strateginen TUTKIMUS

Eettinen tekoälyn hyödyntäminen yhteiskunnan ohjauksessa (ETAIROS)

Tilannekuvaraportti 2019

SUOMEN AKATEMIA | Hakaniemenranta 6 | PL 131 | 00531 Helsinki | Puh. 029 533 5000 | etunimi.sukunimi@aka.fi | www.aka.fi/stn

Tiivistelmä

Tekoäly (AI) on maailmaa muokkaava teknologia, joka vaikuttaa lähes kaikkiin elämänalueisiin. ETAIROS-hankkeen tavoitteena on kehittää tapoja ja menetelmiä tekoälyn eettiseen hyödyntämiseeni yhteiskunnan ohjauksessa.

ETAIROSin tavoite on

1) kehittää teoreettisesti ja empiirisesti perusteltu lähestymistapa tekoälyn ja sen yhteiskunnallisten vaikutusten kehityksen ja käytön ohjaamiseen;
2) luoda ymmärrystä tekoälyn eettisen ja vastuullisen käytön kontekstikohtaisista haasteista, mahdollisuuksista ja esteistä;
3) tuottaa testattuja suunnittelu- ja koneopetusprosesseja eettiselle tekoälylle, ja
4) tuottaa empiirisesti perusteltuja tekoälyn hallintotapoja ja -käytäntöjä.

Tulokset tarjoavat välineitä sosiaalisesti kestävään strategiseen suunnitteluun, poliittiseen päätöksentekoon ja tekoälyn sääntelyyn.

ETAIROS tukee julkisen, yksityisen ja kolmannen sektorin toimijoiden eettistä itsesääntelyä ja yhteiskunnan ohjausta hyödyntämällä erilaisia yhteistyömuotoja, kuten yhteisinnovaatiofoorumeita, joista ensimmäinen pidettiin 7.10.2019.

Tekoäly muuttaa yhteiskuntaa

Rajoitettuun tekoälyyn liittyvät odotukset ovat nopeassa kasvussa. Samalla on alettu aiempaa selvemmin tiedostaa tekoälyn ongelmia ja rajoituksia. Gartnerin ns. “Hype cycle for Artificial Intelligence” -kuvassa (Kuva 1) tekoäly on vahvasti “kasvavien odotusten” käyräosuudella.

Figure 1. Gartnerin “Hype cycle for Artificial Intelligence” (gartner.com)

Tekoälyteknologioiden odotetaan mullistavan maailmamme. Tekoälytoimijat ovat “digitaalisia tietokoneita tai tietokoneen ohjaamia robotteja, jotka suorittavat yleensä älykkäisiin toimijoihin yhdistettyjä tehtäviä”, ja yhdistelevät kykyä oppia, ohjata päätöksiä, ratkaista ongelmia ja käsitellä kieltä (Copeland 2018). Tekoälyteknologioiden päättelymenetelmien taso ja sovellusalueiden laajuus vaihtelevat. Rajoitettua tekoälyä käytetään algoritmisissa asiantuntijajärjestelmissä, suurten datamassojen hyödyntämisessä, syväoppimisessa mm. terveydenhuollon sovelluksissa, pankkien ja vakuutusyhtiöiden analytiikassa ja dokumenttien käsittelyssä, julkisissa palveluissa, teollisessa tuotannossa, turvallisuuden hallinnassa, autonomisissa ajoneuvoissa ja sosiaalisessa mediassa (O’Neil 2016; OECD 2018). Yleinen tekoäly (Artificial General Intelligence, AGI) eli järjestelmät, joilla on ”jonkintasoinen itseymmärrys ja autonominen itsehallinta” (Saariluoma 2015), ovat näköpiirin ulkopuolella.

