Anna-Mari Wallenberg

Tekoälyn etiikka: Hyödytön, hampaaton vai keskeneräinen?

Älykkäät teknologiat nostavat esiin mutkikkaita kysymyksiä niiden yhteiskunnallisista ja eettisistä vaikutuksista. Monimutkaiset, aiempaa itsenäisemmin oppivat tekoälyjärjestelmät ovat kognitiivisesti aktiivisia tavoilla, jotka erottavat ne muusta teknologiasta. Tekoälyjärjestelmät voivat muokata käytäntöjä, rakenteita ja toimintaympäristöä tavoilla, jotka nostavat esiin uudenlaisia kysymyksiä oikeusvaikutuksista, oikeudenmukaisuudesta ja hyvinvoinnin jakautumisesta. Perinteinen teknologian arviointi toimivuuden, käytettävyyden, tehokkuuden tai teknisen luotettavuuden näkökulmasta ei riitä. Sen sijaan lisääntyvän tekoälyn käyttö herättää tarpeen sen eettisestä arvioinnista. Viime vuosina tekoälyn etiikkaa, ja erityisesti tekoälyn eettisiä ohjeistuksia, on kuitenkin arvosteltu hyödyttömyydestä ja hampaattomuudesta. Kirjoituksessa ehdotetaan, että kyse on pikemminkin etiikkatyön keskeneräisyydestä.

AVAINSANAT: Tekoälyn etiikka, algoritmit, teknologian arviointi

Miksi tekoälyn etiikasta puhutaan?

Keinotekoisiin, ajatteleviin tai älykkäisiin koneisiin on liitetty paljon odotuksia ja pelkoja. Filosofi Gottfried Leibniz haaveili 1700-luvulla “universaalista laskukoneesta”, joka olisi yhteydessä yleiseen, kaiken tieteellisen tiedon sisältävään tieteen ensyklopediaan. Kone kykenisi päättelemään vastaukset mihin tahansa kysymykseen (Roinila, 2021). Pari sataa vuotta myöhemmin chatGPT konkretisoi sekä Leibnizin haavetta, että sen ongelmia. Kone kyllä päättelee vastauksia, mutta ei osaa arvioida, pitävätkö ne paikkansa.

ChatGPT, ja generatiivinen tekoäly ylipäänsä, osoittaa myös, miksi realistinen keskustelu teknologiasta on tärkeää. Se havainnollistaa, kuinka nopeasti ja räjähdysmäisesti algoritmiset teknologiat voivat arkipäiväistyä. Vielä vuosi sitten suurin osa ihmisistä ei ollut edes kuullut koko teknologiasta. Nyt kielimallit ja kuvaeditorit ovat internetyhteyden päässä.

Älykkäät teknologiat nostavat esiin hankalia kysymyksiä niiden yhteiskunnallisista ja eettisistä vaikutuksista. Monimutkaiset, koko ajan itsenäisemmin oppivat tekoälyjärjestelmät ovat kognitiivisesti aktiivisia tavoilla, jotka erottavat ne muista teknologioista. Tekoälyjärjestelmät voivat muokata käytäntöjä, rakenteita ja toimintaympäristöä. Ne nostavat esiin uudenlaisia kysymyksiä oikeusvaikutuksista, oikeudenmukaisuudesta, ja hyvinvoinnin jakautumisesta.

Perinteinen teknologian arviointi toimivuuden, käytettävyyden, tehokkuuden tai teknisen luotettavuuden näkökulmasta ei riitä. Viime vuosina onkin toistuvasti todettu, että lisääntyvästä tekoälyn käytöstä seuraa tarve sen eettiselle arvioinnil-

le. On puhuttu paljon tekoälyn etiikasta ja siitä, mitä tekoälyn eettisesti hyväksyttävä kehittäminen, käyttäminen ja hyödyntäminen edellyttävät.

