Algoritminen syrjintä ja yhdenvertaisuuden edistäminen

Arviointikehikko syrjimättömälle tekoälylle

Atte Ojanen, Otto Sahlgren, Juho Vaiste, Anna Björk, Johannes Mikkonen, Kai Kimppa, Arto Laitinen, Nea Oljakka

VALTIONEUVOSTON SELVITYS- JA TUTKIMUSTOIMINNAN JULKAISUSARJA 2022:54

tietokayttoon.fi

Selvitys- ja tutkimustoiminta
Utrednings- och forskningsverksamhet

VALTIONEUVOSTO
STATSRÅDET

Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja 2022:54

Algoritminen syrjintä ja yhdenvertaisuuden edistäminen

Arviointikehikko syrjimättömälle tekoälylle

Atte Ojanen, Otto Sahlgren, Juho Vaiste, Anna Björk, Johannes Mikkonen, Kai Kimppa, Arto Laitinen, Nea Oljakka Helsinki

Demos Helsinki, Tampereen yliopisto, Turun yliopisto

Valtioneuvoston kanslia Helsinki 2022

Julkaisujen jakelu
Distribution av publikationer

Valtioneuvoston julkaisuarkisto Valto

Publikations-
arkivet Valto

julkaisut.valtioneuvosto.fi

Julkaisumyynti
Beställningar av publikationer

Valtioneuvoston verkkokirjakauppa

Statsrådets
nätbokhandel

vnjulkaisumyynti.fi

ISBN pdf: 978-952-383-404-0
ISSN pdf: 2342-6799

Taitto: Valtioneuvoston hallintoyksikkö, Julkaisutuotanto

Helsinki 2022

Kuvailulehti

23.8.2022

Algoritminen syrjintä ja yhdenvertaisuuden edistäminen
Arviointikehikko syrjimättömälle tekoälylle

Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja 2022:54

Julkaisija	Valtioneuvoston kanslia		
Tekijä/t	Atte Ojanen, Otto Sahlgren, Juho Vaiste, Anna Björk, Johannes Mikkonen, Kai Kimppa, Arto Laitinen, Nea Oljakka		
Toimittaja/t	Demos Helsinki		
Yhteisötekijä	Demos Helsinki, Tampereen yliopisto, Turun yliopisto		
Kieli	suomi	**Sivumäärä**	169

Tiivistelmä

Tekoälyn kehitys on edennyt viime vuosina ennennäkemättömällä vauhdilla. Koneoppimiseen perustuvia algoritmeja pidetään lupaavina koulutuksen, terveydenhuollon, rekrytoinnin ja monen muun palvelun parantajana. Algoritminen päätöksenteko luo kuitenkin samaan aikaan myös uhkia yhdenvertaisuudelle ja syrjimättömyydelle. "Tekoälyn vinoumien välttäminen: suomalainen arviointikehikko syrjimättömille tekoälysovelluksille" tutkimus- ja selvityshankkeen tavoitteena oli:

- Tuottaa uutta tietoa siitä, millaisia koneoppimiseen pohjautuvia tekoälyjärjestelmiä Suomessa on käytössä, millaiseen vaikutustenarviointiin ne pohjaavat ja mitä syrjiviä vaikutuksia niillä saattaa olla.
- Arvioida kriittisesti algoritmisten tekoälyjärjestelmien syrjiviä ja perusoikeudellisia vaikutuksia kartoituksen perusteella, huomioiden yhdenvertaisuuslain asettamat velvoitteet.
- Yhteiskehittää tutkimuksen perusteella arviointikehikko tekoälysovellusten syrjivien piirteiden tunnistamiseksi ja välttämiseksi sekä yhdenvertaisuuden edistämiseksi, ja luoda politiikkasuosituksia algoritmien sääntelyn kehittämiseksi.

Tämä loppuraportti kokoaa yhteen hankkeen tulokset ja esittelee sen päätuotteena syntyneen arviointikehikon tekoälyjärjestelmien syrjinnän ja yhdenvertaisuuden arviointiin. Lisäksi tarjoamme politiikkasuosituksia arviointikehikon integrointiin osaksi valtionhallintoa ja vastuullista tekoälyn kehittämistä.

Asiasanat tutkimus, tutkimustoiminta, algoritmit, tekoäly, syrjintä, yhdenvertaisuus

ISBN PDF	978-952-383-404-0	**ISSN PDF**	2342-6799
Julkaisun osoite	https://urn.fi/URN:ISBN:978-952-383-404-0		

Presentationsblad

23.8.2022

Algoritmisk diskriminering och främjande av jämlikhet
Bedömningsram för icke-diskriminerande AI

Publikationsserie för statsrådets utrednings- och forskningsverksamhet 2022:54

Utgivare	Statsrådets kansli		
Författare	Atte Ojanen, Otto Sahlgren, Juho Vaiste, Anna Björk, Johannes Mikkonen, Kai Kimppa, Arto Laitinen, Nea Oljakka		
Redigerare	Demos Helsinki		
Utarbetad av	Demos Helsinki, Tammerfors universitet, Åbo universitet		
Språk	finska	**Sidantal**	169

Referat

Utvecklingen av artificiell intelligens har gått framåt i en aldrig tidigare skådad takt under de senaste åren. Algoritmer för maskininlärning ses som lovande för att förbättra utbildning, hälsovård, rekrytering och många andra tjänster. Samtidigt utgör dock algoritmiskt beslutsfattande ett hot mot jämlikhet och icke-diskriminering. Forsknings- och analysprojektet "Undvika AI-bias: en finländsk utvärderingsram för ickediskriminerande AI-tillämpningar" syftade till att:

- Att skapa ny kunskap om vilka typer av maskininlärningsbaserade AI-system som används allmänt i Finland, vilka konsekvensbedömningar de bygger på och vilka diskriminerande effekter de kan ha.
- Att kritiskt bedöma de diskriminerande konsekvenserna och konsekvenserna för de grundläggande rättigheterna av algoritmiska AI-system baserade på kartläggningen, med hänsyn till de skyldigheter som följer av lagen om ickediskriminering.
- På grundval av studien medutveckla en ram för bedömning för att identifiera och undvika diskriminerande inslag i AI-tillämpningar och främja jämlikhet, samt ta fram politiska rekommendationer för att förbättra regleringen av algoritmer.

I denna slutrapport sammanfattas resultaten av projektet och presenteras dess huvudprodukt, en utvärderingsram för att bedöma diskriminering och jämlikhet i AI-system. Vi ger också politiska rekommendationer för att integrera utvärderingsramen i den statliga förvaltningen och i en ansvarsfull utveckling av artificiell intelligens.

Nyckelord	orskning, forkningsverksamhet, AI, diskriminering, jämlikhet		
ISBN PDF	978-952-383-404-0	**ISSN PDF**	2342-6799
URN-adress	https://urn.fi/URN:ISBN:978-952-383-404-0		

Description sheet

23 August 2022

Algorithmic discrimination and the promotion of equality
Assessment framework for non-discriminatory AI

Publications of the Government´s analysis, assessment and research activities 2022:54

Publisher	Prime Minister's Office		
Author(s)	Atte Ojanen, Otto Sahlgren, Juho Vaiste, Anna Björk, Johannes Mikkonen, Kai Kimppa, Arto Laitinen, Nea Oljakka		
Editor(s)	Demos Helsinki		
Group author	Demos Helsinki, University of Tampere, University of Turku		
Language	Finnish	**Pages**	169

Abstract

The development of artificial intelligence has advanced at an unprecedented pace in recent years. Machine learning algorithms are seen as having a potentially positive impact on education, healthcare, recruitment and many other services. At the same time however, algorithmic decision-making poses threats to equality and non-discrimination. The main objectives of the research project "Avoiding AI biases: a Finnish assessment framework for non-discriminatory AI systems" were:

- To generate new knowledge about the types of machine learning-based AI systems that are used in Finland, the impact assessment they employ, and their possible discriminatory risks.
- To critically assess the discriminatory and fundamental rights impacts of algorithmic systems based on the mapping, taking into account the obligations imposed by the Non-Discrimination Act.
- To co-produce, on the basis of the study, an assessment framework to identify and avoid discriminatory features of AI applications and to promote equality, as well as generate policy recommendations to improve the regulation of AI systems

This final report summarises the results of the project and presents its main output, the assessment framework for assessing discrimination and equality in AI systems. We also introduce policy recommendations for integrating the assessment framework into the public sector governance and responsible AI development.

Provision This publication is part of the implementation of the Government Plan for Analysis, Assessment and Research. (tietokayttoon.fi) The content is the responsibility of the producers of the information and does not necessarily represent the view of the Government.

Keywords research, research activities, AI, non-discrimination, equality

ISBN PDF	978-952-383-404-0	**ISSN PDF**	2342-6799
URN address	https://urn.fi/URN:ISBN:978-952-383-404-0		

Sisältö

Julkaisun ulkopuoliset liitteet on tallennettu omana tiedostonaan osoitteeseen http://urn.fi/URN:ISBN:978-952-383-404-0

JOHDANTO: TEKOÄLYN SYRJINTÄRISKIEN ARVIOINTI JA YHDENVERTAISUUDEN EDISTÄMINEN SUOMESSA

Tekoäly viittaa kehittyneisiin ohjelmistoihin ja tietokonesovelluksiin, jotka pystyvät toimimaan yhä autonomisemmin ja älykkäämmin ihmisten asettamien tavoitteiden saavuttamiseksi, myös vaihtuvissa olosuhteissa. Viimeisen viiden vuoden aikana tekoälyn kehitys on edennyt ennennäkemättömällä vauhdilla. Nykyistä tekoälyn aaltoa leimaavat datan runsas saatavuus, laskentatehon kasvu ja erityisesti edistyneet koneoppimisteknologiat, kuten neuroverkot ja syväoppiminen.

Vaikka nykyiset tekoälysovellukset eivät vielä edusta niin sanottua yleistä älykkyyttä, on niillä jo suuria yhteiskunnallisia vaikutuksia aina ihmisten perusoikeuksiin asti. Tekoälysovellusten tuottamat ja ylläpitävät syrjivät rakenteet ja käytännöt ovatkin yksi merkittävistä haasteista, joita myös julkisen hallinnon keinoin tulee ratkaista. Kun tekoälyä hyödynnetään esimerkiksi automatisoidussa päätöksenteossa, vinoutunut tai syrjivä harjoitusdata saattaa uusintaa ja vahvistaa yhteiskunnallista eriarvoistuutta varsin näkymättömästi. Algoritmiset vinoumat ja niihin liittyvä syrjintä voivat johtua useasta seikasta läpi tekoälyn elinkaaren:

1. Kyseenalaiset perusteet järjestelmän kehittämiselle ja huonosti määritellyt tavoitteet ilman asianosaisten ihmisten kuulemista suunnitteluvaiheessa.
2. Datavinoumat, eli puutteet tai virheet algoritmin opetusdatan edustavuudessa, kuten otanta- ja nimikevinoumat.
3. Algoritmin tai mallin määrittäminen, opettaminen, testaaminen ja jälkiprosessointi ja siihen liittyvät huonosti valitut ennustemuuttujat, reiluusmetriikat ja vertailuluokat.
4. Järjestelmän käyttöönotto suunnittelemattomassa ympäristössä tai väestössä, virheellinen käyttö, riittämätön ylläpito ja läpinäkymättömyys järjestelmän toiminnassa.

Tämä julkaisu esittelee valtioneuvoston selvitys- ja tutkimustoiminnan hankkeen "Tekoälyn vinoumien välttäminen: suomalainen arviointikehikko syrjimättömille tekoälysovelluksille" tulokset. Hankkeen tavoitetteena oli kartoittaa, millaisia riskejä perusoikeuksille ja syrjimättömyydelle Suomessa käytössä ja suunnitteilla olevat, erityisesti koneoppimiseen pohjautuvat tekoälysovellukset saattavat sisältää. Hankkeen keskeisenä tavoitteena oli laatia arviointikehikko ja politiikkasuositukset tekoälysovellusten syrjimättömyyden

varmistamiseksi eri käyttökonteksteissa. Arviointikehikon tavoitteena on tukea yhdenvertaisuuslain ja yhdenvertaisuuden edistämisvelvoitteen (YVL 5-7 §) toimeenpanoa tekoälysovellusten kohdalla.

Julkaisun ensimmäinen osio esittelee tulokset Turun yliopiston toteuttamasta kansallisesta kartoituksesta perusoikeuksiin liittyvistä tekoälysovelluksista Suomessa. Kartoitus, osoittaa, että tekoälyjärjestelmien käyttöönotto on edelleen Suomessa varhaisessa vaiheessa. Vaikka tekoälyjärjestelmiin liittyvään algoritmiseen syrjintään on herätty kohtuullisesti, sen kitkemiseen ei ole vakiintunut selkeää viranomaisyhteistyön mallia tai työkaluja.

Julkaisun toinen osio "Syrjinnän riskit ja yhdenvertaisuusperiaate tekoälyjärjestelmissä" käsittelee Tampereen yliopiston tuottamaa analyysia tekoälyjärjestelmien syrjintäriskeistä, syrjivien vaikutusten teknologisista ja yhteiskunnallisista syistä, niiden tunnistamiseen ja ehkäisemiseen kehitetyistä menetelmistä, sekä haasteista, joita menetelmien käyttöön liittyy. Analyysin perusteella syrjivät vinoumat syntyvät useimmiten tekoälyjärjestelmien arvoketjussa erilaisten sosio-teknisten tekijöiden yhteisvaikutuksesta, eikä niitä voida ratkaista vain teknisin menetelmin ilman kontekstuaalista ja tapauskohtaista harkintaa.

Julkaisun kolmas osio esittelee Demos Helsingin johtaman työn tuloksena kehitetyn arviointikehikon syrjimättömille tekoälyjärjestelmille. Kehikko auttaa tunnistamaan ja hallitsemaan erityisesti julkisen sektorin tekoälyjärjestelmiin liittyviä syrjintäriskejä sekä edistämään yhdenvertaisuutta tekoälyn käytössä. Kolmas osa sekä koko julkaisu päättyy politiikkasuosituksiin, joiden päätavoitteena on varmistaa arviointikehikon hyödynnettävyys ja käyttö osana julkisen hallinnon toimintaa. Politiikkasuosituksilla pyritään edistämään:

- Algoritmista syrjintää koskevan yleisen tietoisuuden lisääntymistä
- Poikkisektoraalista yhteistyötä tekoälyjärjestelmien vastuullisessa kehittämisessä
- Yhdenvertaisuuden edistämisen mahdollistavaa sääntelyä ja työkaluja tekoälyn käytössä.

KESKEISET KÄSITTEET JA SANASTO

Raportin oikeudelliskäsitteellisenä pohjana toimii Suomen yhdenvertaisuuslaki (1325/2014), joskin joitakin viittauksia tehdään myös tasa-arvolakiin (Laki miesten ja naisten välisestä tasa-arvosta 609/1986). Yhdenvertaisuuslain tarkoitusta heijastaen raportissa heijastetaan osatehtävien löydöksiä erityisesti kolmea yhdenvertaisuuslain asettamaa velvollisuutta vasten:

I. Yhdenvertaisuuden edistämisen velvollisuus
II. Syrjinnän kielto
III. Syrjinnän kohteiksi joutuneiden oikeusturvan tehostaminen.

Tosiasiallisen yhdenvertaisuuden edistämisen velvollisuudet koskevat viranomaisia, koulutuksentarjoajia sekä työnantajia. Syrjinnän kielto koskee kaikkia toimijoita niin julkisen kuin yksityisen toiminnan piirissä, joskaan yhdenvertaisuuslakia ei sovelleta yksityis- eikä perhe-elämään kuuluvaan toimintaan eikä uskonnonharjoitukseen. Viittaamme paikoin myös erilaisen kohtelun oikeuttamisperusteisiin (YVL 11-12 §).