Vaikka osa tekoälyinnostuksesta on ylimitoitettua, tekoälyllä on valtava taloudellinen ja sosiaalinen potentiaali muuttaa yhteiskuntaa (Brynjolfsson & Mcafee 2017). On todennäköistä, että nämä teknologiat muuttavat sosiaalisia rakenteita, työelämää ja jopa elämäämme ihmislajina (Bostrom, 2014). Viimeaikaisten selvitysten mukaan tekoälyn vaikutukset ulottuvat mitä erilaisimmille elämänalueille ja ilmiöihin, kuten politiikkaan (Helbing et al, 2017), sodankäyntiin (Cummings, 2017) tulonjakoon, ja työllistymiseen (e.g. Korinek & Stiglitz 2017).

Hallitukset ovat tunnistaneet tekoälyn mahdollisuudet. Melkein kaikkien kehittyneiden maiden hallituksilla on tekoälystrategia, ja ne kilpailevat toistensa kanssa tekoälyn kehittämisessä ja käyttöönotossa (Dutton 2018). Vuonna 2018 EU muodosti tekoälyasiantuntijaryhmän tukemaan EU:n tekoälystrategiaa (High-Level Expert Group on Artificial Intelligence 2019). Suomessa Business Finland käynnisti tekoälyohjelman ja hallitus hyväksyi raportin “Ethical information policy in an age of artificial intelligence.”

Tekoäly käyttöön eettisesti ja sosiaalisesti kestävällä tavalla

Vaikka tekoälytekniikoilla on monia myönteisiä vaikutuksia yhteiskuntaamme ja elämäämme terveydenhuollosta oppimiseen ja päätöksentekoon, ne tuovat ratkaistavaksi myös lukuisia eettisiä kysymyksiä jo senkin vuoksi, että tekoälyn suunnitteluun ja käyttöön liittyy väistämättä sosiaalisia ja kulttuurisia näkökohtia (Kitchin 2017, 18).

Tutkimukset ovat osoittaneet (esimerkiksi Eubanks ,2017), että koneoppimismenetelmät aiheuttavat ohjelmistovalinnoista ja järjestelmien opettamiseen käytetystä tiedosta johtuvaa sosiaalista harhaisuutta. Harhat vinouttavat tekoälyä hyödyntävän järjestelmän toimintaa niin, että sen avulla tehdyt päätökset ovat esimerkiksi sukupuolen ja etnisen taustan suhteen syrjiviä ja heikentävät siksi ihmisten välistä yhdenvertaisuutta (Weber 2018). Koska järjestelmät ovat niin monimutkaisia, tällaisia poikkeamia on erittäin vaikea tunnistaa, eikä ”vinoutuneisuutta poistavia” algoritmeja ruveta siksi kehittämään (Brynjolfsson & Mcafee 2017).

AI-pohjaisia järjestelmiä käytetään jo nyt monella tavalla väärin. Algoritmisia tekniikoita on käytetty demokraattisten prosessien horjuttamiseen mm. vaalivaikuttamisen kautta (Cadwallar & Graham-Harrison 2018) ja myös monilla tavoin kansalaisten illegitiimiin tarkkailuun ja hallintaan. Pahimmassa tapauksessa syntyy yhteiskuntia, joissa panoptinen valvonta (Müller 2016) ja erilaiset social credit -järjestelmät (Citron & Pasquale 2014) seuraavat ja ohjaavat elämäämme. Vieläkin synkemmissä dystopioissa ihmiskunta kohtaa loppunsa yleisen tekoälyn toteuduttua ja otettua vallan tai kun AI:n rikollisen käytön tuhoisat seuraukset hävittävät järjestyneet yhteiskunnat (Naudé & Nicola 2018).

Näyttää siltä, että olemme siirtymässä ”uuteen laajalle levinneen algoritmisen hallinnan aikakauteen, jossa algoritmien rooli vallankäytön osana kasvaa jatkuvasti” (Kitchin 2017, 15). Muutoksen mukanaan tuomia ongelmia voi vaikeuttaa se, että koneen ajattelu ei ole eikä todennäköisesti voi olla ihmisen harjoittaman päättelyn jäljitelmä tai jatke (Lake 2017).