Tekoälyn etiikan lyhyt historia: eettiset ohjeistukset

Kun kymmenisen vuotta sitten kone-oppimiseen (ML) ja syväoppiviin verkkoihin (DNN) perustuvat kaupalliset sovellukset yleistyivät, nousivat esiin huolet niihin liittyvistä riskeistä ja mahdollisista haitoista. Eettisiä ohjeita ja periaatteita alettiin laatia, koska uskottiin, että niiden avulla haittoja voitaisiin ehkäistä ennalta. Toivottiin, että esitetyt periaatteet toimisivat abstraktioina, joiden pohjalta algoritmien kehitystä ja käyttöä voitaisiin ohjata ja rajoittaa.

Nykyinen tekoälyn etiikka ei varsinaisesti perustu akateemiselle tutkimukselle, vaan on syntynyt näiden ohjeistusten ja periaatteiden pohjalta. Ohjeistukset ovat rajanneet ja synnyttäneet viitekehyksen, jonka sisällä viime vuosien keskustelu tekoälyn etiikasta on pitkälti käyty. Ohjeistukset ovat valikoineet käsitteet, kuten “läpinäkyvyys” tai “vastuullisuus”, joilla tekoälyn etiikasta keskustellaan sekä sanallistaneet periaatteet, joiden nojalla tekoälysovellusten hyväksyttävyyttä edelleen arvioidaan.

Tällä hetkellä ohjeistuksia on jo yli 200. Niitä ovat julkaisseet niin kansalliset (esim. Kanadan hallitus) kuin monikansalliset organisaatiot (Euroopan komissio, OECD, Unesco), yleishyödylliset toimijat ja järjestöt (mm. IEEE), erilaiset virastot ja yksiköt (Suomessa esim. Vero ja Kela, Helsingin kaupunki), akateemiset (Future of Life Institute) ja uskonnolliset yhteisöt (mm. katolinen kirkko) sekä yritykset (mm. Microsoft, Google, Facebook).

Suositusten laatimisprosesseja ei ole juurikaan tutkittu. Tiedetään, että tyypillisesti suositukset on laadittu ns. ad hoc -työryhmissä (esim. EU:n High Level Expert Group). Ryhmistä, niiden kokoonpanosta tai työskentelystä ei kuitenkaan ole kokonaiskuvaa. Ryhmien tehtävänannoista, käsiteltyjen aiheiden valikoinnista tai työtavoista ei myöskään ole tutkimustietoa. Ei myöskään tiedetä, kuinka ryhmissä on hahmotettu tekoälyohjeistusten tehtävä, onko ja jos on, niin miten mahdollinen laadunvarmistus on tehty, tai missä määrin ryhmissä on keskusteltu, vastaavatko ohjeistukset tosiasiallisia tekoälyn käyttöön liittyviä haasteita.

Ohjeistuksien laatu vaihtelee lisäksi varsin paljon. Osa sisältää heikosti toisiinsa kytkettyjä yksittäisiä sääntöjä. Niissä käytettyjä käsitteitä ei määritellä, taustoja ei avata, eikä periaatteiden välillä ole ilmeistä yhteyttä. Osa taas on varsin korkealaatuisia ja hyvin muodostettuja, toisiinsa kytkettyjä ja jäsenneltyjä periaatekokoelmia. Ohjeita on esimerkiksi perusteltu laajoilla, osittain asiantuntijatyönä koostetuilla yksityiskohtaisilla ja monitieteisillä selvityksillä, joiden laatimisprosessi on myös jossain määrin julkisesti raportoitu (esim. Unesco).

Eettisten ohjeistusten sisältö: Big Five

Ohjeistuksien sanasto ja painotukset vaihtelevat jonkin verran. Pinnallisista eroista huolimatta niissä kuitenkin käsitellään pitkälti samoja perusteemoja. Tavoiteltavina tekoälyn hyödyntämisen piirteinä useimmissa ohjeistuksissa nähdään mm. myönteinen demokratiakehitys, yhteiskunnallinen oikeudenmukaisuus ja yksilönsuoja (mm. Jobin, 2019; Morley ym., 2019). Ohjeistuksissa korostetaan ihmiskeskeisyyttä, luotettavuutta ja vastuullisuutta keskeisinä hyveinä. Niissä toistuu myös etiikan paikallisuus, eikä esimerkiksi tekoälykehityksen ilmastovaikutusten, globaalien tuotantoketjujen tai hyvinvoinnin jakautumisen kysymyksiä mainitse kuin kourallinen suosituksista (Wallenberg, valmisteilla).