Raportissa hyödynnetään myös runsaasti tutkimuskirjallisuutta sekä aiheen tutkimuskirjallisuudelle ominaista käsitteistöä, jota esittelemme alla lyhyesti. Pyrimme avaamaan keskeiset termit tarvittaessa tekstin sisällä.

Tekoäly kattaa terminologisesti tässä raportissa kaikki jossain määrin autonomisesti toimivat ohjelmistot, jotka tuottavat tulosteita ihmisen asettamien ja määrittelemien tehtävien mukaisesti. Tulosteet voivat olla muodoltaan esimerkiksi sisältöä, ennusteita, suosituksia tai päätöksiä, jotka vaikuttavat ympäristöön, jonka kanssa ohjelmisto tai laite, jonka toimintaan ohjelmisto vaikuttaa, on vuorovaikutuksessa. Menetelmät ja tekniikat, jotka voidaan lukea tämän määritelmän alle sisältävät muun muassa koneoppimismenetelmät sekä muut tilastolliset lähestymistavat, mukaan lukien Bayesilaiset menetelmät ja haku- ja optimointimenetelmät[1].

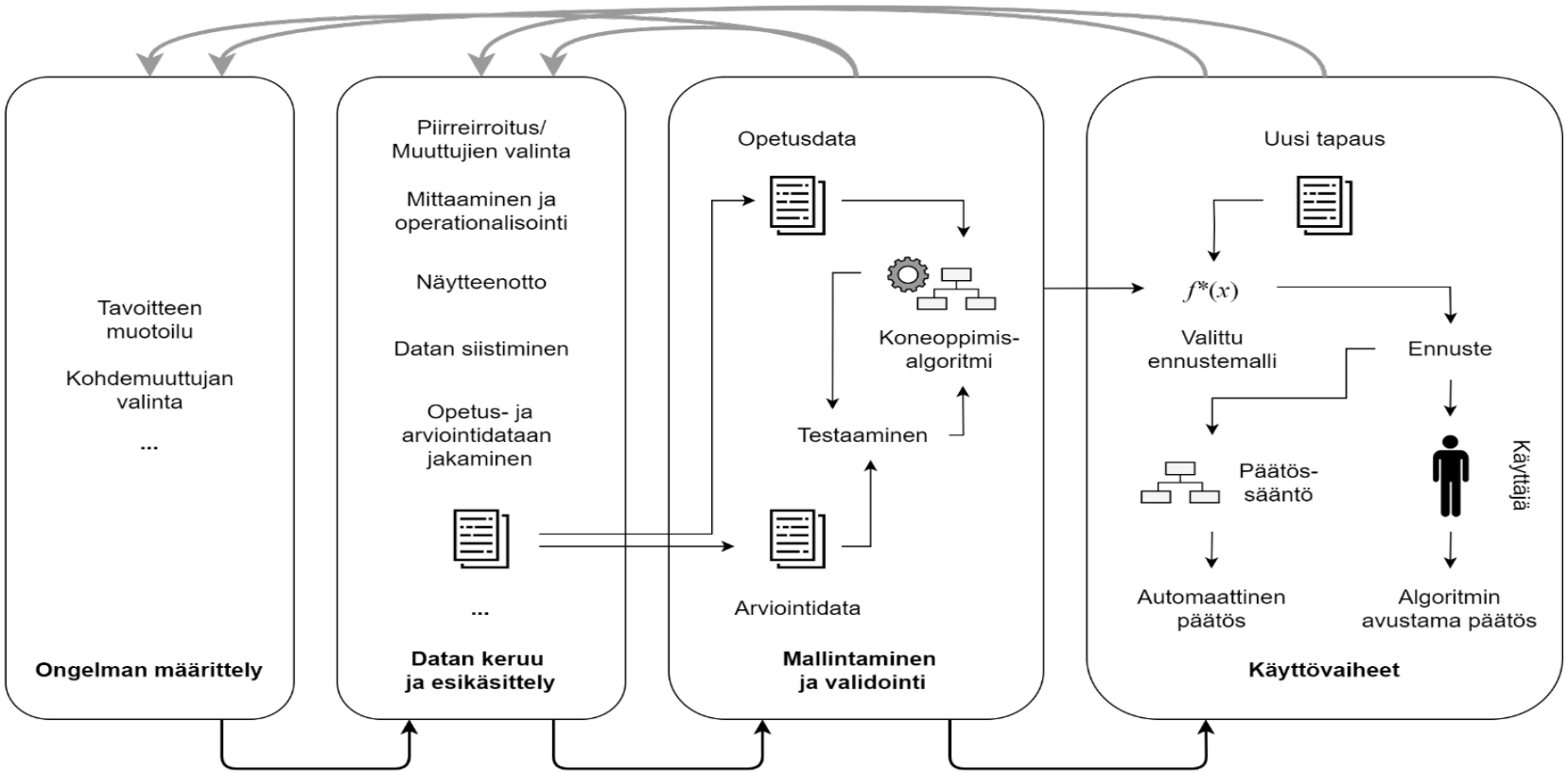
1 Raportin löydökset ovat sovellettavissa monelta osin kuitenkin laajemmin. Suuri osa löydöksistä on merkityksellistä myös (i) muiden ohjelmistojen, kuten asiantuntijajärjestelmien, jotka hyödyntävät symbolista laskentaa sekä induktiivisiin, deduktiivisiin ja logiikka- tai tietopohjaisiin lähestymistapoihin pohjautuvien ohjelmistojen sekä (ii) tilastolliseen dataan perustuvan päätöksenteon näkökulmasta.

Koneoppimisalgoritmi on joukko laskentatoimenpiteitä, joilla tuotetaan harjoitusdatan pohjalta **ennuste- tai päätöksentekomalli**, jota puolestaan käytetään tietyn ihmisen asettaman ja määrittelemän tehtävän toteuttamisen apuna (vrt. päätöksenteko tai päätöksenteon apu). Opetusvaiheessa hyödynnetty malli tunnistaa datasta toistuvia hahmoja, kuvioita ja tilastollisia korrelaatioita ja auttaa tekemään niiden perusteella ennusteita (paljon käytettyjä mallityyppejä ovat mm. lineaarinen ja logistinen regressio, päätöspuut, klusterointi ja syväoppivat neuroverkot). Koneoppimismenetelmät voidaan karkeasti jakaa (i) ohjattuun, (ii) ohjaamattomaan, (iii) puoliohjattuun ja (iv) vahvistusoppimiseen riippuen siitä, miten oppimisprosessia ohjataan ja näytetäänkö algoritmille opetusvaiheessa harjoitusesimerkkien nimikkeitä (engl. *data label*). Nimikkeet kuvaavat esimerkiksi ennustustehtävissä pohjatotuutta (engl. *ground truth*). Koneoppimisalgoritmeja voidaan käyttää erilaisiin tarkoituksiin, kuten kuvaamaan, luokittelemaan tai tuottamaan hahmoja tai sisältöä tai tuottamaan ennusteita.

Datalla viitataan kaikkeen informaatioon, joka on kerätty tai tallennettu representationaalisessa muodossa, kuten digitaalisesti. Data voi koostua esimerkiksi kuvista, videokuvasta, äänestä, symboleista ja numeroista, tekstistä ja tilastoista. Dataa voidaan myös kerätä erilaisin keinoin, kuten sensorein tai manuaalisella kirjauksella. Koneoppimisalgoritmien opettamiseen käytettävää dataa kutsutaan **harjoitus-** tai **opetusdataksi**, kun taas algoritmin testaamiseen ja arviointiin käytettävää dataa kutsutaan **arviointi-** tai **testidataksi**. Koneoppimisjärjestelmää opetettaessa ja testattaessa harjoitus- ja testidatajoukot erotetaan yhdestä suuremmasta datajoukosta soveliaaksi katsotussa suhteessa esimerkiksi ottamalla 80 % datasta harjoitusdataksi ja 20 % testidataksi. Kyseistä suhdetta kutsutaan jakosuhteeksi (engl. *train-test split*). Viittaamme tekoälysovelluksen käyttövaiheessa malliin syötettyyn dataan termillä **syötedata** ja mallin tuottamaan dataan, kuten suosituksiin tai ennusteisiin, yleisesti termillä **tuloste**.

Tekoälysovelluksen arvoketjulla viitataan yleisesti toimenpiteiden joukkoon, jonka tuloksena tekoälyjärjestelmä otetaan käyttöön. Järjestelmän käyttövaiheet voidaan lukea osaksi **tekoälytuotteen- tai palvelun elinkaarta** sikäli, kun järjestelmää päivitetään (epä)säännöllisesti tai se hyödyntää jatkuvaa oppimista (engl. *continuous* tai *online learning*). Lisäksi elinkaaren keskeinen osa on tekoälytuotteen tai -palvelun käytön lopettaminen, joka voi itsessään olla merkityksellinen yhdenvertaisuusvaikutusten arvioinnin näkökulmasta. Elinkaarinäkökulma korostaa, että koneoppimisjärjestelmien tuotanto- tai arvoketjut ovat tyypillisesti luonteeltaan monivaiheisia, iteratiivisia ja/tai syklisiä (Kuva 1). Voidaan myös erottaa kaksi päätöksentekoprosessin muotoa riippuen tekoälyjärjestelmän roolista päätöksenteossa: (a) **automaattinen tai algoritmivetoinen päätöksenteko**, jossa tekoälyjärjestelmän tuloste toimii päätöksentekoprosessin päätepisteenä eli päätöksenä ja (b) **algoritmiavusteinen päätöksenteko**, jossa tekoälyjärjestelmän tuloste toimii ihmispäätöksentekijän päätöksenteon tukena.

Kuvio 1. Algoritmiavusteinen- ja vetoinen päätöksenteko kuvattuna vaiheittain.



Vinoumalla tai **algoritmisella vinoumalla** viitataan oikeudellisesti ja moraalisesti neutraalissa mielessä kaikkiin tapauksiin, joissa jokin koneoppimismalli tai laajemmin tekoälysovellus kohtelee tiettyä ihmisryhmää systemaattisesti epäsuotuisammin. Ihmisryhmät spesifioidaan tietyn demografisen ominaisuuden (esim. ikä), fenotyypin (esim. ihonväri) tai muun ominaisuuden (esim. poliittinen suuntaus) perusteella. Epäsuotuisa kohtelu ymmärretään esimerkiksi systemaattisesti epätarkempina ennusteina tai päätöksinä, joilla on negatiivisia seurauksia yksilöille. Paikoin viittaamme vinoumilla myös tapauksiin, joissa jokin datajoukko ilmentää eroavaisuuksia ihmisryhmien välillä jonkin mitattavan tai muutoin arvioitavan asian, kuten taudin tai luottokelpoisuuden, (todennäköisyys) jakaumassa. Vinoumia voi esiintyä kuitenkin myös muussa mielessä, jotka voivat vaikuttaa järjestelmien toimintaan ja suorituskykyyn – esimerkiksi kasvojentunnistusalgoritmi saattaa toimia heikommin vähäisen valon olosuhteissa, jolloin kyse on myös vinoumasta sikäli, kun sen käyttöympäristöön liittyvän tekijä vaikuttaa sen toimintaan tai osumatarkkuuteen epätoivottavalla tavalla.

Algoritminen syrjintä tarkoittaa välitöntä syrjintää algoritmin tai tekoälyjärjestelmän käytössä ja/tai välillisesti syrjiviä vaikutuksia, joita siitä seuraa. Käsitteellisessä mielessä kaikki algoritmiset vinoumat eivät ole oikeudellisesti tai moraalisesti ongelmallisia ja ”algoritmi voi olla moraalisessa, oikeudellisessa tai sosiaalisessa mielessä vinoutunut riippuen sovelletusta normatiivisesta standardista”[2]. Algoritmiseen syrjintään johtavat vinoumat tuleekin ymmärtää erityistapauksina vinoumista termin yleisessä mielessä. Mikäli malli kuitenkin tuottaa systemaattisesti epäsuotuisia ennusteita tai luokitteluja ihmisryhmälle perustuen välittömästi tai välillisesti kiellettyyn syrjintäperusteeseen, on mahdollisesti kyseessä syrjivä algoritmi[3]. Lisäksi, kun malli tuottaa systemaattisesti epäsuotuisia ennusteita ihmisryhmälle, voidaan silti löytää periaatteessa tapauksia, joissa tämä kohtelu on moraalisesti arveluttavaa, vaikka erilainen kohtelu ei perustuisi kiellettyyn syrjintäperusteeseen. Käsitteellisesti voidaankin erottaa kolme eri tapausluokkaa (Taulukko 1). Näiden tapausten erotteleminen toisistaan on kuitenkin useissa tapauksissa monimutkaista, mikä tekeekin algoritmisen syrjinnän todentamisesta haastavaa.

2 Fazelpour & Danks, 2021.

3 Kielletyiksi syrjintäperustoiksi luetaan ikä, alkuperä, kansalaisuus, kieli, uskonto, vakaumus, mielipide, poliittinen toiminta, ammattiyhdistystoiminta, perhesuhteet, terveydentila, vammaisuus, seksuaalinen suuntautuminen tai muu henkilöön liittyvä syy (YVL 8 §).

Taulukko 1. Vinoumat, erilainen kohtelu ja syrjintä.

Käsite	Määritelmä
Systemaattinen erilainen kohtelu	Algoritmin käytössä esiintyvä systemaattinen erilainen kohtelu on oikeutettua (täyttää välttämättömyyden ja oikeasuhtaisuuden kriteerit). *Esim. rekrytoinnissa käytettävä algoritmi, joka seuloo hyvällä tarkkuudella epäpätevien hakijoiden hakemukset.* *Esim. lääketieteellisen diagnoosin tukena käytetty algoritmi, joka ilmoittaa, mikäli potilaalla on korkea riski tietylle taudille.*
Algoritminen syrjintä	Algoritmin tosiasiallisessa käytössä esiintyvä systemaattinen erilainen kohtelu, joka perustuu välittömästi tai välillisesti kiellettyyn syrjintäperusteeseen. *Esim.* Svea Ekonomin luottoluokitus-tapaus. *Esim.* SyRI-tapaus Haagin käräjäoikeudessa.
Moraalisesti ongelmallinen erilainen kohtelu	Algoritmin käytössä esiintyvä systemaattinen erilainen kohtelu, joka ei perustu kiellettyyn syrjintäperusteeseen, mutta joka on muutoin moraalisesti ongelmallista. *Esim. algoritmi evää tiettyjä palveluita yksilön ilmaistujen preferenssien, kuten Facebook-tykkäysten, perusteella4.* *Esim. suosittelualgoritmi, joka suosittelee polarisoivaa, väkivaltaista tai seksuaalista sisältöä personalisoidulla tavalla henkilön kuluttaman uutissisällön mukaan.*

4 Mann & Matzner, 2019.

1 Kansallinen kartoitus perusoikeuksiin liittyvistä tekoälysovelluksista

Tämä osio käsittelee "Tekoälyn vinoumien välttäminen: suomalainen arviointikehikko syrjimättömille tekoälysovelluksille" - hankkeen ensimmäisen vaiheen tuloksia. Hankkeen ensimmäinen osa on kansallinen kartoitus, joka toteutettiin haastattalemalla merkittäviä tekoälyä käyttäviä suomalaisia julkisia ja yksityisiä organisaatioita.