Viime vuosina on julkaistu lukuisia tutkimuksia, joissa on pyritty luonnostelemaan tekoälyn eettisen käytön suuntaviivoja. AI:n eettisten ohjeiden soveltamisesta ja erilaisten AI-järjestelmien käytännön suunnittelusta ja implementoinnista niiden erilaisissa käyttöolosuhteissa on kuitenkin vain muutamia tutkimuksia.

Tarvitaan toisaalta yksityiskohtaista sääntelyä, toisaalta on selvitettävä toimijoiden itsesääntelyn ala. Myös säätelyn toteuttaminen on haasteellista. Charbit'n (2011, 15-16) mukaan esimerkiksi riittämätön ja epäsymmetrinen informaatio sidosryhmien välillä, huonosti toteutettu poliittisten päämäärien koordinointi ja toisistaan poikkeavat näkemykset politiikan eri aloilla sekä toiminnan läpinäkymättömyys ja kansalaisten osallistumisen mukanaan tuomat erilaiset vaikeudet voivat estää pehmeiden ohjelmien toteuttamisen.

Sääntelyn ja hallinnon eri tasojen välillä tulisi olla koordinointia eurooppalaiselta kansalliselle ja alueelliselle / paikalliselle tasolle (horisontaalinen ja monitasoinen hallinto, esim. Conzelmann 2008). On kyettävä myös luomaan sopiva tasapaino "kovan" ja "pehmeän" sääntelyn välillä. Ymmärrämme hallinnon laajasti a) AI-toimijoiden poliittisena ja hallinnollisena ohjauksena ja koordinointina, b) ei-hierarkkisina koordinointitapoina ja c) erilaisina koordinointimuotoina.

ETAIROS

ETAIROS-projektissa tutkimme ja kehitämme käytännön prosesseja ja puitteita, jotka auttavat julkisen, yksityisen ja kolmannen sektorin toimijoita parantamaan tekoälyä hyödyntävän toimintansa eettistä kestävyyttä. Erityisenä tavoitteenamme on tutkia ja luoda yhdessä hallintokehyksiä ja työkaluja, joiden avulla voidaan luoda tapoja käyttää tekoälyä sosiaalisesti kestävällä tavalla julkisissa palveluissa ja hallinnossa sekä yksityisellä sektorilla.

Tavoitteen saavuttaminen on monitieteellinen tehtävä.

1. Tekoälyn eettisiä ongelmia on tiedostettu, mutta niiden arviointi ja evaluointi on edelleenkin haaste. Eettisiä ongelmia ei aina tunnisteta. Etiikan asiantuntemusta tarvitaan tekoälyn etiikan ymmärtämiseen.
2. Eettiset näkökulmat tulee ottaa huomioon jo tekoälyjärjestelmiä suunniteltaessa. Tuolloin on tarpeen roolittaa oikein ihmiset ja tekoälytoimijat sekä muut järjestelmät. Tekoälyjärjestelmät ovat monimutkaisia, ja parempia työkaluja tarvitaan sekä eettisten aspektien mallintamiseen että järjestelmien mallintamiseen, analysointiin ja simulointiin.
3. Merkittävä osa tunnistetuista tekoälyn riskeistä liittyy koneoppimisen ja siinä käytetyn datan mukanaan tuomiin ongelmiin ja vinoutumiin. Näiden tunnistaminen ja välttäminen kuuluu hankkeen tutkimusteemoihin.
4. Yhteiskunnan ohjaus tekoälyn hyödyntämisessä ja tekoälyä hyödyntäen on merkittävä tulevaisuuden haaste jossa sekä riskit että potentiaali on suuri. Toimijoilla on oltava yhteinen ja koordinoitu ymmärrys yleisistä periaatteista, joiden varassa tekoälyn käyttöä toteutetaan ja säännellään.