Ohjeistuksissa heijastuu myös monia perinteisen eurooppalaisen yhteiskunta- ja moraalifilosofisen ajattelun piirteitä. Painotus ilmenee niin käytetyissä käsitteissä, keskeisissä päättelyissä kuin teemojen valikoinnissakin. Ohjeistuksissa esimerkiksi tulkitaan perusoikeudet länsimaisittain, ja korostetaan länsimaisille demokratioille keskeistä ajatusta valtion vallankäytön rajoittamisesta kansalaisen oikeuksien suhteen. Kansalais- ja yksilöoikeuksien painotus näkyy myös siten, että yhdenvertaisuus- ja omistusoikeuksia (esim. dataa koskeva omistusoikeus), psyykkistä ja fyysistä koskemattomuutta (algoritminen manipulaatio) sekä tieto- ja yksilönsuojaa tulkitaan pääsääntöisesti yksilöiden oikeuksien näkökulmasta.

Ohjeistuksille on leimallista myös niiden riskipohjaisuus. Tavoitteena on lähes aina ennaltaehkäistä harmeja ja haittoja. Tekoälyn mahdollisia myönteisiä vaikutuksia ei mainita kuin muutamassa ohjeistuksessa (Wallenberg valmisteilla; Rusanen, 2019). Tämä heijastuu myös ohjeistuksissa käytettyjen käsitteiden tulkintaan. Esimerkiksi perusoikeudet, kuten yhdenvertaisuus tai yksilönsuoja, tulkitaan ohjeistuksissa pääsääntöisesti ns. negatiivisesti. Painopiste on oikeuksiin kohdistuvien loukkauksien, kuten yhdenvertaisuutta rikkovan syrjinnän ennaltaehkäisyssä, ei oikeuksien toteutumista tosiasiallisesti tukevien velvoitteiden asettamisessa (ns. positiivinen tulkinta).

Taulukko 1. Tekoälyn etiikan keskeiset periaatteet (Jobin ym., 2019).

1. vahinkojen välttäminen, eli “älä aiheuta harmia-periaate” (Riskien ennaltaehkäisy).
2. vastuullisuus, eli “kuka kantaa vastuun tekoälyn toiminnasta” (Vastuun kantilainen tulkinta.)
3. läpinäkyvyys ja selitettävyys, eli “miten tiedetään, kuinka tekoäly toimii” (Seuraa sekä vastuullisuudesta että yksilön oikeudesta saada selitys häntä koskevan päätöksen perusteluista.)
4. oikeudenmukaisuus ja yhdenvertaisuus eli “tekoälyn tulee toimia reilusti, eikä se saa syrjiä”
5. ihmisoikeuksien, kuten yksityisyyden ja turvallisuuden, kunnioittaminen.

Suositusten peruskäsitteet rakentuvat pitkälti eurooppalaisen moraalifilosofian periaatteiden päälle. Esimerkiksi vastuullisuus (engl. accountability) tulkitaan ohjeistuksissa ns. perinteisen kantilaisen luennan pohjalta. Vastuu on aktiivisen toimijan vastuuta teoista ja valinnoista. Vastuu edellyttää toimijalta sekä valinnanvapautta että tiedollista kykyä ennakoida valintojen seurauksia. Toimija on vastuussa vain teoista (tai tekemättä jättämisistä), jotka hän valitsee vapaasti ja joiden kohdalla hän on kyennyt arvioimaan, mitä valitusta teosta voi seurata.