Alustavan kartoituksen keskeiset löydökset olivat:

- Tekoälyjärjestelmien käyttöönotto on edelleen maltillisella tasolla
- Tekoälyjärjestelmiin liittyvän syrjinnän tematiikkaan on herätty kohtuullisesti
- Syrjimättömyyden minimoimiseen ei ole vakiintunut selkeää viranomaisyhteistyön mallia kuten esimerkiksi tietosuojavaltuutetun kohdalla
- Julkisen ja yksityisen sektorin vaatimuksissa on merkittäviä eroja
- Globaaleilta markkinoilta hankittavien tekoälyjärjestelmien kohdalla syrjimättömyyteen on kiinnitettävä erityistä huomiota
- Vakuutusala ja turvallisuusviranomaiset erottuivat sektoreina, joilla tutkimusta tulee laajentaa

1.1 Johdanto

Tämä osio keskittyy hankkeen ensimmäiseen osatehtävään "Kansallinen kartoitus: Perusoikeuksiin liittyvät tekoälysovellukset Suomessa", joka perustuu asiantuntijahaastatteluihin relevanttien kansallisten toimijoiden kanssa. Osatehtävästä päävastuussa on Turun yliopisto. Se tarkastelee seuraavia tutkimuskysymyksiä:

1. Millaisia koneoppimiseen pohjautuvia tekoälysovelluksia on käytössä Suomessa, erityisesti julkisella sektorilla?
2. Mihin konkreettisiin syrjiviin vaikutuksiin käytetyt algoritmit ovat jo johtaneet?
3. Miten ihmisoikeudet ja syrjimättömyys on otettu huomioon käytetyissä sovelluksissa, esimerkiksi algoritmisin vaikutustenarvioinnein?

Seuraava alaluku kertoo hankkeen taustasta ja kysymyksenasettelusta, liittyen tekoälyn ja syrjinnän tärkeimpiin kysymyksiin. Toisessa luvussa esittelemme osatehtävän tärkeimmät tutkimustulokset haastatteluihin pohjautuvasta kansallisesta kartoituksesta. Lopuksi teemme yhteenvedon löydöksistä ja pohjustamme seuraavia vaiheita.

1.1.1 Tekoäly ja syrjintä

Viimeisen viiden vuoden aikana tekoälyn kehitys on edennyt ennennäkemättömällä vauhdilla. Nykyistä tekoälyn aaltoa leimaavat datan runsas saatavuus, laskentatehon kasvu ja erityisesti edistyneet koneoppimisteknologiat, kuten neuroverkot ja syväoppiminen.[5] Koneoppiminen viittaa algoritmeihin, jotka pystyvät runsaan datan avulla itsenäisesti oppimaan ja muuttamaan toimintaansa. Ne eivät siis vaadi valmista ohjelmointia eri tilanteisiin vaan mukautuvat prosessoidun opetusdata perusteella. Olemassa olevilla tekoälyjärjestelmillä on jo suuria yhteiskunnallisia vaikutuksia aina ihmisten perusoikeuksiin asti. Tekoälysovellusten aiheuttamat syrjivät seuraukset ovatkin yksi merkittävistä haasteista, jotka julkisen hallinnon tulee huomioida uuden teknologian sääntelyssä.

Koneoppimiseen perustuvia algoritmeja pidetään lupaavina koulutuksen, terveydenhuollon, rekrytoinnin, luotonannon ja monen muun palvelun parantajana. Tämä perustuu oppivien tekoälysovellusten mahdollisuuksiin tuottaa tehokkaasti yksilöllisiin elämäntapahtumiin perustuvia palveluita näitä kohdentamalla ja räätälöimällä. Samaan aikaan oppivat algoritmit kuitenkin luovat uhkia yhdenvertaisuudelle ja syrjimättömyydelle. Kun teknologiaa hyödynnetään automatisoidussa päätöksenteossa, saattaa yksipuolinen tai puutteellinen harjoitusdata johtaa syrjintään varsin huomaamattomasti. Esimerkkejä ovat kasvojentunnistusalgoritmit, jotka eivät luotettavasti tunnista tummaihoisia ihmisiä opetusdatan suppeuden vuoksi[6] sekä rekrytointijärjestelmät, jotka syrjivät naisia aiempien palkkauspäätöksien myötä. Epäedustavan harjoitusdatan ohella myös algoritmisen päätöksenteon prosessi voi itsessään johtaa vinoutumiin huonosti valittujen muuttujien ja luokittelun myötä.[7] Myös tekoälyjärjestelmän käyttöönotto epäoikeudenmukaisessa tai erilaisessa yhteiskunnallisessa kontekstissa kuin suunniteltua voi johtaa syrjintään. Algoritmiset vinoumat ovat erityisen haitallisia, sillä osana automatisoitua päätöksentekoa ne ovat vaikeasti havaittavissa ja uusintavat jo olemassa olevaa epäoikeudenmukaisuutta.

Koska koneoppiminen perustuu suurille datamäärille, voi tekoäly myös välillisesti syrjiä joitakin henkilöitä tai väestöryhmiä yhdistelemällä sinällään neutraalia tietoa, jotka liittyvät syrjintäperusteisiin. Esimerkiksi rekrytointi-algoritmi voi oppia syrjimään hakijoita perustuen heidän postinumeroonsa, joka saattaa korreloida henkilöiden etnisen ja sosioekonomisen taustansa kanssa. Algoritmisen syrjinnän ollessa pääsääntöisesti epäsuoraa haaste on, että yhdenvertaisuuteen liittyvä lainsäädäntö ei anna selkeää ja helposti sovellettavaa standardia syrjinnän arvioimiseksi. Välillinen algoritminen syrjintä on usein vaikea näyttää *prima facie* toteen, sillä se on luonteeltaan näkymätöntä ja tilastollista, eikä välttämättä

5 Ailisto ym., 2019.

6 Buolamwini & Gebru, 2018.

7 Ntoutsi ym., 2020.

noudata perinteisiä syrjintäperusteita.[8] Erityisen ongelmallista syrjinnästä tekoälykontekstissa tekee se, että algoritmit ovat käyttäjän näkökulmasta usein läpinäkymättömiä ja selittämättömiä. Jos esimerkiksi luotonantajan algoritmi hylkää lainahakemuksen automaattisesti, ei hakijan ole välttämättä mahdollisesta saada tietoa hylkäyksen syistä. Näin mahdollisen syrjinnän kohteeksi joutuneiden on vaikea näyttää syrjintää toteen ja hakea korjausta tapahtuneelle. Ratkaisuksi läpinäkymättömyyteen on tarjottu muun muassa puolueettomien ulkopuolisten tahojen toteuttamia lain vaatimia auditointeja algoritmeille ja niiden harjoitusdatalle.[9]

Yksi tunnetuimmista eurooppalaisista esimerkeistä on Alankomaiden sosiaali- ja työministeriön käyttämä SyRI-algoritmi sosiaaliturvapetoksien tunnistamiseksi. Algoritmi käsitteli suuria määriä viranomaisten keräämiä henkilötietoja arvioidakseen kansalaisten petosriskiä. Sitä käytettiin ensisijaisesti vain maan köyhimmissä naapurustoissa, vaikuttaen näin suhteettomasti maahanmuuttajiin ja sosio-ekonomisesesti huono-osaisiin. Algoritmi luokitteli useita perheitä virheellisesti syylliseksi petokseen, näin eväten sosiaalietuudet, joihin he olivat oikeutettuja. Erittäin ongelmallista oli järjestelmän läpinäkymättömyys, mikä teki sen tarkastelusta ja kansalaisten korvausvaatimuksista haastavaa. Alkuvuodesta 2020 tuomioistuin Haagissa tuomitsi algoritmin käytön laittomaksi, nojaten perusoikeuksien rikkomuksiin, erityisesti yksityisyyden ja mahdolliseen syrjinnän osalta.[10]

Huomautettakoon, että tutkimuskirjallisuudessa on keskusteltu myös mahdollisesta tarpeesta päivittää syrjintälainsäädäntöä tekoälyyn liittyvien eettisten riskien valossa[11]. Jotkut asiantuntijat ovat myös pitäneet eurooppalaisen lainsäädännön tapauskohtaisen arvioinnin ja kontekstuaalisen ymmärryksen edellytystä haasteena tekoälyn syrjintäriskien suoraviivaiselle tunnistamiselle ja esittäneet, että tutkimuskirjallisuudessa ehdotetut tilastolliset testit syrjintäriskien tunnistamiseksi tarjoavat lähtökohtaisesti vain osittaisen keinon yhdenvertaisuusvaikutusten arviointiin[12]. Tekoälysovellukset voivat skaalautuvassa käytössä kohdella suuria ihmisryhmiä epäsuotuisasti perustuen näennäisesti neutraaleihin ominaisuuksiin (esim. kulutuskäyttäytyminen) tai edellä mainittujen yhdistelmiin/intersektioihin muiden ominaisuuksien kanssa. Tämä on herättänyt keskustelua siitä, onko staattinen kiellettyjen syrjintäperusteiden lista riittävä tunnistamaan algoritmista epäoikeudenmukaista kohtelua, joka saattaa rajoittaa ihmisten perusoikeuksien toteutumista.

8 Borgesius, 2018.

9 Kim, 2017.

10 Tuomioistuimen päätös, https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:1878

11 Ibid.; ks. myös esim. Gerards & Xenidis, 2021.

12 Wachter, Mittelstadt & Russell, 2020.

Huomattakoon kuitenkin, että Suomen yhdenvertaisuuslaissa esitetty kiellettyjen syrjintäperusteiden lista on avoin (vrt. YVL 8 § ja ”muu henkilöön liittyvä syy”) ja saattaa täten tarjota paremmat välineet myös tällaisten tapausten käsittelyyn.

Miten algoritmiset vinoumat syntyvät?

Algoritmiset vinoumat voivat juontaa useasta lähteestä, kuten epäedustavasta opetusdatasta tai käytetyistä mallin muuttujista. Tutkimuskirjallisuudesta löytyy useita systemaattisia esityksiä näistä lähteistä, ja vinoumia on luokiteltu niissä eri tavoin[13].

Yksi keskeinen ero voidaan tehdä nk. tilastollisien ja yhteiskunnallisten vinoumien välillä[14]. Algoritmin voidaan ymmärtää olevan tilastollisesti vinoutunut, kun se ei estimoi mallintamaansa ilmiötä (esim. populaation ominaisuuksia) riittävällä tarkkuudella – ts., kyse on mittausvirheestä, epäedustavuudesta otoksessa tai muulla tapaa epätarkasta mallista. Yhteiskunnallisesta vinoumasta voidaan puhua, kun algoritmi ei ole vinoutunut tilastollisessa mielessä, mutta sen käyttämä malli heijastaa olemassa olevia syrjiviä rakenteita (esim. rekrytoinnissa esiintyvä syrjintä) tai tosiasiallisesta epätasa-arvoa (esim. tuloerot).

Keskeisiä tuote- tai palvelukehitysprosessin askeleita, joihin algoritmisia vinoumia on jäljitetty tutkimuskirjallisuudessa[15], ovat muun muassa seuraavat:

1. *Tehtävän määritteleminen*. Tekoälyjärjestelmän laskennallisen tehtävän ja tavoitefunktion määritteleminen.
2. *Data*. Opetusdatajoukon rakentaminen, mukaan lukien esimerkiksi datan tuottaminen ja/tai kerääminen, annotaatio eli nimeäminen ja muu datan esikäsittely opetusprosessia varten.
3. *Algoritmit ja mallit*. Algoritmien ja/tai mallien valinta ja määrittäminen, opetusprosessi, mallin testaaminen ja tulosteiden jälkiprosessointi.
4. *Käyttö*. Opetetun mallin tosiasiallinen käyttäminen, mukaan lukien virheellinen käyttö ihmiskäyttäjien toimesta, “takaisinkytkentä-loopit”, kohdepopulaation tai -kontekstin muutokset, kolmansien osapuolien suorittamat hyökkäykset ja hakkerointi.

13 Olteanu ym., 2019; Suresh & Guttag, 2020; Paullada ym., 2020; Mitchell ym., 2021; Mehrabi ym., 2021.

14 Mitchell ym., 2021.

15 Ks. yllä.

Vastaavasti mahdollisia oikeudellisesta tai moraalisesta näkökulmasta ongelmallisia vinoumia voidaan ehkäistä tai korjata eri osissa kehitys- ja käyttöprosesseja. Tutkimuskirjallisuudessa on muun muassa esitetty erilaisia matemaattisia ja proseduraalisia mittareita, joilla vinoumia pyritään tunnistamaan ja/tai kvantifioimaan mallin testaus- ja käyttövaiheissa[16]. Lisäksi löytyy erilaisia toimenpiteitä, joiden avulla vinoumia voidaan minimoida ennen ja jälkeen koneoppimisprosessin tai sen aikana[17].

Syrjintälainsäädännön suhde tekoälyyn

Aihetta käsittelevässä kirjallisuudessa voidaan erottaa toisistaan ainakin kaksi diskurssia: teknisempi algoritminen reiluus sekä algoritminen syrjintä. Näistä ensimmäinen eli algoritminen reiluus keskittyy erityisesti järjestelmien vinoumiin tietojenkäsittelytieteen ja tilastotieteen näkökulmista. Reiluus, vastuullisuus ja läpinäkyvyys (engl. fairness, accountability and transparency) koneoppimisessa on keskittynyt erilaisiin teknisiin työkaluihin vinoumien välttämiseen algoritmien suunnittelussa. Tutkimussuuntaus on ollut erityisesti suurten yhdysvaltalaisten teknologiayhtiöiden rahoittama, keskittyen erilaisiin tapoihin varmistaa algoritmien tarkkuus ja yhdenmukainen kohtelu eri väestöryhmien (esim. rotu ja sukupuoli) välillä. Algoritminen syrjintä on puolestaan oikeustieteeseen ja julkiseen päätöksentekoon nojaava diskurssi, joka on saanut enemmän huomiota Euroopassa. Syrjintä on oikeudellinen termi, joka viittaa ihmisten välittömään tai välilliseen epäoikeutetuttaan kohteluun jonkin syrjintäperusteen, kuten sukupuolen, iän tai etnisen alkuperän mukaan. Syrjintää voidaankin pitää mahdollisena seurauksena tietynlaisista vinoumista tekoälyjärjestelmissä ja/tai järjestelmien tosiasiallisesta käytöstä (ks. yllä).

Algoritminen syrjintä on erityisen ongelmallista, koska nykyisillä oikeusjärjestelmillä on vaikeuksia vastata uuteen ilmiöön. Tekoälyn merkittävä hyödyntäminen ylittää usein myös kansainväliset rajat, jolloin lainsäädännön ohjaus ei noudata ainoastaan kansallisia rajoja. Kuten Janneke Gerards ja Raphaële Xenidis[18] ovat osoittaneet, EU:n syrjintälainsäädäntö kohtaa monia haasteita algoritmisen syrjinnän kohdalla:

- Oikeudelliset aukot EU:n lainsäädännössä: vaikka koneoppimista käytetään laajalti tavaroissa ja palveluissa, mainonnassa ja rahoituksessa, EU:n syrjintä- ja tasa-arvolainsäädäntöä ei sovelleta niihin kaikkien kiellettyjen syrjintäperusteiden osalta.

16 Ks. Mitchell ym., 2021.

17 Ks. esim. Kamiran & Calders, 2012; Kamishima ym., 2012; Zhang, Lemoine & Mitchell, 2018; Kamiran, Karim & Zhang 2012; Hardt, Price & Srebro, 2016.

18 Gerards & Xenidis, 2021.

- Intersektionaalinen syrjintä: algoritminen profilointi voi johtaa syrjintään useiden eri ominaisuuksiin pohjalta välillisten muuttujien ja niiden yhdistelmien vuoksi, mutta EU:n lainsäädännössä otetaan huomioon vain yksittäiset syrjintäperusteet
- Välillinen syrjintä: vaikka suurin osa algoritmisesta syrjinnästä on todennäköisesti epäsuoraa, sen arvioiminen on tilannekohtaista ja haastavaa. On osoitettava, että näennäisesti neutraali sääntö haittaa suhteettomasti jotakin ryhmää. Lisäksi välillinen syrjintä voi myös olla perusteltua, jos se saavuttaa objektiivisen ja oikeutettavan tavoitteen, kuten algoritmin korkean osumatarkkuuden.
- Toimeenpano: prima-facie todisteiden esittäminen erilaisesta kohtelusta ja syrjinnästä on vaikeaa tekoälyn läpinäkymättömyyden vuoksi. Vastuuseen asettaminen algoritmisesta syrjinnästä on myös haastavaa tekoälyjärjestelmien monimutkaisuuden, päällekkäisyyden sekä moninaisten toimijoiden (kehittäjät, käyttöönottajat ja käyttäjät) vuoksi.

