). https://doi.org/10.1162/99608f92.8cd550d1

(22) Mehrabi N, Morstatter F, Saxena N, ym. A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) 2021;54(6):1-35. https://doi.org/10.1145/3457607

(23) Sahlgren O. Action-guidance and AI ethics: the case of fair machine learning. AI and Ethics, 2024:1-13. https://doi.org/10.1007/s43681-024-00437-2

(24) Mittelstadt B. Principles alone cannot guarantee ethical AI. Nat Mach Intell 2019;1(11):501-507. https://doi.org/10.1038/s42256-019-0114-4

(25) Balayn A, Gürses S. Beyond Debiasing: Regulating AI and its inequalities. EDRi:n raportti 2021. https://edri.org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online.pdf

(26) Ojanen A, Sahlgren O, Vaiste J, ym. Algoritminen syrjintä ja yhdenvertaisuuden edistäminen: Arviointikehikko syrjimättömälle tekoälylle. Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja 2022:54. http://urn.fi/URN:ISBN:978-952-383-404-0

(27) Narayanan A, Kapoor S. AI safety is not a model property. AI Snake Oil -verkkoblogi 12. maaliskuuta 2024. Luettu 18.3.2024. https://www.aisnakeoil.com/p/ai-safety-is-not-a-model-property

(28) Sahlgren O. Ethics in the AI Lifecycle: Theoretical Perspectives, Practical Resources and Recommendations. 2023. https://urn.fi/URN:ISBN:978-952-03-2777-4

(29) Vrudhula A, Hughes JW, Yuan N, ym. The Impact of Task Set-up in Algorithm Design: Regression versus Classification. NEJM AI, AIcs 2024;2300176. DOI: 10.1056/AIcs2300176

(30) Ropohl G. Philosophy of socio-technical systems. Society for Philosophy and Technology Quarterly Electronic Journal 1999;4(3):186-194. https://doi.org/10.5840/techne19994311

(31) Celikates R, Flynn J. Critical Theory (Frankfurt School). The Stanford Encyclopedia of Philosophy, Edward N. Zalta & Uri Nodelman (toim.) 2023. Luettu 20.3.2024. https://plato.stanford.edu/archives/win2023/entries/critical-theory/

(32) Whittlestone J, Nyrup R, Alexandrova A, ym. The role and limits of principles in AI ethics: Towards a focus on tensions. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society 2019:195-200. https://doi.org/10.1145/3306618.3314289

(33) Parviainen J, Rantala J. Ennakoiva analytiikka ja tekoälyn etiikka: Miten ennakoivat teknologiat taipuvat hallintajärjestelmäksi?. Futura 2020;39(1):61–70

Otto Sahlgren
Filosofian väitöskirjatutkija
Tampereen yliopisto