Ohjeistuksissa tiedollinen ennakointivelvoite tulkitaan usein järjestelmien läpinäkyvyytenä ja selitettävyytenä. Päättely etenee seuraavasti: Jos järjestelmän on oltava ennakoitava, sen toiminta on osattava selittää. Jotta järjestelmän toimintaa voidaan selittää, on sen oltava läpinäkyvää (esim. Morley ym., 2019). Toiminta on puolestaan läpinäkyvää, jos se voidaan kuvata riittävällä tarkkuudella askel askeleelta. Tästä tulkinnasta on kuitenkin keskusteltu varsin paljon viime vuosina. On esimerkiksi todettu, että aito läpinäkyvyys voi myös vaatia kuvauksen ymmärrettävyyttä. Sen mukaan myös henkilöiden, joilla ei ole tietoteknistä koulutusta, on kyettävä hahmottamaan, miten heitä koskevat päätökset on koneellisesti tehty. Lisäksi on huomautettu, että selitettävyys ei välttämättä tarkoita samaa kuin yksilön oikeusturvan kannalta keskeinen läpinäkyvyys. Päätöksenteon kohteena olevan yksilön kannalta keskeisempää on usein päätöksen oikeutus kuin sen selitettävyys (Robbins, 2019). Jos vaikkapa sosiaalietuus tai oleskelulupa evätään koneellisesti (tai ihmisen toimesta), yksilölle ei ole keskeistä se, millaisilla algoritmeilla (tai hermosoluilla) päätös on tehty. Keskeisempää on, perustellaanko päätös oikeilla pykälillä tai laintulkinnalla (Wachter ym., 2018; Robbins, 2019).

Myös vinouman käsitteen tulkinta on herättänyt arvostelua. Julkisissa keskusteluissa usein oletetaan, että monille koneoppiville järjestelmille, kuten luokittelualgoritmeille määritelmällinen, niiden matemaattisista ominaisuuksista johtuva vinoumaherkkyys on automaattisesti moraalisessa mielessä syrjivää, tai johtaa syrjintään. Oletus toistuu monissa ohjeistuksissa. Vinoumia on kuitenkin useita eri lajeja, ja ne voivat johtua monista eri syistä (Belenguer, 2022). Kaikki vinoumat eivät syrji, sillä aidosti syrjivien vinoumien on oltava moraalisesti merkityksellisiä esimerkiksi perusoikeusvaikutuksien vuoksi (Rusanen, 2020). Ne tyypillisesti ilmenevät käyttöyhteyksissä esimerkiksi mahdollisena epätasa-arvoisena kohteluna tai oikeuksien eväämisenä (Belenguer, 2022).

Vinoumaherkkyys ja läpinäkyvyys ovat esimerkkejä siitä, kuinka ohjeistuksissa ei ole aina ajateltu asioita aivan loppuun asti. Pinnallisuus näkyy myös epämääräisyytenä, eli ”tekoäly ei saa syrjiä”- tyylisinä ohjeina, jotka eivät kiinnity mihinkään. Pahimmillaan ohjeistukset voivat olla sisäisesti ristiriitaisia. Saatetaan esimerkiksi kieltää yksilöiden erotteleminen luokittelualgoritmeilla syrjinnän estämiseksi yhdessä periaatteessa. Seuraavassa voidaan kuitenkin teknologian inklusiivisuuden (eli erityispiirteiden huomioimiseen perustuvan teknologian) lisäämiseksi vaatia kuitenkin sitä, että luokittelualgoritmien tulee tunnistaa ja huomioida yksilöiden erot.

Ohjeistuksiin liittyy myös muita ongelmia. Eräs on niiden ruokkima virhetulkinta, että kyse olisi aivan uudesta, muusta moraalikoodistosta irrallisesta, vain tekoälyä koskevasta säännöstöstä. On jäänyt huomaamatta, että useimmat ohjeistusten teemoista, kuten yhdenvertaisuus, oikeudenmukaisuus tai vastuullisuus, eivät varsinaisesti ole tekoälyspesifejä, vaan pätevät mihin tahansa eettiseen arviointiin (Floridi ym., 2018). Seurauksena on mielikuva, että vain tekoälyn etiikka asettaisi eettisiä velvoitteita tekoälyn hyväksyttävälle käytölle. Mielikuva johtaa epärealistisiin odotuksiin tekoälyn etiikan tehtävistä ja synnyttää turhautumista, kun odotuksia ei kuitenkaan ohjeistuksien tasolla täytetä.