Ennakkotapaukset tasa-arvo- ja syrjintälainsäädännön soveltamisesta algoritmisiin päätöksiin puuttuvat edelleen pitkälti. EU:n lainsäädännön epäjohdonmukaisuudet ja epäselvyydet mahdollisesti rajoittavat sen kykyä käsitellä algoritmista syrjintää. Eurooppalaisen syrjintälainsäädännön perustavoitteena ei kuitenkaan ole ainoastaan estää juridista syrjintää vaan myös edistää tosiasiallista tasa-arvoa.[19] Tämä tarkoittaa muun muassa yhteiskunnan historiallisten eriarvoisuuksien huomioimista ja kitkemistä. Myös Suomessa yhdenvertaisuuslaki velvoittaa viranomaiset, oppilaitokset ja työnantajat arvioimaan toimintansa yhdenvertaisuusvaikutuksia ja edistämään yhdenvertaisuuden toteutumista.[20] Riskien lisäksi on siis huomioitava myös tekoälyn potentiaali yhdenvertaisuuden edistämiseksi. Tekoälyä voidaankin hyödyntää esimerkiksi a) nykyisen päätöksenteon syrjivien piirteiden tunnistamiseksi b) vähemmistöjen ja huono-osaisten oikeuksien toteutumista tukevien työkalujen tuottamiseksi ja c) johdonmukaisempiin ja avoimempiin päätöksenteon prosesseihin.

Yhdenvertaisuuden edistämisvelvoitteen kannalta erityisen tärkeää on tekoälyjärjestelmän yhdenvertaisuusvaikutusten arviointi sen suunnittelun, hankinnan ja käytön aikana. Yhdenvertaisuusvaltuutettu on korostanut ennakoivan vaikutustenarviointien olennaisuutta sekä jatkuvaa valvontaa oppivien algoritmien kohdalla, jo käyttöä suunniteltaessa.[21] Viranomaisten edistämisvelvoite koskee myös tekoälysovelluksia, jotka on ulkoistettu tai

19 Wachter, Mittelstadt & Russell, 2021.

20 Ks. YVL 5-7 §.

21 Yhdenvertaisuusvaltuutetun uutinen, https://syrjinta.fi/-/yhdenvertaisuusvaltuutetun-huomioita-tekoalyn-yhdenvertaisuusvaikutuksista-1.

hankittu yksityisiltä toimijoilta. Vaaditaankin lisää tietoa siitä, miten yhdenvertaisuussuunnittelun mukaisesti viranomaiset voivat poistaa esteitä yhdenvertaisuuden toteutumiselle sekä tekoälysovelluksissa että niitä hyödyntäen. Kysymys ei ole pelkästään se, toimiiko jokin tietty tekoälyjärjestelmä sisäiseltä logiikaltaan juridisesti syrjivästi, sillä tekoälyn käytön tulee hyödyttää tasavertaisesti kaikkia. Tekoälyn syrjintä- ja yhdenvertaisuusvaikutuksia onkin arvioitava aina suhteessa laajempaan yhteiskunnalliseen kontekstiin. Huomiota tuleekin kiinnittää tekoälyn käyttötarkoitukseen jo suunnitteluvaiheessa, kuten Alankomaiden esimerkki tähdentää: käytetäänkö tekoälyä apua tarvitsevien tunnistamiseen ja tukemiseen vai sääntöjen vastaisten tukien valvontaan.

1.1.2 Tilanne Suomessa

Suomella on maine digitaalisten innovaatioiden edelläkävijänä, myös tekoälyn kohdalla. Suomi on esimerkiksi kolmas Oxford Insightsin globaalissa Government AI Readiness -indeksissä vuodelta 2020.[22] Samaten Suomi on myös kolmas investoinneissa tekoälyyn komission EU-maiden vertailussa (asukasta kohti, 2018).[23] Tekoälyä hyödyntävien suomalaisten yritysten osuus onkin noussut yli kolminkertaiseksi muutamassa vuodessa.[24] Toisaalta Ipsosin komissiolle tekemässä yritysten tekoälyn hyödyntämisen vertailussa Suomi jää EU:n keskiarvon alapuolelle (Suomi: 36% vs EU27: 42%).[25] Suurissa yrityksissä tekoälyä kehitetään ja hyödynnetään enemmän kuin pienissä. Sekä Suomessa että EU:ssa julkisen rahoituksen rooli on vielä tärkeä, mm. Horisontti Eurooppa ja Business Finland tekoälyohjelmat. Vaikka tarkkaa tilannekuvaa tekoälyn käytöstä on vaikea luoda, voitaneen sanoa, että itse teknologian käyttöönotto on retoriikkaa jäljessä. Suomen edut mainituissa vertailuissa perustuivat erityisesti koulutukseen, infrastruktuuriin ja hallintoon, itse teknologian kehityksen ja yksityisten investointien seuratessa jäljessä.

Nykyiseen hallitusohjelmaan on kirjattu, että "hallitus seuraa tekoälyn käytön vaikutuksia ihmisten yhdenvertaisuuteen ja pyrkii varmistamaan, ettei tekoälyjärjestelmissä hyödynnetä välittömästi tai välillisesti syrjiviä toimintamalleja". Lisäksi mainitaan, että "Suomeen luodaan ohjeistus tekoälyn eettisestä käytöstä". Työ- ja elinkeinoministeriön vuosina 2017–2019 organisoiman kilpailukykyyn keskittyneen Tekoälyaika-ohjelman yhteydessä toimi etiikka-työryhmä sekä julkistettiin etiikkahaaste, joka aktivoi yritykset tekoälyn eettiseen hyödyntämiseen. Valtioneuvoston selonteko tietopolitiikasta ja tekoälystä

22 Oxford Insights, 2020.
23 Nepelski & Sobolewski 2020.
24 Finland's AI Accelerator, 2020.
25 Euroopan komissio, 2020.

julkaistiin joulukuussa 2018.[26] Myöhemmin tekoälystrategiatyötä on jatkanut marraskuussa 2020 lanseerattu Tekoäly 4.0 -ohjelma. Valtiovarainministeriön alaisuudessa toimii sekä Tekoälyn ja digitalisaation tutkimuksen valtakunnallinen asiantuntijaryhmä että kansallinen tekoälyohjelma AuroraAI, joka pyrkii ennakoivasti tekoälyä hyödyntäen parantamaan ihmisten eri elämänvaiheiden palvelutarjontaa. Myös sen yhteydessä toimii etiikkaryhmä. Viranomaisista erityisesti yhdenvertaisuusvaltuutettu sekä tietosuojavaltuutettu ovat huomioineet automaattisen ja tekoälyyn perustuvan päätöksenteon mahdolliset syrjivät vaikutukset.

Perusoikeuksien kunnioittamista ei kontekstuaalisuutensa vuoksi voida suoraan automatisoida tekoälyjärjestelmiin, vaan käyttötapaukset on arvioitava tilannekohtaisesti. (mm. tekoälyjärjestelmien kompleksisuus, automatisaation taso, mahdolliset virheet, niiden mittakaava ja käyttöala). Esimerkiksi oikeus sosiaaliturvaan on keskeistä sosiaalihuollossa, kun taas kokoontumisvapaus ja yksityisyyden suoja on huomioitava viranomaisten hyödyntäessä kasvojentunnistusteknologiaa. Euroopan unionin perusoikeusviraston (FRA) tutkimushanke[27] kartoitti hiljattain perusoikeuksiin mahdollisesti vaikuttavia tekoälyratkaisuja myös Suomessa, keskittyen neljään käyttökohteeseen: sosiaaliedut, ennakoiva poliisitoiminta, terveyspalvelut ja kohdistettu mainonta. Erityisesti julkisorganisaatioiden ja finanssialan toimijoiden tekoälyratkaisut herättivät tutkimuksessa huomiota. Finanssialalla onkin Suomessa nähty jo yhdenvertaisuutta rikkovia tekoälyjärjestelmiä, joihin muun muassa yhdenvertaisuusvaltuutettu on puuttunut. Vuonna 2018 eduskunnan yhdenvertaisuus- ja tasa-arvolautakunta piti kuluttajan luottokelpoisuuden tilastotieteellistä arviointia perustuen kiellettyihin syrjintäperusteisiin, kuten asuinpaikkaan, ikään ja sukupuoleen ilman yksilöllistä arviointia syrjintänä, ja asetti kieltopäätöksensä tehosteeksi merkittävän uhkasakon.[28]

Keskustelu automatisoidusta päätöksenteosta on jatkunut Suomessa myös tämän jälkeen. Muun muassa eduskunnan apulaisoikeusasiamies on todennut, että verohallinnon automatisoitu päätöksentekomenettely ei täytä perustuslain vaatimuksia, sillä se on avoimen, täsmällisen ja hyvän hallinnon periaatteiden vastainen.[29] KELA, Verohallinto ja Maahanmuuttovirasto ovat yhteisesti argumentoineet, että automatisaatio on välttämättömiä virastojen toiminnalle, ja sen poistaminen vaatisi tuhansittain uusien käsittelijöiden

26 Valtioneuvoston selonteko tietopolitiikasta ja tekoälystä (2018): Eettistä tietopolitiikkaa tekoälyn aikakaudella https://vm.fi/tietopoliittinen-selonteko.

27 Euroopan unionin perusoikeusvirasto, 2020.

28 Yhdenvertaisuus- ja tasa-arvolautakunta, täysistunto, dnro 216/2017.

29 Lehdistötiedote 26.11.2019, https://www.oikeusasiamies.fi/en/-/verohallinnon-automatisoitu-paatoksentekomenettely-ei-tayta-perustuslain-vaatimuksia.

rekrytointia.[30] He myös korostavat automaatiota käytettävän vain, kun päätös ei vaadi tilannekohtaista arviointia. Myös oikeuskansleri ja eduskunnan perustuslakivaliokunta ovat huomauttaneet automatisoituun päätöksentekomenettelyyn liittyvän useita säätelemättömiä kysymyksiä. Tästä johtuen oikeusministeriössä onkin aloitettu hallinnon automaattista päätöksentekoa koskevan yleislainsäädännön valmistelu.

Lainsäädännöllisesti suuri osa relevantista sääntelystä on odotettavissa EU-tasolta. Esimerkiksi automaattisen päätöksentekoon vaikuttaa EU:n yleinen tietosuoja-asetus eli GDPR[31], jonka mukaan kansalaisilla on oikeus olla joutumatta pelkästään automaattisen päätöksenteon, kuten profiloinnin kohteeksi, sikäli tällä on heille oikeusvaikutuksia. Henkilöitä on myös informoitava läpinäkyvästi automatisoidusta päätöksenteosta ja sen logiikasta.[32] Tekoälyn kohdalla Euroopan komissio ja parlamentti ovat vuodesta 2017 lähtien korostaneet enenevässä määrin ihmiskeskeisestä ja luotettavaa lähestymistapaa. EU:n korkean tason asiantuntijaryhmän (High Level Expert Group on Artificial Intelligence, AI HLEG) keväällä 2019 julkaistut eettisten suositukset[33] nostivat monimuotoisuuden, syrjimättömyyden ja oikeudenmukaisuuden yhdeksi luotettavan tekoälyn vaatimuksiksi.

Komission huhtikuussa 2021 julkaisema ehdotus Euroopan parlamentin ja neuvoston asetukseksi tekoälyn harmonisoiduksi säännöksiksi[34] pohjaa riskiperustaiseen näkemykseen tekoälyn sääntelystä, jossa erotellaan toisistaan ei-hyväksyttävät, suuren riskin, rajoitetun riskin ja vähäriskiset tekoälyjärjestelmät. Suurin osa sovelluksista on vähäriskisiä, eikä niitä erikseen säädellä. Suuririskisiä tekoälysovelluksia (liittyen esim. koulutukseen, rekrytointiin ja lainvalvontaan) koskevat erityiset vaatimukset muun muassa datan laadusta, järjestelmän tarkkuudesta sekä läpinäkyvyydestä ja niiltä vaaditaan näiden vaatimustenmukaisuuden arviointia. Huomionarvoista on, että vaikka lakiesityksessä viitataan syrjimättömyyteen useaan kertaan, ei sitä erikseen nosteta suuririskisten järjestelmien vaatimukseksi. Lakiesitys on kerännyt myös arvostelua; esimerkiksi Euroopan tietosuojaneuvosto ja Euroopan tietosuojavaltuutettu ovat tukeneet esitystä vahvempaa kieltoa kaikelle ihmisten biometriselle identikaatiolle (kuten kasvojentunnistukselle) julkisissa tiloissa, vedoten

30 Yle Uutiset 17.12.2019, https://yle.fi/uutiset/3-11122069.

31 Euroopan parlamentti ja Euroopan unionin neuvosto. (2016). Regulation (EU) 2016/679.

32 Automaattisen päätöksenteon kieltoon on kuitenkin GDPR:ssä välttämättömyyteen ja suostumukseen liittyviä poikkeuksia (artikla 22 ja 25(1)). Yleisesti tietosuojalainsäädännöllä on myös vaikeuksia puuttua syrjintään, sillä rajoittuu vain yksityishenkilöiden tietoihin, eikä täten päde ennustaviin algoritmeihin, jotka koskettavat ihmisjoukkoja eivätkä tunnistettavia yksilöitä. Kts. Borgesius, 2020.

33 Euroopan komissio, AI HLEG, 2019.

34 Valtioneuvoston U-kirjelmä U 28/2021 vp: https://www.eduskunta.fi/FI/vaski/Kirjelma/Sivut/U_28+2021.aspx

juuri perusoikeuksiin. Lisäksi he suosittelevat tekoälyn kieltämistä ihmisten tunnistamiseen ja jaotteluun kiellettyjen syrjintäperusteiden tai tunteiden perusteella perussopimuksen artiklan 21 mukaisesti.[35]

1.2 Kansallinen kartoitus ja haastattelut liittyen tekoälyn syrjiviin vaikutuksiin

Esittelemme tässä luvussa hankkeen ensimmäisessä vaiheessa tehtyä kartoitusta Suomessa käytössä olevista tekoälysovelluksista ja niihin liittyvistä syrjinnän riskeistä.

Haastattelut ja niiden metodologia

Hankkeen ensimmäisessä vaiheessa kartoitettiin tekoälysovellusten käyttöä yhteiskunnan ja yrityselämän eri sektoreilla haastatteluin, erityisesti keskittyen syrjinnän kysymyksiin. Kartoitusvaihe suoritettiin huhti-kesäkuussa 2021. Kartoituksesta ja haastatteluista vastasi Turun yliopisto, kartoitusraporttia on täydennetty myös muiden konsortiopartnereiden Demos Helsingin ja Tampereen yliopiston toimesta. Huhti-kesäkuussa 2021 kerättyä materiaalia on täydennetty hankkeen myöhemmissä vaiheissa soveltuvin ja tarvittavin osin.

Alustavan kartoituksen avulla tunnistetaan ja validoidaan keskeisiä syrjintään liittyviä ongelmia ja kysymyksiä eri toimialoilla. Yhdessä ohjausryhmän kanssa kartoitusvaiheen kriittisimmiksi sektoreiksi määriteltiin julkiset organisaatiot ja viranomaiset, finanssiala, terveydenhuolto sekä rekrytointiala/-toimi. Näiltä sektoreilta oli löydettävissä aikaisemman tutkimuskirjallisuuden perusteella eniten mahdollisia ongelmia syrjimättömyyden osalta.