Etairosissa yhdistyvät monipuolinen asiantuntemus ja yhteistyö

ETAIROS-hanke vastaa edellä kuvattuihin haasteisiin ottamalla kaikki asiaan liittyvät keskeiset toimijat (viranomaiset, asiantuntijat, kansalaiset, yksityinen sektori) läpinäkyvästi ja tavoitteellisesti toteutettuun tietoiseen yhteisinnovaatioprosessiin, joka sisältää uusia käytännön hallintokehyksiä ja työkaluja (mukaan lukien sääntelyehdotukset). Hankkeeseen liitetyt konkreettiset käyttötapaukset edistävät yhteisen ymmärryksen muodostumista haasteiden luonteesta mutta myös ratkaisujen löytämistä.

Eettisten AI-sovellusten kehittäminen on vaativa ja laaja-alainen sosiotekninen haaste

(Geels & Schot 2007). Siihen liittyvien uhkien torjuminen vaatii monitieteistä lähestymistapaa (Ailisto ym. 2018). AI: n suunnittelu ei ole vain moniteknologista työtä, vaan sen onnistunut toteuttaminen edellyttää sosiaalisten, psykologisten, taloudellisten, poliittisten ja oikeudellisten näkökohtien huomioon ottamista. Koska tekoälyjärjestelmillä on todennäköisesti syvällekäyvä vaikutus erilaisiin yhteiskunnallisiin ulottuvuuksiin (Lucivero 2016), AI:n tutkiminen edellyttää tutkimusryhmää, jolla on hallussaan laaja valikoima erilaisia tutkimuksellisia lähestymistapoja ja osaamista.

ETAIROS-konsortio yhdistää monipuolisen asiantuntemuksen, mukaan lukien tietotekniikan, moraalifilosofian, tilastotieteen, oikeustieteen, psykologian, kognitiotieteen, turvallisuustieteen, ennakointitekniikoiden, suunnittelumallintamisen, ihmisen ja tekniikan vuorovaikutuksen, riskienhallinnan ja turvallisuuskulttuurin, julkisen johtamisen, hallinnon ja sosiaalisen oppimisen näkökulmat sekä tiedonhallinnan, ja tiede-, teknologia- ja innovaatiotutkimuksen. Hankkeen osaksi on liitetty myös moniulotteinen yhteistyö ja -toiminta yhteiskunnallisten sidosryhmien ja kansalaisten kanssa.

ETAIROS-hankkeen odotettavissa olevat sosiaaliset vaikutukset kattavat laajan joukon toimia ja prosesseja. Hanke lisää ymmärrystämme erilaisista tekoälyn konkreettisista käyttökonteksteista ja niihin liittyvistä uudenlaisista haasteista mutta myös AI:n mahdollisuuksista kehittää konkreettisia eettisiä suunnittelu- ja arviointikehyksiä ja työkaluja erilaisiin tehtäviin ja ympäristöihin sekä luonnostella yleisiä hallintotapoja ja -käytäntöjä, jotka tukevat julkisia, yksityisiä ja kolmansia osapuolia.

Temaattiset alueet

Julkiselle sektorille hanke tuottaa tekoälyn hyödyntämistä ja käyttöä koskevia ehdotuksia ja käytäntöjä AI:n suunnitteluun ja hallintoon. Tavoitteena on edistää kestävää, avoimuuteen perustuvaa ja osallistaa yhteiskunnallista kehitystä. Kansalaisten kannalta

hanke lisää tekoälyn käytön läpinäkyvyyttä, eettisesti hyväksyttävien AI-järjestelmien yleistä ymmärtämistä sekä mahdollisuuksia tietoon perustuvaan julkiseen keskusteluun ja vaikuttamiseen.