Turhautuminen tekoälyn etiikkaan?

Ohjeistuksia onkin viime vuosina arvosteltu ankarasti, että niiden tosiasiallinen vaikutus on suhteellisen heikkoa. Ohjeistuksia on kuvattu mm. abstrakteiksi periaatelistoiksi, jotka eivät kiinnity käytäntöihin tai ohjaa niitä (McNamara ym., 2018; Hagendorff, 2020). Ohjeistuksia on kutsuttu “hyödyttömiksi” ja tekoälyn etiikkaa “hampaattomaksi” (Resseguier ym., 2020; Munn, 2022). Ongelma usein on, että ohjeet eivät itsessään yksilöi konkreettisia mekanismeja, työkaluja tai välineitä niiden toteuttamiseksi (Morley ym., 2019).

Abstraktien periaatteiden kytkeminen käytäntöön on haasteellista, olipa kyse mistä alasta tahansa. Normatiivisia sääntöjä on vaikeaa muuntaa teknistä kehitystyötä tai käyttöä ohjaaviksi periaatteiksi siten, että ne säilyttäisivät normatiivisen luonteensa. Tekoälyn eettisten ohjeistusten epämääräisyys tekee muuntamisesta vielä vaikeampaa. Yritykset jalkauttaa eettiset ohjeistukset “käännöstyökalujen” avulla ovatkin usein epäonnistuneet (Morley ym., 2021).

Ei kuitenkaan voida sanoa, että eettiset ohjeistukset olisivat olleet täysin turhia. Ohjeistukset ovat antaneet alustavan sanallisen ilmiasun toiveille, joita tekoälyn kehitykseen ja käyttämiseen liitetään (Morley ym., 2019). Niiden pohjalta on myös viime vuosina kehitetty lukemattomia erilaisia toimintamalleja, standardointialoitteita, etiikkatyökaluja ja arvioinnin menetelmiä kansalaispaneeleista riskien arviointilomakkeisiin ja koulutusohjelmiin (Kinder ym., 2023). Eettiset teemat ovat myös jatkuvasti otsikoissa, ja niihin liittyvä yleinen osaamistaso on noussut valtavasti (Morley ym., 2021). Myös moni merkittävä sääntelyaloite, kuten EU:n tekoälysäädös (AI Act) on hakenut inspiraatiota eettisistä ohjeistuksista.

Lisäksi tekoälytutkimuksen ja -tuotekehityksen parissa on jo nyt paljon alueita, joilla on ainakin välillisesti edistytty eettistä päämääriä tukevien teknisten ratkaisujen kehittämisessä. Esimerkiksi algoritmista reiluutta lisääviä ratkaisuja on kehitetty useita, ja ala on vakiinnuttanut asemansa tutkimuskentällä (Ratra ym., 2021; Wang ym., 2022). Algoritminen yksityisyyden- ja tietosuoja on myös ottanut valtavia harppauksia nimenomaan teknisten ratkaisujen vuoksi (Xu ym., 2021). Viime vuosina on myös kehitetty aktiivisesti erilaisia teknisiä keinoja koneilla generoidun väärän tiedon tunnistamiseksi

(Naitali ym., 2023). Hiljattain on edistytty myös kielimallien hallusinointimenetelmien kitkemisessä (Li ym., 2022).

Teknosolutionismia?

Paradoksaalista sinänsä, mutta moni yhteiskuntatieteissä suositun ns. kriittisen teknologiatutkimuksen edustaja näkee kuitenkin teknisten ratkaisujen kehittämisen paheksuttavana "teknosolutionismina". Teknosolutionismia syytetään siitä, että eettiset ongelmat tulkitaan "vain" teknisinä ongelmina, joita yritetään ratkaistaan teknisten pikaratkaisujen avulla. Monimutkaiset, rakenteelliset yhteiskunnalliset ongelmat "häivytetään" teknologisen keskeneräisyyden taakse, ja niitä ratkaistaan puuttuvina lisäosina, suunnittelu- tai ohjelmointivirheinä.