Kartoitusta suoritettiin yleisen tiedonkeruun menetelmien, syvähaastatteluiden sekä kevyemmällä menettelyllä suoritettujen puhelinhaastatteluiden avulla. Syvähaastatteluja suoritettiin kuusi kappaletta. Yksi sovittu haastattelu peruuntui haastateltavan kieltäydyttyä. Haastattelut suoritettiin puolistruktuoituina haastatteluina joko online-yhteyksillä tai puhelinyhteyksillä. Haastatteluihin osallistui haastateltava sekä yksi tai kaksi Turun yliopiston tutkijaa. Haastattelut nauhoitettiin ja litteroitiin tarpeellisilta osiltaan. Lisäksi kevyempiä kartoittavia haastatteluita suoritettiin puhelimitse kuusi kappaletta. Näistä puhelinhaastatteluista tehtiin muistiinpanot kartoitustyötä varten.[36]

35 EDPB:n lehdistötiedote 21.6.2021, https://edpb.europa.eu/news/news/2021/edpb-edps-call-ban-use-ai-automated-recognition-human-features-publicly-accessible_fi.

36 Tarkemmat haastattelukysymykset ja sitaatit on listattu liiteosiossa raportin lopussa.

Kartoitusvaiheen aluksi määriteltiin ideaaliksi haastateltavaksi tekoälyjärjestelmien kehityksestä vastaavan osaston tai tiimin johtaja/päällikkö. Tällä tavoiteltiin soveltuvinta osaamista niin eettis-juridisista näkökulmista kuin ajankohtaisista teknologioista. Haastateltavien profiilit ovat lähellä määriteltyä ideaalia. Haastattelut luvattiin suorittaa anonyymeina henkilö-/asiantuntijahaastatteluina. Haastateltavia ei pyydetty ilmoittamaan sukupuoltaan.

Määritellyistä sektoreista julkisista organisaatioista ja viranomaisista saatiin tavoiteltu kartoitustaso. Peruuntuneet haastattelut niin finanssialalta kuin terveydenhuoltoalalta kavensivat saatua kartoituskuvaa näiltä sektoreilta, mutta näiden sektorien osalta haastatteludataa täydennetään mahdollisesti vielä elo-syyskuussa 2021. Rekrytointialalta saatiin tietoa, joka puhuisi laajemman kartoituksen ja kirjallisuuteen pohjautuvan analyysin puolesta: tämä johtuen sekä rekrytointitoimen hajanaisuudesta että käytettävien rekrytointialan järjestelmien ulkomailta tuonnista johtuen. Lisäksi kartoittavilla asiantuntijahaastatteluilla (6kpl) täydennettiin ja laajennettiin tilannekuvaa.

Haastattelujen sektorikohtainen lukumäärä sulkee pois määrällisen tutkimuksen menetelmät, joten kartoitusvaiheen analyysissa hyödynnetään sisällöllisiä ja laadullisia menetelmiä. Ensimmäisen osahankkeen tarkoitus ei ole tuottaa lopullista kuvaa, vaan tukea tutkimusprojektin seuraavia vaiheita. Lukumäärästä johtuen myös yleistettävyyden tasolla kyseessä on alustava kartoitus, jonka avulla jatkotutkimusta sekä tässä hankkeessa että mahdollisesti sen ulkopuolella voidaan luodata.

1.2.1 Haastattelut toimialoittain

Tutkittavat yhteiskunnan sektorit valikoituivat tutkimusprojektin ohjausryhmän kanssa käytyjen keskusteluiden sekä aikaisemman tutkimuskirjallisuuden perusteella. Sektoreista finanssiala ja julkisorganisaatiot olivat tutkimuksen kohteina myös Euroopan perusoikeusviraston tutkimuksessa, jossa nämä sektorit nostettiin esimerkkeinä tekoälyn soveltamisen osalta seurattavina ja ajankohtaisina.

Turvallisuusviranomaiset valikoituivat mukaan erityisesti kasvojentunnistusjärjestelmien soveltamisen perusteella. Kasvojentunnistusjärjestelmien käytöstä on tunnistettu useita perusoikeuksia mahdollisesti heikentäviä riskejä, niin yksityisyyden kuin syrjimättömyyden näkökulmasta. Rekrytointialalla käytettävät järjestelmät ovat samalla tavalla olleet aktiivisen keskustelun aihe, muun muassa Euroopan komission tekoälyä koskevassa lakiehdotuksessa[37].

37 Euroopan komissio. (2021). COM/2021/206 final.

Terveydenhuolto valikoitui tutkimuskohteeksi yhteiskunnallisen merkittävyyden ja tunnistettujen mahdollisesti syrjiviin vaikutuksiin johtavien riskien vuoksi. Tutkimuskirjallisuudessa on käyty keskustelua muun muassa diagnostiikkajärjestelmien kehitykseen käytettävän datan vinoumista[38] ja terveydenhuollon resurssien kohdistamiseen käytettävistä järjestelmistä[39].

Julkiset organisaatiot ja viranomaiset

Julkisten organisaatioiden osalta keskityttiin erityisesti turvallisuusviranomaisiksi luokiteltaviin tahoihin, ja haastatteluiden valikoitui Suomen Poliisi ja Maahanmuuttovirasto (Migri). Lisäksi haastateltiin Kansaneläkelaitoksen edustajaa. Rajavartiolaitos jatkoi vaikeasti tavoitettavana tahona, kuten myös Euroopan perusoikeusviraston tutkimuksessa vuonna 2020. Tähän on mahdollisesti syynä se, että Rajavartiolaitos jatkaa ainoana turvallisuusviranomaisena, jolla on oikeudet reaaliaikaisten kasvojentunnistusjärjestelmien käyttöön.

Finanssiala

Finanssialalla fokus oli vakuutustoiminnassa. Tavoitteena oli haastatella kahta vakuutusalan toimijaa, joista toinen toteutui tässä vaiheessa. Toisen toimijan kanssa haastattelu saatiin sovittua, mutta se selkeästi peruuntui yrityksen sisällä käydyn keskustelun myötä. Samainen toimija kieltäytyi samoilla sanoilla myös Euroopan perusoikeusviraston haastattelusta keväällä 2020. Toistunut kieltäytyminen herättää kysymyksiä kyseisen toimijan ja vakuutusalan tekoälyjärjestelmien soveltamista koskien, mikä on syytä huomioida tutkimusprojektin tuloksissa.

Terveydenhuolto

Terveydenhuollon osalta saatiin haastatteludataa julkisen tahon toimijalta Helsingin ja Uudenmaan Sairaanhoitopiiriltä (HUS). Arvokasta tietoa kertyi erityisesti tekoälyjärjestelmien käytöstä erilaisissa diagnoosi- ja analyysitehtävissä. Yksityisistä terveydenhuoltoalan yrityksistä ei vielä onnistuttu tavoittamaan ketään. Kartoitusvaihe jäi vielä puutteelliseksi erityisesti koskien terveydenhuollon resurssiallokaation järjestelmiä sekä erilaisia

38 Garattini ym., 2019..

39 Morley, ym. 2020.

OmaLääkäri-sovelluksia. Sosiaali- ja terveysministeriön maaliskuussa 2022 järjestämä seminaari ”Tekoäly STM:n hallinnonalalla - kehityssuuntia ja tulevaisuuden näkymiä” mahdollisti kartoituksen tietojen täydentämisen.[40]

Taulukko 2. Haastattelujen jakautuminen toimialoittain

Toimiala	Haastattelujen lukumäärä	Organisaatiot	Huom.
Julkiset organisaatiot ja viranomaiset	3	Suomen Poliisi Maahanmuuttovirasto Kansaneläkelaitos	Tavoiteltu myös: Rajavartiolaitos
Finanssiala	1	Vakuutustoimija 1 Vakuutustoimija 2	Toimija 2. kieltäytynyt toistuvasti haastattelusta
Terveydenhuolto	1	Helsingin ja Uudenmaan Sairaanhoitopiiri (HUS)	Yksityisiä terveydenhuolto-alan toimijoita ei tavoitettu
Rekrytointi/ rekrytointijärjestelmät	1	Teknologiatoimittaja	
Kartoittavat haastattelut	6	AuroraAI Yhdenvertaisuusvaltuutettu Opetus/koulutus Rekrytointi (2kpl) Kaupungit (Turku)	Tavoiteltiin asiantuntijoita, jotka tuntevat laaja-alaisesti tekoälyjärjestelmien soveltamista omalla sektorillaan
Haastattelut yhteensä:	6 + 6		

1.2.2 Haastatteluaineisto

Syvähaastatteluista ja kartoittavista puhelinhaastatteluista saatua aineistoa on esitelty tässä kappaleessa teemoittain asetettujen tutkimuskysymysten mukaan jäsenneltynä.

40 Sosiaali- ja terveysministeriö (2022) Tekoäly STM:n hallinnonalalla - kehityssuuntia ja tulevaisuuden näkymiä -seminaaritallenne. https://stm.fi/-/tekoalyn-tulevaisuudesta-keskusteltiin-sosiaali-ja-terveysministerion-sidosryhmatilaisuudessa.

Millaisia koneoppimiseen pohjautuvia tekoälysovelluksia on käytössä Suomessa, erityisesti julkisella sektorilla?

Taulukko 3. Käytössä olevat tekoälysovellukset

Toimija	Järjestelmä
Vakuutustoimija 1	Avustava tekoälytyökalu korvaushakemusten käsittelyyn, tunnistaa esimerkiksi puuttuvia kohtia hakemuksista ja luokittelee korvaushakemuksia valmiiksi korvauskäsittelijälle.
	Suosittelumoottori, joka avustaa myyjien työskentelyä. Kohdentaa ja suosittelee asiakkaille tai asiakkaisiin yhteydessä oleville myyjille sopivia lisävakuutuksia.
Poliisi	Kasvojen tunnistusjärjestelmä Kastu.[41] Järjestelmä muuttaa jpg:n kuvan biometrisesti käsiteltävään muotoon.
	Äänentunnistusjärjestelmä, jota käytetään kuulustelujen litterointiin. Saman tyyppinen järjestelmä kuin lääkäreillä on käytössä.
Migri	Tekoäly/NLP-pohjainen chat, käytössä asiakaspalveluun. Vielä kehitteillä, että chat-botin kanssa voisi keskustella myös puheella. Pohdinnassa on chat-botin laajentaminen myös asiointiin.
	Koneoppimismalli ollut käytössä myös skenaario- ja simulaatiorakentamisessa. Muuttuvasta datasta mallinnettu skenaarioita ja simulaatioita tulevaisuuteen. 2016 alussa mallinnettiin turvapaikkahakijoiden tilanteesta. Datan perusteella muuttuva, ei staattinen malli.
HUS	Algoritmeja, jotka nopeuttavat harvinaissairauksien hoitoa. Toisiolakia saa käyttää tähän tutkimukseen. Jos on potilaalle etua, yksityisyyteen saa puuttua. Tekoäly voi löytää potentiaalisia sairaita, jotka ovat hoidettavissa – seuraus potilaalle, jos diagnoosi viivästyy radikaali, kuolema, pysyvä laitoshoito, tjms. – kun tekoäly on antanut sairausdiagnoosin, HUSin lääkäri arvioi pseudonymisoidun tiedon. Tekoälyn osumatarkkuus 90%. Tällöin eettisesti, medikolegaalisti, ja -ekonomisesti syytä ottaa yhteyttä.
Opetus/koulutus	Oppimisanalytiikan työkaluja laajasti käytössä, esimerkkinä ViLLE-oppimisjärjestelmä[42]. Ei laajassa mittakaavassa vielä tekoälyä, mutta koneoppimismalleja ja datavisualisointiratkaisuja käytössä. Järjestelmän soveltamisessa painottuu tutkimuskäyttö.

41 Hokkanen, 2021.

42 ViLLE-oppimisjärjestelmä, https://oppimisanalytiikka.fi/ville.

Toimija	Järjestelmä
AuroraAI	Suosittelumoottori, joka yhdistää käyttäjät suositeltujen palvelupolkujen kanssa. Pääasiassa julkispalveluita ja niiden suosittelu käyttäjän eri parametrien mukaan.
	Chatbot, NLP-teknologiaa. Ulkoinen integraatiopalikka.
KELA	Konenäköön perustuvia ratkaisuja, jotka helpottavat ja tehostavat rutiininomaisia tehtäviä esimerkiksi hakemusliitteiden tai -tositteiden käsittelyssä. Luokittelee ja valmistelee liitteitä ja dokumentteja ennen siirtämistä hakemuskäsittelijälle.
Lisäys kevät 2022	KEHA-keskuksen sekä työ- ja elinkeinoministeriön toteuttama uudenlainen työelämäpalvelu Työmarkkinatori. Palvelussa käytetään tekoälyä räätälöimään käyttäjäkokemusta sekä parantamaan työnantajien ja työnhakijoiden välistä kohtaantoa.[43]

Taulukko 4. Suunnitteilla olevat tekoälysovellukset

Toimija	Järjestelmä
Poliisi	Koekäytössä hahmontunnistusjärjestelmä. Voidaan asettaa hakemaan tiettyä objektia valvontakameradatasta. Esimerkkinä annettu tietynlainen reppu ryöstötilanteessa.
Migri	Järjestelmä, joka arvioi eri hakemusten kompleksisuutta. Päätöksenteossa on kokeiltu, kuinka hyvin tekoälypohjainen malli arvioisi hakemuksen kompleksisuutta. Todennäköisyyksiä sen pohjalta. Vasta testivaiheessa. 4 RPA-tyyppistä automaatiota ja käsittelyä.
Rekrytointijärjestelmät	Hakemukset menevät järjestelmään, jossa on määrämuotoinen malli rekrytoinnissa. Tunnistetaan tiettyjä asioita, avainsanoja, tiettyjä ehdottomia vaatimuksia, löytyykö. Päätöksenteon tukitietoa. Ei tietoa onko suoraan rekrytoimista ehdottavaa järjestelmää.
KELA	Luonnollisten kielen tulkintaa: puheesta tekstiksi tai tekstistä puheeksi. Enemmänkin organisaatiolle rakennettava kyvykkyys, jota voi soveltaa eri tehtäviin.
Turun kaupunki	Turun kaupunki kommentoi olevansa pitkälti vielä datankeruuvaiheessa.

43 Niittylä, H. (2021) Vieraskynä: Tekoälyn hyödyntäminen työpaikkojen haussa. Tyomarkkinatori.fi, https://tyomarkkinatori.fi/blogit/vieraskyna-niittyla-tekoalyn-hyodyntaminen.

Toimija	Järjestelmä
Lisäys kevät 2022 STM:n seminaari	Mm. KELA ja Etelä-Karjalan Eksote ovat pilotoineet ratkaisuja, joissa dataan perustuen ennakoidaan sosiaali- tai muun tuen tarvetta nuorilla. Data-analyysiin perustuen voi tehdä ennakoivia toimenpiteitä, joita suorittaa ihmis-asiantuntija.[44, 45]
Lisäys kevät 2022 STM:n seminaari	Kotona asumisen teknologiat ikäihmisille (KATI) -ohjelmassa pilotoidaan eNero-alustaa, joka pohjautuu Gillie AI -nimiseen kaupalliseen tekoälysovellukseen. eNero-alustalla monitoroidaan kotihoitoasiakkaiden hoidon ja avun tarvetta, esimerkiksi niin kutsuttujen huoliherättäiden avulla.[46]
Lisäys kevät 2022 Opetus/koulutus	Korkeakoulutuksen puolella esimerkiksi oppimisanalytiikan mahdollisuuksia tutkittu ja edistetty AnalytiikkaÄly-hankkeessa.[47]

Mihin konkreettisiin syrjiviin vaikutuksiin käytetyt algoritmit ovat jo johtaneet?

Haastateltavilla ei ollut esittää omista organisaatioistaan tekoälyratkaisuja, jotka olisivat johtaneet syrjiviin vaikutuksiin. Yhdenvertaisuusvaltuutettu nosti esille aikaisemmin muun muassa Euroopan perusoikeusviraston tutkimuksessa esitellyt tapaukset.[48] Yhdenvertaisuusvaltuutetun edustaja nosti esiin sen, että lainsäädäntö velvoittaa julkisia organisaatioita yhdenvertaisuuden edistämisessä, mutta että yksityisellä sektorilla käytännöt voivat olla vaihtelevampia. Julkisorganisaatiot nostivatkin esiin laista tulevat velvoitteet esille. Lisäksi joitakin mahdollisia tulevia riskialueita tunnistettiin.