Etiikka ja vastuu ovat ETAIROS-hankkeen läpikäyviä periaatteita. Noudatamme tutkimuksessa vastuullisen innovoinnin perusajatuksia, joihin kuuluvat innovaatiotoiminnan vaikutusten ennakoiminen, oman työn ja sen tavoitteiden jatkuva arvioiminen, osallisuuden edistäminen selvittäessämme sidosryhmien ja kansalaisten etuja ja arvoja sekä jatkuvaan oppimiseen perustuva työtapa, jossa hyviä käytäntöjä jaetaan ja edistetään (Owen et al. 2013).

Noudatamme myös mm. Euroopan komission suosituksia, joita on laadittu korkean tason tekoälyasiantuntijaryhmille ja eurooppalaisen etiikkaryhmän IEEE: n maailmanlaajuista aloitetta autonomisten ja älykkäiden järjestelmien etiikasta ja suosituksia tutkimuseettisistä käytännöistä.

Yhteiskunnallinen vakaus -teema-alueessa tarkastellaan yhteiskunnan turvallisuutta edistäviä valvontakohteita, joissa kansalaiset ovat osallisina. Näissä tekoäly toimii turvallisuuden hallinnassa ihmisen partnerina. ETAIROSin erityisenä kiinnostuksen kohteena on, pystytäänkö rakentamaan sellainen tekoälyjärjestelmä, joka käyttää reaaliaikaista dataa, oppii samalla koko ajan, ja muodostaa jatkuvasti tilannekuvaa ihmisten toiminnasta, sekä miten tällainen toimintamalli voi olla eettinen. Tärkeä kysymys on myös, voidaanko tekoälyn toiminnassa tunnistaa eettisesti arveluttavia tekijöitä.

Autonomiset järjestelmät -teema-alueessa tarkastellaan autonomisten järjestelmien ja ihmisen uudenlaista vuorovaikutusta ja tämän vuorovaikutuksen käyttötapojen merkitystä turvallisuuden ja päätöksenteon kannalta. Hankkeessa tarkastellaan erityisesti tekoälyn mahdollisuuksia ja eettistä käyttöä erilaisten liikennejärjestelmien yhteensovittamisessa, sekä liikennettä palveluna ja tekoälyä työkaverina.

Kansallinen turvallisuus - teema-alueessa, joka painottuu kansalliseen turvallisuuteen ja maan puolustukseen, korostuu erityisesti se, että samalla kun tekoälyn käyttömahdollisuudet lisääntyvät, kompleksisuus kasvaa ja riskit lisääntyvät. ETAIROSissa tarkastellaan tällä teema-alueella erityisesti ovatko kansalliset turvallisuutta koskevat, olemassa olevat eettiset säädökset ja toimintamallit riittäviä, ja millaista sosioteknistä uudistusta nämä toimintamallit kaipaavat, ts. millaisia paranuksia toimintamalleihin tulisi tehdä uusien uhkien pohjalta, kansalaisten ja yhteiskunnan näkökulmasta.

Sosiaali- ja terveyspalvelut - teema-alueessa tarkastellaan tekoälyn eettisen käytön näkökulmia sosiaali- ja terveyspalvelujen kentässä. ETAIROSia kiinnostaa tällä teema-alueella erityisesti se, miten erilaiset tekoälyjärjestelmät voivat luoda sellaista yhteistoimijuutta, joka tukisi asiakkaiden tai potilaiden omaa aktiivisuutta ja ylläpitäisi ennaltaehkäisevästi terveyttä ja hyvinvointia.
Ikääntymisen teema-alue painottuu kotihoidon ja palveluasumisen konteksteihin. Tässä teema-alueessa tarkastellaan tekoälyn käyttötapauksia ja eettisiä kysymyksiä. Tarkastelun kohteena on tekoälyn eettinen hyödyntäminen ja käyttö kotihoidon asiakkaan/ läheisen / hoitohenkilökunnan / ja palvelun tuottajan näkökulmasta.