Kritiikin mukaan tämä ohjaa keskittymään "liian kapeasti" läpinäkyvyyden, hallusinoinnin ja vinoumien kaltaisiin "prosessuaalisiin kysymyksiin" ja sivuuttamaan teknologian "sosiotekninen konteksti" (Zalnieriute, 2021). Sosioteknisellä kontekstilla tarkoitetaan tässä yhteydessä "yhteiskunnallisia valta-, rooliodotus-, identiteetti- ja sortorakenteita", joissa teknologiaa tuotetaan, käytetään ja hyödynnetään (Zalnieriute, 2021).

Näkökulma on kuitenkin turhan yksipuolinen. On vaarallisen lähellä, että väitteissä sosioteknisen kontekstin sivuuttamisesta, se tosiasiassa sivuutetaan itse. Teknologian käyttöympäristöt ovat aina kerroksellisia ja moniulotteisia. Teknologiaa ei käytetä, kehitetä tai hyödynnetä vain valtarakenteissa, identiteetti- tai rooliodotuksissa. Myös muut, aivan yhtä merkitykselliset tekijät, kuten juridinen normisto, informaatioympäristö, ihmisen ja koneen välinen tiedollinen vuorovaikutus tai siihen perustuva työnjako, määrittävät ja rakentavat sosioteknis(-kognitiivista) kontekstia (Wallenberg, valmisteilla). Eettisten ohjeistusten tehtävä voikin olla sosioteknisen normatiivisen kontekstin ylläpito, ei sen sivuuttaminen.

Eettisten ohjeistusten rooli julkisessa hallinnossa

Monilla tekoälyn käyttöalueilla, kuten julkisessa hallinnossa, on jo omat eettiset ja juridiset norminsa. Ne ovat keskeisiä julkiselle hallinnolle sosioteknisenä kontekstina. Tekoäly tai sen etiikka eivät itsessään kumoa, määrittele tai muuta julkisen hallinnon normiperustaa, sen tavoitteita tai ideaaleja.

Jos julkisessa hallinnossa tai viranomaistyössä käytetään tekoälyteknologiaa, sitä koskee aivan samat säännöt ja periaatteet kuin kaikkea muutakin viranomais- tai hallintotoimintaa. Teknologia ei itsessään muuta sitä, että viranomaisen vastuulla on esimerkiksi toimia puolueettomasti, riippumattomasti ja tasapuolisesti. Ne eivät myöskään vaikuta siihen, että viranomaisen toiminnan on oltava avointa, tarkoituksenmukaista ja hyväksyttävää. Eivätkä ne kyseenalaista sitä, että viranomaisen on turvattava asianosaisten oikeusturva sekä virkavastuun toteutuminen myös silloin, kun algoritmeja käytetään.

Eettisten ohjeistusten tehtävä on tulkita, kuinka hyvän hallinnon periaatteet voidaan toteuttaa käytettäessä tekoälyä. Kyse ei siis ole sosioteknisen kontekstin häivyttämisestä, vaan sen täydentämisestä. Tekoäly voi esimerkiksi muokata hallintotoiminnan rakenteita ja käytäntöjä tavoilla, joita aiempi normisto ei tunnista. Tekoälyn käytöllä voi myös olla sellaisia välillisiä tai välittömiä vaikutuksia viranomaistoimintaan, joita aiemmilla tietoteknisillä laitteilla ei ole ollut. Eettisten ohjeistuksien tehtävä onkin taata, että tekoälyjärjestelmiä kehitetään ja käytetään olemassa olevan normipohjan kanssa yhteensopivalla tavalla.