> *"Ei, ja tuskin tulevaisuudessakaan. Rikosasiain tietosuojalaki säätää tätä (säädetty 2018); vaikutustenarviointivaatimus, jonka mukaan poliisin on kuultava tietosuojavaltuutettua. Siinä kontrolli, jossa jos vaikka muutetaan biometrisesti tunnistettavaan muotoon kuvia, katsotaan mitä ja miten voi*

44 Koskinen, M. (2021) Tekoäly oppii ennakoimaan tuen tarpeita – olisiko siitä apua palveluihin ohjautumisessa? Sosiaalivakuutus.fi, KELA. https://sosiaalivakuutus.fi/tekoaly-oppii-ennakoimaan-tuen-tarpeita-olisiko-siita-apua-palveluihin-ohjautumisessa/.

45 SSTinfo.fi (2019) Fujitsu ja Eksote selvittävät eteläkarjalaisten nuorten syrjäytymisriskiä tekoälyratkaisulla. https://www.sttinfo.fi/tiedote/fujitsu-ja-eksote-selvittavat-etela-karjalaisten-nuorten-syrjaytymisriskia-tekoalyratkaisulla?publisherId=40135012&releaseId=69868412.

46 Sosiaali- ja terveysministeriö (2022) Tekoäly STM:n hallinnonalalla - kehityssuuntia ja tulevaisuuden näkymiä -seminaaritallenne. https://stm.fi/-/tekoalyn-tulevaisuudesta-keskusteltiin-sosiaali-ja-terveysministerion-sidosryhmatilaisuudessa.

47 Analytiikkaäly.fi-verkkosivusto. https://analytiikkaaly.fi.

48 Euroopan unionin perusoikeusvirasto, 2020.

muuntaa. Vaatimukset tietosuojavaltuutetulta tarkkoja, mutta ainahan sitä voi olla riskejä. Riskejä kuitenkin aina, käytetään järjestelmiä tai ei." - P

"Kaikkien vastuulla tehdä tasa-arvon ja yhdenvertaisuuden edistämistä. Korvausten puolella erityisesti pitää olla tarkkana. Siellä data voisi antaa vääristyneen kuvan helpommin. Jos data ei huomioi jotain, miten se vaikuttaa malliin – ja miten sen voisi korjata." - Va

"On sellainen riski, joka pitäisi ennakoida: 40 erilaista etuutta ja sitä kautta erilaisia asiakasryhmiä keskiössä, niin pitäisi pystyä palvelemaan tasapuolisesta ja tasavertaisesti, niin tuo omanlaisiaan haasteita, joissa pitää olla tarkkana, ettei tule vääristymiä." - K

Haastateltavat nostivat esiin jonkin verran varoittavia esimerkkejä ulkomailta. Opetusalalta raportoitiin Lontoossa tehdystä kokeilusta, jossa oppilaiden mikroilmeitä, poissaoloja ja käyttäytymistä mitattiin ja analysoitiin. Ruotsissa tietosuojavaltuutettu antoi sakon koululle, joka pilotoi kasvojentunnistusjärjestelmää poissaolojen mittaamisessa[49]. Lisäksi haastateltu opetusalan asiantuntija näki merkittäviä riskejä autonomisoiduissa oppilaiden arviointijärjestelmissä, joita on käytetty maissa, joissa on standardisoituihin kokeisiin perustuva arviointijärjestelmä.

Rekrytointialan haastateltavat raportoivat, että heidän tietojensa mukaan Suomessa ei laajasti vielä käytettäisi tekoälyteknologioihin perustuvia rekrytointijärjestelmiä. Tekoälyavusteisten järjestelmien kehittäminen tai käyttöönotto voi kuitenkin olla suunnitteilla monessa organisaatiossa. Rekrytointialan tekoälyjärjestelmiin on kirjallisuudessa liitetty merkittäviä syrjinnän riskejä[50], ja myös haastateltavat olivat tietoisia näistä riskeistä.

Miten ihmisoikeudet ja syrjimättömyys on otettu huomioon käytetyissä sovelluksissa, esimerkiksi algoritmisin vaikutustenarvioinnein?

Haastateltavat tunsivat ja tunnistivat yleisluontoisesti hyvin tekoälyn soveltamiseen riskejä ihmisoikeuksien, syrjimättömyyden ja etiikan näkökulmasta. Useasti mainittiin tietosuoja, yksityisyys ja yhteistyö tietosuojavaltuutetun kanssa. Läpinäkyvyyden edistämiseen haastateltavat nostivat viestinnän avoimuuden ja tekoälyn käytöstä kertovat ilmoitukset. Kaikki vastaajat korostivat, että tekoälyn käytöstä huolimatta ihmisen tulisi aina olla tekemässä lopulliset päätökset. Syrjimättömyys teemana tunnistettiin tutkittujen organisaatioiden kaikkeen toimintaan vaikuttavana asiana ja lain kieltämät syrjintäperusteet tunnettiin.

49 Datainspektionen, 2021.

50 Köchling & Wehner 2020.

Tekoälyjärjestelmiin liittyvien syrjivyyden riskien minimoimiseksi ei juuri ollut käytössä erityisiä työkaluja tai arviointikehikoita. Syrjivyyden minimoimiseen liittyvää vaikutuksenarviointia tarkasteltiin pääosin testaus- ja riskienhallintaprosessien osana. Syrjivyyden mahdollisiksi syiksi tunnistettiin tekoälyjärjestelmän kehitykseen käytetty data ja järjestelmiä kouluttavien ihmisten vinoumat. Myös itseoppivien järjestelmien suuremmat riskit verrattuna esimerkiksi sääntöpohjaisiin järjestelmiin tunnistettiin useamman vastaajan toimesta, kuten alla olevissa vastauksissa kysymyksiin:

1. Minkälaista arviointia mahdollisista syrjivistä vaikutuksista on tehty? (Onko mahdollisten syrjivien vaikutusten löytämiseksi ja minimoimiseksi laadittu strategiaa tai suunnitelmaa? Onko arvioitu syötettävän datan ja algoritmien suunnittelun vaikutuksia, yhdessä tai erikseen?)

   ”Aikalailla rajoittuu tekoälyn testaamiseen. Tuodaan testikeissejä ja katsotaan mitä vasteita tulee; mustaa laatikkoa ei voi ”ymmärtää”, joten täytyy käyttää todellisia käyttötapauksia joita järjestelmälle voitaisiin syöttää. Yleinen testaussuunnitelma. Tulee sattumaltakin vastaan ongelmakohtia, aina testaus ei riitä. Ihan kaikkea ei voida ottaa ennalta huomioon. Syrjimättömyyden testaussuunnitelmaa ei ole.” – V

2. Minkälaisia prosesseja, työkaluja ja kehikoita syrjivyyden minimoimiseksi on käytetty tai suunnitellaan käytettävän? (Minkälaisia testaus- tai monitorointikeinoja käytetään mahdollisten syrjivien vaikutusten löytämiseksi ja arvioimiseksi? Onko prosessit suunniteltu syrjivyyden arvioimiseksi tekoälyjärjestelmien koko elinkaaren ajaksi?)

   “Ei tietoinen tästä näkökulmasta; pitää käydä haastattelemassa kehittäjiä aiheesta!” -Va

   “Meillä on tuollainen etiikka-alusta/-kehys Saidot.ai:n kanssa tehtynä, ja sitä sovelletaan soveltuvin osin. Ja ylätasolla eettiset periaatteet, johon liittyy myös erilaisten vastuiden hahmottamista. Nämä muodostavat jotakuinkin työkalun/prosessit arviointiin.” - K

7. Mistä organisaationne käyttämien tekoälyjärjestelmien riski syrjiä muodostuu? Mihin syrjinnän riskit liittyvät?

   “Itseoppivia kun ei ole, niin vähenee. Riski piilee tekoälyn kouluttamisessa. Vastuu mistä datasta itse oppii, on ihmisen vastuulla. Ei voi sanoa että ”nyt meidän tekoäly vaan oppi näin”.” -K

1.2.3 Keskeiset löydökset

1. **Tekoälyjärjestelmien käyttöönotto on edelleen vaatimattomalla tasolla**

Kuten aikaisemmissa kartoituksissa on käynyt ilmi (FRA 2020, Osaamiskartoitus, Tekoälyaika), on tekoälyn soveltaminen Suomessa vielä alussa. Erityisesti julkisella sektorilla tekoälyn soveltaminen on pilotointiasteella, ja tämä tilanne voi jatkua vielä seuraavat 2–4 vuotta. Julkisella sektorilla tilanteeseen vaikuttaa erityisesti automaattista päätöksentekoa koskevan lainsäädännön uudistushanke, jonka tuloksia moni organisaatio vaikuttaa odottavan.

Yksityisellä sektorilla suurin osa tekoälykokeiluista tai -käyttöönotoista toimivat korostetun kapealla alueella tai kohdistuvat yhteiskunnallis-eettiseltä merkitykseltään vaarattomampiin käyttökohteisiin (esim. tuotanto- ja prosessioptimointi). Yksityisellä sektorilla olisi tärkeä kartoittaa laajemmilla tutkimuksilla, esimerkiksi kyselytutkimuksin, missä määrin tekoälyn soveltaminen jakaantuu itse ja/tai teknologiapartnerin kanssa tuotettavien tekoälyjärjestelmien varaan, että toisaalta ulkopuolelta valmiina järjestelminä ostettaviin palveluihin/tuotteisiin.

Tekoälyn soveltaminen alkeellinen taso ei kuitenkaan vähennä tämän tutkimusprojektin merkitystä, päinvastoin se antaa mahdollisuuden varautua tekoälyn käytöstä johtuvan syrjinnän minimoimiseen jo etukäteen. Kaikki tutkitut organisaatiot ovat kuitenkin etenemässä tekoälyn soveltamisen kanssa, keskimäärin muutaman seuraavan vuoden aikana.

2. **Tekoälyjärjestelmiin liittyvän syrjinnän tematiikkaan on herätty kohtuullisesti**

Kaikki vastaajat ja heidän edustamansa organisaatiot olivat tietoisia tekoälyjärjestelmiin liittyvistä mahdollisista syrjinnän ongelmista. Jokainen haastateltava osasi myös esittää joitakin toimia, joihin on ryhdytty syrjinnän estämiseksi teknologiaratkaisujen yhteydessä. Monen organisaation kohdalla toimet olivat kuitenkin yleisluontoisia (esim. keskustelut ja koulutus syrjimättömyydestä) tai symbolisia (esim. ”tämä on otettu huomioon”).

Erilaisia keinoja tai työkaluja syrjimättömyyden estämiseksi käytetään vähän tai enintään vakiintumattomasti. Esimerkiksi erilaiset arviointikehikot tai edes tarkistuslista-tyyppiset työkalut eivät olleet vakiinnuttaneet paikkaansa tutkituissa organisaatioissa. Tähän saattaa vaikuttaa tekoälyteknologioiden soveltamisen vähäisyys (johtopäätös 1). Myöskään luontevaa viranomaisyhteistyön mallia ei ollut muodostunut tekoälyteknologian ja syrjimättömyyden välisiin kysymysten ratkaisemiseksi samalla tavoin kuin tietosuoja-asioissa tietosuojavaltuutetun kanssa (johtopäätös 3).

Suurin osa haastateltavista ei ollut syvällisesti perehtynyt tämän tutkimusprojektin kysymyksiin. Haastateltavat käsittelivät syrjimättömyyttä kuitenkin organisaatioita laajasti läpileikkaavana teemana. Haastattelukysymysten lisäosuutta eli hieman erikoistuneempia kysymyksiä ei ollut juurikaan hedelmällistä käydä läpi haastateltavien kanssa. Tämä on tosin siinä mielessä ymmärrettävää että, haastateltujen kommentteja referoiden, tekoäly- tai teknologiaratkaisuihin liitetyt syrjimättömyyden kysymykset ovat tulleet monelle tutuksi vasta viimeisen 1-2 vuoden aikana.

3. **Syrjimättömyyden minimoimiseen ei ole vakiintunut selkeää viranomais-yhteistyön mallia kuten esimerkiksi tietosuojavaltuutetun kohdalla**

Haastatteluissa tietosuojavaltuutetun kanssa tehty yhteistyö mainittiin useampaan otteeseen, mutta syrjimättömyyden alueelta yhtä luontaisia yhteistyötahoja tai -malleja ei löytynyt. Tietosuojavaltuutetun kanssa yhteistyötä oli tehty niin lainsäädännön velvoittamana kuin vapaamuotoisemmin: useimmiten riskien- tai vaikutustenarvioinnin nimikkeellä.

Laajempaa keskustelua ansaitseva kysymys on, että minkälainen yhdenvertaisuusvaltuutetun rooli tulisi olla tekoälyn soveltamiseen liittyvissä syrjimättömyyskysymyksissä. Yhdenvertaisuusvaltuutetun edustajan kanssa käyty keskustelu todensi myös sen, että valtuutetulla ei ole näihin kysymyksiin osoitettua roolia eikä välttämättä tarpeeksi oikeanlaisia resursseja roolin perustamiseksi. Tutkimusprojektin edetessä tulisi lisäksi perehtyä yksityiskohtaisemmin minkälaisia yhteistyön muotoja eri organisaatiot ovat kehittäneet tietosuojavaltuutetun kanssa.

Yhteistyötä yliopistojen tai tutkimusinstituutioiden kanssa ei myöskään mainittu. Yhtäläisesti voidaan pohtia, onko tutkimusinstituutioiden ja esimerkiksi julkisten organisaatioiden välillä toimivia yhteistyömalleja tekoälyjärjestelmiin liittyvän mahdollisen syrjinnän, tai laajemmin tekoälyn eettisten näkökulmien, kaltaisten nopeasti merkittäväksi nousevien ilmiöiden käsittelemiseksi. Olemassa olevista yhteistyömalleista on syytä mainita valtiovarainministeriön asettama tekoälyn ja digitalisaation tutkimuksen valtakunnallinen asiantuntijaryhmä.[51]

51 Valtiovarainministeriö, Tekoälyn ja digitalisaation asiantuntijaryhmä https://vm.fi/tekoalyn-ja-digitalisaation-asiantuntijaryhma.

4. **Julkisen ja yksityisen sektorin vaatimuksissa on merkittäviä eroja**

Erityisesti julkisorganisaatiot mainitsivat suuresta erosta julkisen ja yksityisen sektorin organisaatioille esitettävistä vaatimuksista, esimerkiksi syrjimättömyyden estämiseen liittyen. Tämä juontuu selkeästi lainsäädännöstä ja yhdenvertaisuuslain asettamista velvoitteista.

Toisaalta yksityisten ja julkisten palvelun raja teknologioiden kehittyessä voidaan nähdä häilyvän. Näin erityisesti globaalien tuotantoketjujen ja mikrotyön lisääntymisen myötä (esim. Amazon Turk), sekä valmiina ostettavien tekoälyjärjestelmien ja niiden yhdistelmien myötä, joita myös julkiset toimijat hyödyntävät. Samalla myös suurimmat digitaaliset alustat ovat saaneet niin ratkaisevan markkina-aseman ja suuren yhteiskunnallisen roolin, että ne voidaan merkitykseltään rinnastaa julkisiin toimijoihin.

Jatkotutkimuksissa olisi hyvä tarkastella, että muuttuuko tai kasvaako tämä ero julkisten ja yksityisten toimijoiden välillä erityisesti tekoälyteknologioita sovellettaessa.

5. **Globaaleilta markkinoilta hankittavien tekoälyjärjestelmien kohdalla syrjimättömyyteen on kiinnitettävä erityistä huomiota**

Erityisesti yksityisellä sektorilla tullaan tulevina vuosina soveltamaan laajasti globaaleilta markkinoilta hankittavia tekoälyjärjestelmiä. Soveltaville organisaatioille näiden järjestelmien käyttö muodostaa mustan laatikon ongelmia sekä kysymyksiä vastuun jakautumisesta.