Tekoälyn eettisyyden kokonaisvaltainen ohjaus ja sääntely -teema-alue tuottaa synteesiä muista edellä esitetyistä teema-alueista. Se painottaa yhteiskunnallisia ja oikeudellisia näkökohtia tekoälyn eettisessä hyödyntämisessä.
Mielenkiintoinen, kaikkia teema-alueita läpileikkaava teema on, miten yrityksille voi syntyä kilpailuetua eettisen tekoälyn kehittämisen kautta.

Etairos-konsortio tutkii ja kehittää yhteistyössä kattavan yhteistyöverkoston kanssa, sisältäen mm ministeriöitä, kuntia, ja merkittäviä julkisia palveluita tuottavia organisaatioita. Tässä merkittävä työkalu on Co-Innovation Forum), joista ensimmäinen on jo pidetty ja saimme sinne erittäin hyvin osanottajia. Etairos järjestää teema-alueiltaan myös kaikille

avoimia tilaisuuksia (Open Dialogue Forum).

Lähdeluettelo

Ailisto et al. (2018) Tekoälyn kokonaiskuva ja osaamiskartoitus. Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja 46/2018. Valtioneuvoston kanslia. Bostrom, N. (2014) Superintelligence. Oxford University Press.Brynjolfsson, E. & Mcafee, A. (2017) The business of artificial intelligence. Harvard Business Review, 7.Charbit, C. (2011) Governance of Public Policies in Decentralised Contexts: OECD Regional Development Working Papers, 2011/04, OECD Publishing. Cadwallar, C. & Graham-Harrison E. (2018) How Cambridge Analytica turned Facebook 'likes' into a lucrative political tool. The Guardian, March 17. Citron, D. K. & Pasquale, F. A. (2014) The scored society. Washington Law Review, 89, 2014; U of Maryland Legal Studies Research Paper No. 2014-8. Conzelmann, T. (2008) A New Mode of Governing? In T. Conzelmann, & R. Smith (Eds.), Multi-level Governance in the European Union: Taking Stock and Looking Ahead, pp. 11-30. Baden-Baden: Nomos.Copeland, B.J. (2018) Artificial intelligence. Encyclopaedia Britannica.Cummings, M. L. (2017) Artificial intelligence and the future of warfare. London: Chatham House.Dutton, T. (2018) An overview of national AI strategies. https://medium.com/ Eubanks V,2017, Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor, St Martin's Press, NYGeels, F. & Schot, J. (2007) Typology of sociotechnical transition pathways. Research Policy, 36(3), 399-417.Helbing et al. (2017) Will democracy survive Big Data and Artificial Intelligence? Scientific American, February 25. High-Level Expert Group on Artificial Intelligence (2019) Ethics guidelines for trustworthy AI. European Commission.Kitchin, R. & Dodge, M. (2011). Code/Space. Cambridge, MA: MIT Press. Korinek, A. & Stiglitz J. (2017) Artificial Intelligence and its implications for income distribution and unemployment. NBER Working Paper.Lake, B., Ullman, T., Tenenbaum, J. & Gershman, S. (2017) Building machines that learn and think like people. Behavioral and Brain Sciences, 40. Müller, V. (Ed.). (2016) Risks of AI. CRC Press, Taylor & Francis.Naudé W. & Nicola, D. (2018) The race for an artificial general intelligence: Implications for public policy. UNU-MERIT Working Papers, Maastricht. Lucivero, F. (2016) Ethical assessments of emerging technologies. The International Library of Ethics, Law and Technology, Vol 15. OECD (2018) AI: Intelligent machines, smart policies: Conference summary. OECD Digital Economy Papers, No. 270, OECD Publishing, Paris.Owen et al..

(2013) A Framework for Responsible Innovation. In R.Owen, J. Bessant & M.Heintz, (Eds.) Responsible Innovation. John Wiley, 27-50. Saariluoma, P. (2015) Four challenges in designing autonomous systems. In: Williams et al. (Eds.) Autonomous Systems: Issues for Defence Policymakers. NATO. Weber, J. (2018) Auto-management as governance? Predictive analytics in counter-insurgency and marketing. EASST 2018, Lancaster University.