Jos tekoälyn etiikan rooli nähdään ylläpitävänä ja täydentävänä, on luontevaa, että ohjeistuksien avulla ohjataan etsimään vastauksia "prosessuaalisiin kysymyksiin". Ohjeistuksien tehtävä ei ole kyseenalaistaa koko normipohjaa, vaan ohjata teknistä kehittämistä ja teknologian käyttöä niin, että niiden varassa tapahtuva viranomaistoiminta noudattaa hyvän hallinnon periaatteita. Ohjeistusten tehtävä on rakentaa viitekehys, jolla hallinnon periaatteet voidaan operationalisoida muotoon, joita voidaan soveltaa tekoälyn käytössä ja kehittämisessä.

Ovatko ohjeistukset onnistuneet tässä? Onko ohjeistuksia pystytty operationalisoimaan niin, että ne ovat – edes välillisesti – auttaneet esimerkiksi parantamaan julkisten palveluiden laatua?

Kysymykset eivät ole aivan yksinkertaisia, eikä niihin ole tyhjentäviä vastauksia. Kyse ei ole siitä, etteikö julkista hallintoa ja sen tekoälyn käyttöä olisi tutkittu ja selvitetty. On kymmeniä, ellei satoja raportteja ja tutkimuksia tekoälyn hyödyntämiseen liittyvistä odotuksista, eettisen tekoälyn hyödyntämisen malleista, niiden pilottikokeiluista ja vaikutuksista (Zuiderwijk ym., 2021). Tutkimukset ja selvitykset tarkastelevat kuitenkin tekoälyn käytön seurauksia ja tekoälyn käyttöön liittyviä riskejä usein esimerkiksi osallisuuden ja luottamuksen, niiden rakentumisen ja niihin kohdistuvien odotusten näkökulmista (Henman, 2020; Desouza ym., 2019). Eettisten ohjeistuksien operationalisointikysymyksiä ei kuitenkaan ole tutkittu, eikä kokonaiskuvaa ole (Gianni ym., 2022).

Aihepiiriä mutkistaa myös, että julkinen hallinto ei ole monoliitti, vaan monikerroksinen rakenne. Useissa maissa julkisen hallinnon eri kerroksilla on omat lakisääteiset toimialueensa, ja ohjeistukset ovat alisteisia myös näille rakenteille. Tonttijaon vuoksi ohjeistuksien on oltava sensitiivisiä kunkin toimijan toimialueelle ja sallittava, että kukin toimialue operationalisoi lopulta käytäntöjen tasolla ohjeistukset suhteessa niitä koskevaan sääntelyyn, käytännön vaatimuksiin ja tehtäviin. Toki julkisen hallinnon ylimmillä kerroksilla, kuten ministeriöillä, on käytössä vahvat välineet: Se voi omalla tasollaan operationalisoida eettiset ohjeistukset muotoon, jossa se käyttää rahoitusta, informaatio-ohjausta ja sääntelyä ohjaamisen mekanismeina. Teknologian tosiasialliseen käyttöön liittyvät kysymykset kuitenkin konkretisoituvat usein vasta virastojen tai kuntien tasolla.

Lopuksi

On totta, että julkinen hallinto on jossain määrin oma erityiskysymyksensä. Sen normistoa koskevia väitteitä ei voida automaattisesti yleistää koskemaan alueita, joilla ei ole yhtä samanlaista, usein lainsäädännöllä vahvistettua perustaa. Vastaavasti myöskään muita alueita koskevia huomioita ei voida suoraviivaisesti yleistää koskemaan julkista hallintoa. On toki myös totta, että julkinen hallintokaan ei ole vielä onnistunut operationalisoimaan eettisiä ohjeita aina niin, että ne olisivat hyödyllisiä tekoälyn kehittämisen tai käyttämisen näkökulmasta. Ohjeistukset ovat usein liian ympäripyöreitä ja ylimalkaisia. Niiden abstraktiotaso voi myös olla väärä, ja niistä puuttuu riittävä kohdennus tekijöihin, tekemisiin ja tekemisen kohteisiin. Ohjeistukset voivat tulkita väärin teknologiaan liittyvät haasteet ja ohittavat monia muita oleellisia tekijöitä. Nämä puutteet saattavat kuitenkin olla ennen kaikkea signaaleja etiikkatyön keskeneräisyydestä, eivät niinkään sen turhuudesta.