Myöskään julkisorganisaatiot ja viranomaiset eivät ole täysin immuuneja ostettavien algoritmien ongelmille. Esimerkiksi Suomen Poliisi soveltaa ostettua algoritmia.[52] Vaikka Poliisin käytössä olevan algoritmin hankinta- ja tutkimusprosessi vaikutti haastattelun perusteella harkitulta, haasteena on pitää hankintaprosessi yhtä hyvänä a) kaikissa hankinnoissa b) monimutkaisempien järjestelmien hankinnassa c) muissa julkis- ja viranomaisorganisaatioissa.

Euroopan unionin sääntely liittyen suuririskisiin tekoälysovelluksiin voi olla yksi avain asian ratkaisemiseksi. Kansainvälisen yhteistyön ohella arviointikehikot, esimerkiksi tämän hankkeen kohdalla voivat antaa työkaluja läpinäkyvään laadun varmistamiseen.

52 Tarkennettakoon, että tässä ei ole kyse kohua herättäneestä ClearView-kokeilusta, joka oli ilmeisesti muutaman henkilön harkitsematon kokeilu heidän omalla datallaan.

6. **Vakuutusala ja turvallisuusviranomaiset erottuivat sektoreina, joilla kartoitusta ja tutkimusta tulee laajentaa**

Vakuutusalan tekoälyn soveltamisesta on ollut haastavaa saada tarkkoja tietoja. Yksi vakuutusalan toimija kieltäytyi niin tämän tutkimusprojektin kuin Euroopan ihmisoikeusviraston haastatteluista, jotakuinkin kategorisesti käyttäen samoja viestimuotoiluja.

Haastateltu suomalaisen vakuutusyhtiön edustaja avasi joitain käyttökohteita ja -tapoja, joissa he soveltavat ja suunnittelevat soveltavansa tekoälyä, mutta nämä kohdistuivat pääosin vielä eettisesti tarkasteltuna vähemmän merkityksellisiin, lähinnä rutiininomaisiin lomaketarkistuksiin. Joitain viitteitä siitä, että tekoälyä saatettaisiin tulevaisuudessa käyttää myös esimerkiksi vakuutuskorvaushakemusten käsittelyssä, kuitenkin löytyi. Pienemmät toimijat saattavat liikkua teknologiakäyttöönotoissaan hitaammin ja rajoitetummin, joten syrjinnän riskit voivat olla voimakkaampia alan isojen toimijoiden ratkaisuissa. Mahdolliseksi syrjinnän riskialueeksi havaittiin kieli: automaattiset korvaustekstin lukujärjestelmät lukevat paremmin virheetöntä ja hyvin muotoiltua kieltä. Tutkimusprojektin jatkovaiheissa on tarpeen vielä laajentaa kartoitusta vakuutusalan tekoälyratkaisujen riskeistä sekä mahdollisista suositeltavista ratkaisumalleista esimerkiksi läpinäkyvyyden lisäämisessä.

Turvallisuusviranomaiset ovat puolestaan julkisia toimijoita ja heidän toimintaansa rajoittavat monet lait. Haastatteluissa ei ole havaittu ongelmallisia prosesseja tai asennoitumista tekoälyn soveltamista kohtaan, vaan suunnittelu, käyttöönotto ja käyttö vaikuttavat asianmukaiselta. Sektorilta löydettävät tekoälyn sovellusalueet ovat kuitenkin kriittisempiä myös syrjimättömyyden näkökulmasta. Esimerkiksi kasvojentunnistusteknologia on tunnistettu tutkimuskirjallisuudessa erityisen ongelmalliseksi käyttökohteeksi yhdenvertaisuuden kannalta. Sektorin lisäkartoituksessa tulisi pyrkiä syvemmälle teknisiin yksityiskohtiin ja teknisen toteutusosaston tekemisiin.

Rajavartiolaitoksen käyttämistä tekoälyratkaisuista on ollut haastava saada tietoa. Rajavartiolaitoksella on edelleen ainoana turvallisuusviranomaisena oikeudet reaaliaikaisen kasvojentunnistusteknologian käyttöön. Vaikka tämän kartoituksen osalta kyseessä ovat todennäköisesti olleet satunnaiset haasteet viranomaista edustavien henkilöiden kontaktoinnissa, ei Rajavartiolaitoksen viestintää tekoälyteknologioiden käytöstä muutenkaan voi tutkijan näkökulmasta pitää järin läpinäkyvänä. Samalla on huomioitava, että tälle vaiteliaisuudelle voi olla hyvät perusteet.

1.3 Osuuden yhteenveto

Tutkimusprojektin kansallinen kartoitusosuus antoi tietoa siitä mikä on tekoälyteknologioiden soveltamisen tilanne projektin kannalta keskeisimmillä sektoreilla. Kartoituksessa tarkasteltiin syrjivien vaikutusten minimoimiseksi keskeisiä asioita, kuten organisaatioiden

näkemystä tekoälyn vastuullisesta ja eettisestä kehittämisestä, asiantuntemusta syrjimättömyyden teemoista ja edistämisestä, sekä tekoälyratkaisujen suunnittelu-, kehitys- ja testausprosesseista. Lisäksi selvitettiin, onko organisaatioihin vakiintunut käytäntöjä tai työkaluja syrjivien vaikutusten arvioimiseksi tai estämiseksi.

Keskeisiä, haastatteluista nousevia jatkokysymyksiä ovat muun muassa seuraavat:

- Miten käsitellä **kansainvälisten tekoälyjärjestelmien soveltamista syrjimättömyyden näkökulmasta?** Ulkomailta ostetut tai hankitut järjestelmät ovat verrattain läpinäkymättömiä, usein ns. mustia laatikoita. Erityisesti yksityisellä sektorilla eri tekoälyjärjestelmien soveltaminen on lähtökohtaisesti vapaata eikä tarkkaan määriteltyä. Haasteena on myös seurata tällaisten järjestelmien käytön yleistymistä.
- Mikä on **yhdenvertaisuusvaltuutetun rooli** tekoälyteknologian syrjimättömyyden varmistamisessa ja organisaatioiden kanssa tehtävässä yhteistyössä? Tällä hetkellä valtuutetulla ei ole juuri resursseja ja asiat on kohdennettu yhden henkilön muihin tehtäviin.
- Miten käsitellä **rakenteellista syrjintää tekoälyn kohdalla?** Tämä tutkimusprojekti tarkastelee pitkälti lainsäädännöllisesti määriteltyä syrjintää ja yhdenvertaisuutta, mutta sikäli kun tekoäly ei noudata laissa määriteltyjä syrjintäperusteita, on käsiteltävä myös laajemmin yhteiskunnallista eriarvoisuutta.
- Kuinka hyvin **turvallisuusviranomaisten, hallinnon ja muiden julkisorganisaatioiden välinen keskusteluyhteys toimii?** Tällä hetkellä tutkijoiden näkökulmasta on hankala saada merkityksellistä tietoa esimerkiksi tunnistusjärjestelmien käytöstä. Turvallisuusviranomaisilla voi olla hyviä perusteita tiedonsaannin rajoituksiin, mutta tutkimuksen ja läpinäkyvyyden näkökulmasta olisi silti tärkeää saada koherentti kuva vallitsevasta soveltamisen tilasta.
- Löytyykö eri **organisaatioista tarvittavaa osaamista** tekoälyjärjestelmien mahdollisesti syrjivien vaikutusten huomioimiseen ja minimoimiseen? Vaikka organisaatioiden asiantuntemus syrjimättömyyden teemoista on kohtuullista, ovat tarkemmat työkalut ja lähestymistavat tekoälyn syrjimättömyyden arviointiin vielä kehittymättömiä.

2 Syrjinnän riskit ja yhdenvertaisuusperiaate tekoälyjärjestelmissä

2.1 Johdanto

Tässä osuudessa kootaan löydöksiä hankkeen toisessa osatehtävässä – ”Syrjinnän riskit ja yhdenvertaisuusperiaate tekoälyjärjestelmissä” – suoritetusta kartoituksesta, jossa kerättiin ja analysoitiin viranomais-, tutkimusryhmien sekä toisen ja kolmannen sektorin toimijoiden raportteja sekä tutkimuskirjallisuutta tekoälyjärjestelmien syrjintäriskeistä, niiden tunnistamiseen ja ehkäisemiseen kehitetyistä menetelmistä sekä mahdollisista haasteista, joita kyseisten menetelmien käyttöön liittyy. Tässä analyysissa tuotettu tietopohja palvelee myös hankkeen kolmannen osatehtävän tuotosta, arviointikehikkoa syrjimättömille tekoälysovelluksille, minkä takia käsittelylukuihin on sisällytetty myös näkökulmia ja suosituksia, joiden pohjalta arviointikehikkoa rakennettiin hankkeen kolmannessa vaiheessa.

Kartoituksen viitekehys on kuvattu osiossa ”Keskeiset käsitteet ja sanasto” -osiossa. Kartoituksen tavoitteena oli vastata seuraaviin kysymyksiin:

1. *Mitä syrjintäriskejä koneoppimista hyödyntävissä sovelluksissa on tunnistettu?*
2. *Mihin kohtaan tekoälysovellusten elinkaarta tai arvoketjua nämä riskit voidaan paikantaa?*
3. *Mitä keinoja riskien ja syrjivien vaikutusten arvioimiseksi ja ehkäisemiseksi on tunnistettu?*
4. *Mitä haasteita liittyy syrjintäriskien arviointia ja ehkäisemistä tukevien menetelmien käyttöön?*

Löydökset on jaoteltu viiteen osioon, joiden kuvaukset ja keskeiset löydökset on esitelty lyhyesti alla.

Tekoälysovellusten syrjintäriskien profiili -osio antaa kuvauksen yleisistä, yhdenvertaisuuslain näkökulmasta merkityksellisistä löydöksistä, jotka auttavat rakentamaan pohjaa tulevien osioiden yksityiskohtaisemmalle tarkastelulle. Osiossa esitellään välittömän, välillisen ja moniperusteisen syrjinnän riskejä tekoälyn käytössä yleisellä tasolla ja

kuvaillaan konteksteja, joissa riskejä on tunnistettu. Osiossa tarkastellaan myös vaikutuksia tosiasialliseen tasa-arvoon sekä kysymyksiä rakenteellisesta ja intersektionaalisesta syrjinnästä.

> **Löydökset:** Kartoituksen perusteella erityisesti välillinen (mukaan lukien välillisesti moniperusteinen) syrjintä nousee keskeiseksi huolenaiheeksi tekoälyn käytössä läpi sektoreiden ja toimialojen. Johtuen välittömän syrjinnän eksplisiittisestä luonteesta, sen riskejä on ainakin periaatteessa huomattavasti helpompi tunnistaa ja arvioida kuin välillisen syrjinnän. Tekoälyjärjestelmien luonne voi kuitenkin mahdollistaa tahallisen syrjinnän naamioimisen käyttämällä nk. *proxy*- tai korvannaismuuttujia päätöksenteossa, jotka korreloivat kiellettyjen syrjintäperusteiden kanssa. Vaikutuksien arvioijan näkökulmasta keskeiseksi haasteeksi muodostuukin rajanveto kielletyn syrjintäperusteen kanssa oikeuttamattomasti ja oikeutetusti limittyvien näennäisesti neutraalien muuttujien välillä. Suomi on erityisasemassa yhdenvertaisuuslainsäädäntönsä erityisluonteen ansiosta sikäli, että kiellettyjen syrjintäperusteiden listan avoimuus mahdollistaa suhteellisen laajat toimet tosiasiallisen yhdenvertaisuuden edistämiseksi. Tekoälyn yhdenvertaisuus- ja tasa-arvovaikutusten ympärillä käyty keskustelu keskittyy huomattavasti syrjinnän kieltoon, kun taas keskustelu mahdollisuuksista edistää tosiasiallisen yhdenvertaisuuden toteutumista on suurelta osin marginaalista.

Tekoälysovellusten syrjintäriskien teknologiset syyt -osio muodostaa hienojakoisemman kuvauksen algoritmisista vinoumista ja niiden lähteistä teknologian tasolla. Kyseessä on tekninen kuvaus mahdollisiin syrjiviin vaikutuksiin johtavista vinoumista sekä (sosio-) teknisistä syistä, jotka ovat merkityksellisiä näiden vaikutusten tunnistamisen ja ehkäisemisen kannalta.

> **Löydökset:** Kartoitus osoittaa, että teknologian tasolla voidaan tunnistaa joukko käsitteellisesti eroteltavia, mutta käytännössä limittyviä ja päällekkäisiä riskejä syrjiville vaikutuksille. Tekoälyjärjestelmien suorittamien tehtävien määrittely, datan keruu ja esikäsittely, hyödynnetyt algoritmit ja mallit sekä niiden mahdollinen, sisäisestä monimutkaisuudesta johtuva selitettävyyden puute osoittautuvat merkityksellisiksi syrjivien vaikutusten tunnistamisen ja ehkäisemisen kannalta. Syrjivät vinoumat syntyvätkin useimmiten tekoälyjärjestelmien arvoketjussa erilaisten vinoumien yhteisvaikutuksesta. Toisaalta arvoketjussa tunnistettuja vinoumia voidaan korjata tarkoituksellisesti rakentamalla järjestelmiin positiivisia yhdenvertaisuusvaikutuksia edistäviä vinoumia. Neutraalia tai ”vinoutumatonta” järjestelmää ei kuitenkaan ole olemassa, ja

yhdenvertaisuusvaikutusten tehokasta ja vaikuttavaa arvioimista voisikin osaltaan edesauttaa ”teknologian neutraalisuuden” diskurssista luopuminen.

Teknologian ulkopuoliset syyt syrjintäriskeille -**osio** tarkastelee syrjivien vaikutusten toteutumiseen vaikuttavia, teknologian ulkopuolisia tekijöitä, jotka ovat merkityksellisiä mahdollisten syrjivien vaikutuksen tunnistamisen ja ehkäisemisen kannalta. Osion sisällössä korostuu, että tekoälysovellusten mahdollisia syrjiviä vaikutuksia ei voida tunnistaa, saati ehkäistä tarkastelemalla yksinomaan niiden teknistä koostumusta ja rakennetta (esim. dataa tai algoritmeja). Myös laajempi sosiotekninen kehys ja sosiaalinen konteksti on otettava huomioon.

Löydökset: Kartoituksesta selviää, että tekoälyjärjestelmien syrjintäriskien luonteeseen ja mahdollisten syrjivien vaikutusten realisoitumiseen vaikuttavat useat sosiaaliset, kulttuuriset ja teknologiset tekijät. Nämä tekijät muodostavat haasteen universaalisti sovellettavalle ja yksiselitteiselle lähestymistavalle syrjivien vaikutusten tunnistamiseen ja ehkäisemiseen. Tällaisiin tekijöihin lukeutuvat muun muassa järjestelmän lähtökohtainen käyttötarkoitus ja sen oikeutettavuus; järjestelmän käyttäjien tulkinnat ja toiminta; kohdepopulaatio; käytön skaala sekä ajassa ilmenevät, dynaamiset muutokset edellä mainituissa. Mahdollisten syrjivien vaikutusten realisoitumisen riippuvaisuus tästä laajemmasta sosioteknisestä kontekstista korostaa yhdenvertaisuusvaikutustenarvioinnin paikallisuutta ja tapauskohtaisuutta. Huomiota on kiinnitettävä myös laajemmin järjestelmän tukemiin institutionaalisiin toimintoihin ja päätöksentekoprosesseihin sekä jatkuvuuteen syrjivien vaikutusten ennakoimisessa, arvioinnissa ja ehkäisemisessä.