LÄHTEET

Anderson, M. & Anderson, S. L. (2011). Machine ethics. Cambridge: Cambridge University Press.

Belenguer, L. (2022). AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. AI and Ethics, 2, 771–787.

Desouza, K.C., Dawson, G.S. & Chenok, D. (2020). Designing, developing, and deploying artificial intelligence systems: Lessons from and for the public sector. Business Horizons, 63(2), 205-213.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P. & Dignum, V. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707.

Gianni, R., Lehtinen, S. & Nieminen, M. (2022) Governance of responsible AI: From ethical guidelines to cooperative policies. Frontiers of Computer Sciences, 48(7), 34-37.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds & Machines, 30, 99–120.

Henman, P. (2020) Improving public services using artificial intelligence: possibilities, pitfalls, governance. Asia Pacific Journal of Public Administration, 42(4), 209-221.

Li, H., Su, Y., Cai, D., Wang, Y. & Liu, L. (2022). A survey on retrieval-augmented text generation. arXiv preprint. arXiv: 2202.01110.

McNamara, A., Smith, J. & Murphy-Hill, E. (2018). Does ACM's code of ethics change ethical decision making in software development? Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, October 2018, ss. 729–733. https://doi.org/10.1145/3236024.3264833.

Morley, J., Floridi, L. & Kinsey, L. (2019). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Science and Engineering Ethics, 26, 2141–2168.

Morley, J., Elhalal, A. & Garcia, F. (2021). Operationalisation of AI ethics. Minds & Machines, 31, 239–256.

Munn, L. (2023). The uselessness of AI ethics. AI and Ethics, 3, 869–877.

Naitali, A., Ridouani, M., Salahdine, F. & Kaabouch, N. (2023). Deepfake attacks: Generation, detection, datasets, challenges, and research directions. Computers, 12(10).

Ratra, R. & Gulia, P. (2020). Privacy preserving data mining: Techniques and algorithms. International Journal of Engineering Trends and Technology, 68, 56-62.

Rességuier, A. & Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. Big Data Society, 7(2).

Robbins, S. A. (2019). Misdirected principle with a catch: Explicability for AI. Minds & Machines, 29, 495–514.

Roinila, M. (2021). Tekoälyn varhaishistoriaa: laskevia koneita ja spiritualisia automaatteja. Teoksessa P. Raatikainen (toim.), Tekoäly, ihminen ja yhteiskunta: filosofisia näkökulmia (ss. 21-37). Helsinki: Gaudeamus.

Rusanen, A-M. (2021). Algoritmien aakkoset. Teoksessa Älykäs huominen - Miten tekoäly ja digitalisaatio muuttavat maailmaa? (ss. 33-39). Helsinki: Gaudeamus.

Zalnieriute, M. (2021). "Transparency-Washing' in the digital age: A corporate agenda of procedural fetishism. Critical Analysis of Law, 8(1), 21–33.

Zuiderwijk, A., Chen, Y. & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. Government Information Quarterly, 38(3), 101577.

Xu, R., Baracaldo, N. & James, J. (2021). Privacy-preserving machine learning: Methods, challenges and directions. arXiv preprint, arXiv:2108.04417.

Wachter, S., Mittelstadt, B. & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 842-887.

Wallenberg, A-M. (valmisteilla). How Ethics of AI went wrong.

ANNA-MARI WALLENBERG *Kognitiotieteen yliopistonlehtori, dosentti (Digitaalisten ihmistieteiden osasto, Helsingin yliopisto) tutkii älykkäiden järjestelmien tiedonkäsittelyominaisuuksia, sekä tekoälyteknologioiden yhteiskunnallisia seurauksia monitieteisestä näkökulmasta. Viime vuosina Anna-Mari on lisäksi työskennellyt tekoälyn etiikan ja tekoälyä koskevan sääntelyn parissa erityisasiantuntijana Valtiovarainministeriössä, sekä tutkinut mm. datalukutaidon merkitystä päätöksenteossa.*