Tekoälysovellusten syrjintäriskien arvioiminen ja hallinta -osio esittelee aineistossa visioituja menetelmiä tekoälysovellusten syrjintäriskien arvioimiseksi ja ehkäisemiseksi. Raporteissa suositellut menetelmät sisältävät niin teknisiä menetelmiä kuin laajempia proseduraalisia, organisaatiotason ja lainsäädännöllisiä keinoja.

Löydökset: Kartoituksen perusteella datan laadunarviointi sekä algoritmien yksilö- ja ryhmäkohtaisten vaikutusten tilastolliset testit, joilla pyritään parantamaan algoritmien ”reiluutta”, nousevat aineistosta esille keskeisinä syrjinnän ehkäisemisen keinoina. Näitä menetelmiä voidaan hyödyntää myös osana laajempia järjestelmien auditointi- ja vaikutuksenarviointiprosesseja, kuten Euroopan yleisen tietosuoja-asetuksen (GDPR) edellyttämää tietosuojavaikutustenarviointia (DPIA), tasa-arvovai kutustenarviointiprosessia (”equality impact assessment”) tai algoritmien vaikutuksenarviointiprosesseja (”algorithmic impact assessment”).

Aineistosta voidaan myös tunnistaa algoritmisten päätöksentekoprosessien läpinäkyvyyden ja niiden riitauttamisen mahdollisuuden merkitys ihmisten perusoikeuksien toteutumiselle.

Haasteita syrjintäriskien hallinnalle -osio tarkastelee kriittisesti edellisessä osiossa esiteltyjä menetelmiä ja kuvaa käytännöllisiä haasteita, joita on tunnistettu liittyvän syrjintäriskien arvioimiseen ja hallintaan.

Löydökset: Kartoituksen perusteella on selvää, että tekoälyjärjestelmien syrjiviä vaikutuksia ei voida tunnistaa yksinomaan teknologian tasolla (esimerkiksi tarkastelemalla opetus-, arviointi-, syöte- ja tulostedataa sekä algoritmeja ja malleja irrallaan niiden käyttökontekstista). Tämä johtuu yhtäältä EU:n ja tarkemmin Suomen yhdenvertaisuuslainsäädäntöön sisäänrakennetusta kontekstuaalisuudesta ja faktuaalisen tarkastelun vaatimuksesta ja toisaalta tekoälyjärjestelmien sosioteknisten käyttöympäristöjen ja kohdepopulaatioiden dynaamisten muutosten asettamista haasteista. Vaikka kartoituksessa tunnistetut keinot voivat parhaimmillaan tukea vaikuttavia arviointi- ja ehkäisemistoimenpiteitä, osoittautuivat lainmukaisuusvaatimuksia (erit. tietosuoja- ja yhdenvertaisuuslainsäädännön osalta) ja toimijoiden vastuita koskevan tiedon puute sekä ohjeistavien standardien uupuminen organisaatiotasolla haasteeksi.

2.2 Metodologia

Kartoituksessa analysoitiin yhteensä 23 raporttia sekä tutkimuskirjallisuutta aiheesta. Kartoituksen löydöksiä tarkastellaan erityisesti yhdenvertaisuuslakia 1325/2014 sekä kentän tutkimuskirjallisuutta vasten sekä paikoin tasa-arvolain 1986/609 näkökulmasta. Aineiston heterogeenisyys maantieteellisten ja juridisten kontekstien suhteen sekä algoritmista syrjintää koskevan oikeuskäytännön puute voidaan kuitenkin mainita mahdollisina rajoitteina joidenkin löydösten yleistettävyydelle Suomen kontekstiin. Löydöksiä pidetään kuitenkin edustavana yleisestä diskurssista, parhaista käytännöistä sekä tutkitusta tiedosta algoritmisen syrjinnän tematiikan ympärillä. Raportin sisältö keskittyy yhdenvertaisuutta ja syrjintää koskeviin kysymyksiin niiden oikeudellisessa merkityksessä, mutta osatehtävän tarkoitusta heijastaen esittelemme myös yleisempiä sosiaalieettisiä kysymyksiä ja näkökulmia aiheen ympäriltä. Raportissa huomioidaan Euroopan komission ehdotus COM(2021)206 tekoälyä koskevista yhdenmukaistetuista säännöistä (”Tekoälysäädös”), jotka oletettavasti tulevat osaltaan vastaamaan tekoälyä ja yhdenvertaisuusperiaatetta koskeviin avoimiin kysymyksiin ja näitä koskevan käytännön muodostumiseen.

Tutkimuskysymyksiä (ks. yllä) heijastavien relevanssikriteerien (ks. alla) tarkoitus oli rajata kartoitus aineistoon, jossa käsiteltiin ensisijaisesti tekoälyä ja syrjintää, erityisesti yhteydessä oikeudellisiin syrjintää koskeviin käsitteisiin. Euroopan yhdenvertaisuuslainsäädännön kontekstissa sekä vertailtavissa yhteyksissä (ml. esim. Yhdistyneen kuningaskunnan ja Yhdysvaltojen yhdenvertaisuuslainsäädännön kontekstissa). Aineistossa ”algoritmisia vinoumia” käsiteltiin kuitenkin monessa merkityksessä ja niin yhdenvertaisuuden kuin tasa-arvonkin näkökulmista. Tämä kuvastanee vallitsevaa epätietoisuutta siitä, missä tapauksissa algoritmiset vinoumat voidaan katsoa syrjiviksi (juridisessa merkityksessä).

Aineisto kerättiin suorittamalla hakuja Euroopan unionin julkaisuportaalissa, Google Scholarissa sekä Googlen hakukoneella. Tarkoituksena oli tuottaa otos, joka sisältäisi edustavasti eri sektoreiden toimijoiden – esim. tutkimusryhmien, viranomaisten sekä yksityisen ja kolmannen sektorin toimijoiden – kokoavia raportteja aiheesta. Haussa käytettiin seuraavien hakusanojen yhdistelmiä: ”algorithm”,”artificial intelligence”, ”AI”, ”machine learning”, ”bias”,”discrimination” ja ”equality”. Haut tuottivat laajasti tuloksia. Esimerkiksi Euroopan unionin julkaisuportaalissa löytyi haulla 27543 tekstiä. Rajasimme haun kussakin hakukoneessa 10 ensimmäiseen hakusivuun. Relevanssikriteerien avulla karsittiin alustavassa kartoitusvaiheessa pois soveltumatonta aineistoa, kuten yksittäisiä tutkimus- ja uutisartikkeleita sekä blogitekstejä. Relevanssikriteereinä käytettiin seuraavia:

1. Tekstin otsikossa kuvauksessa, abstraktissa tai hakusanoissa mainittiin tekoälyyn, koneoppimiseen tai koneoppimismenetelmiin, digitalisaatioon ja dataan tai nouseviin teknologioihin liittyviä sanoja
2. Tekstin otsikossa, kuvauksessa, abstraktissa tai hakusanoissa mainittiin perusoikeuksiin tai yhdenvertaisuus- tai tasa-arvovaikutuksiin liittyviä sanoja
3. Aineisto on selvitys- tai raporttimuotoinen
4. Aineisto on saatavilla englanniksi tai suomeksi

Kartoitus tuotti 22 raporttia, joista alustavan lukemisen jälkeen karsittiin yksi raportti relevanssikriteereiden perusteella. Aineiston analyysivaiheen aikana suoritettiin täydennyshakuja, jotka tuottivat 2 relevanssikriteerit täyttävää raporttia. Lopullinen otos pääkartoituksessa oli täten yhteensä n=23 raporttia. Tuotettu aineisto sisälsi pääosin kansallisten ja kansainvälisten toimijoiden, kuten kansainvälisten järjestöjen, tutkimusryhmien ja kansalaisjärjestöjen, raportteja. Suuri osa aineistosta oli tavoitteiltaan yhteneväisiä hankkeen ja selvityksen tavoitteiden kanssa sikäli, kun niissä pyrittiin tuottamaan tietoa tekoälyn yhdenvertaisuusvaikutuksista ja syrjintäriskeistä.

Tutkimuskysymykset ohjasivat aineiston analyysiä, joka suoritettiin hyödyntäen Atlas.ti-ohjelmistoa. Raporttien sisältöä luokiteltiin tutkimuskysymyksiin pohjautuen ”Riskit”-, ”Keinot”- ja ”Haasteet”-kategorioiden alle, joiden alle syntyi analyysin aikana alakategorioita. Esimerkiksi syrjintäriskien suhteen eriteltiin harjoitusdataan, algoritmeihin ja

mallin opettamiseen sekä tulosteiden tulkintaan liittyviä ongelmia eri koodien alle. Koodit täten heijastivat tutkimuskysymyksiä tuottaen kuitenkin hienojakoisemman luokittelun aineistossa nousseista teemoista. Raporteissa annettuja suosituksia koodattiin myös ”Suositukset”-kategorian alle. Koodien alle luokiteltua sisältöä ei järjestetty niiden esiintyvyyden mukaan, sillä kartoituksen ensisijaisena tarkoituksena oli saada kattava kuva syrjintäriskien, riskinhallintakeinojen sekä haasteiden kirjosta (esim. tiheyden sijaan). Löydökset-osiossa kuitenkin pyrimme kielellisesti tuomaan esille tiettyjen teemojen esiintyvyyttä sikäli, kun se nähdään tarpeelliseksi.

Analyysivaiheessa myös kerättiin ja tarkasteltiin raporteissa viitattua tutkimuskirjallisuutta (esim. suorittamalla pienempiä ”täsmähakuja” esiin nousseiden teemojen pohjalta). Tutkimuskirjallisuutta kartoittavia täsmähakuja suoritettiin muun muassa Google Scholar -hakukoneella sekä ACM Digital Library -verkkosivun hakutoiminnolla. Näin saatiin teoreettisesti vakaampi, tutkimuskirjallisuuden tukema pohja selvityksessä tuotetun tiedon tueksi.

Seuraavissa luvuissa kuvaamme yksityiskohtaisesti kartoituksen löydöksiä. Löydökset ovat jaoteltu viiden teeman alle. Jokaisen luvun alussa on yleistasoinen tiivistelmä osion sisällöstä.

2.3 Tekoälysovellusten syrjintäriskien profiili

Tämä osio tarjoaa aineiston pohjalta tuotetun riskiprofiilin algoritmisesta syrjinnästä, vastaten yleisellä tasolla kartoitusta ohjanneeseen tutkimuskysymykseen: *mitä syrjintäriskejä koneoppimista hyödyntävissä sovelluksissa on tunnistettu?* Raportin muut osiot tarkentavat tätä profiilia kuvaamalla riskiprofiilia teknisestä ja sosioteknisestä näkökulmasta.

- **Yleisesti aineistosta käy ilmi, että algoritmisten vinoumien riskejä voidaan löytää läpi sektoreiden ja toimialojen.** Löydös ei ole uusi, mutta korostaa tekoälysovellusten yhdenvertaisuusvaikutusten arvioinnin tarpeellisuutta läpi teknologian käyttökontekstien. Kontekstit, joissa teknologian käytöllä voi olla oikeudellisia vaikutuksia, luonnollisesti korostuvat.
- **Vinoumat voivat johtaa systemaattiseen epäsuotuisaan kohteluun erityisesti esimerkiksi etnisen alkuperän, sukupuolen ja vammaisuuden takia.** Erityisesti näihin ominaisuuksiin perustuvan välillisen syrjinnän riskit korostuvat aineistossa. Aineistossa tunnistetaan myös riski negatiivisille yhdenvertaisuusvaikutuksille, jotka voivat koskea kompleksisia intersektionaalisia ryhmiä, joita voidaan päätellä tai löytää koneoppimismenetelmin algoritmien hyödyntämästä opetusdatasta.

- **Aineistosta nousee esille, että rajanveto yhtäältä vinoumien ja niistä seuraavan systemaattisen erilaisen kohtelun ja toisaalta** prima facie **syrjinnän välillä on usein vaikeaa**. Tämän voidaan olettaa johtuvan osittain niin yhdenvertaisuutta koskevan lainsäädännön ja oikeuskäytännön kontekstuaalisuudesta ja tapauskohtaisuudesta[53] kuin todennettujen syrjintätapauksien puutteesta tekoälyn kontekstissa. Viime vuosien aikana ilmenneet tapaukset Euroopan sisällä[54] antavat kuitenkin joitakin viitteitä algoritmista syrjintää koskevan oikeuskäytännön kehittymisestä. Välittömän syrjinnän riskien arviointi on lähtökohtaisesti suoraviivaisempaa, mutta kulloisessakin tapauksessa mahdolliset oikeuttamisperusteet täytyy ottaa huomioon. Myös moniperusteisen ja intersektionaalisen syrjinnän riskejä tunnistetaan aineistossa.
- **Johdonmukaisena puutteena aineistossa tunnistetaan, että keskustelu tosiasiallisen yhdenvertaisuuden edistämisestä on marginaalista.** Löydös heijastelee oletettavasti algoritmisten vinoumien ympärillä käydyn keskustelun keskittymistä Yhdysvaltain kontekstiin, jossa syrjinnän kielto omaksuu keskeisen paikan yhdenvertaisuutta koskevissa diskursseissa.

2.3.1 Algoritmisen syrjinnän riskialueet

Raporteissa tunnistettiin algoritmisten vinoumien riskejä pitkälti läpi niin sektorien ja alojen (esim. sosiaali- ja terveyspalvelut, kasvatus ja koulutus, poliisitoiminta ja oikeudenkäyttö, finanssi- ja vakuutusalan toiminta, markkinointi, rekrytointi, asuntomarkkinat) kuin eri koneoppimismenetelmien ja sovellusten käyttötarkoitustenkin (esim. luokittelu, regressio, tiedonhaku ja -järjestäminen, konekääntäminen, hahmon- ja kuvantunnistus, suosittelu ja personalisointi). Raporteissa keskitytään ymmärrettävästi pääosin konteksteihin, joissa algoritmiavusteisella tai -vetoisella päätöksenteolla voi olla perusoikeudellisia vaikutuksia. Huomattakoon, että aineistossa mainitut riskialueet sisältävät kuitenkin myös yksityisen sektorin toimia (esim. finanssiala, rekrytointitoimet, mainonta), jotka suurelta osin kuuluvat yhdenvertaisuuslain soveltamisalaan (YVL 2 §).

Heijastaen osaltaan riskialueiden monipuolisuutta, tekoälyjärjestelmien avulla tuotetut syrjivät vaikutukset voivat erota merkittävästi luonteeltaan ja ne voivat myös kohdistua eri ryhmiin kuuluviin yksilöihin. Terveydenhuollossa käytetyt algoritmit voivat esimerkiksi evätä marginalisoituihin ryhmiin kuuluvilta henkilöiltä pääsyn heidän tarvitsemiinsa

53 Ks. esim. Wachter, Mittelstadt & Russell, 2020; Wachter, Mittelstadt & Russell, 2021.

54 Ks. esim. "Deliveroo Rating Algorithm Was Unfair To Riders, Italian Court Rules". (5.1.2021). Forbes. https://www.forbes.com/sites/jonathankeane/2021/01/05/italian-court-finds-deliveroo-rating-algorithm-was-unfair-to-riders/?sh=212a3c6122a1. Ks. myös vuoden 2020 SyRI-tapaus Haagin käräjäoikeudessa C/09/550982 / HA ZA 18-388

palveluihin, kun taas rekrytoinnissa käytettävät järjestelmät saattavat vahvistaa naisiin ja vammaisiin ihmisiin kohdistuvaa syrjintää. Yhdenvertaisuusvaikutukset voivat olla huomattavan erilaisia riippuen niin järjestelmästä kuin sen tosiasiallisesta käyttötarkoituksestakin (ks. alla). Päätöksentekoa tukevilla järjestelmillä voidaan esimerkiksi profiloida ja hallita ihmisten pääsyä etuuksiin tai palveluihin. Sisältöä moderoivat ja tuottavat algoritmit, kuten luonnollisen kielen prosessoinnin järjestelmät ja syvävärennös-teknologiat, taas ovat alttiimpia tuottamaan loukkaavaa sisältöä tai niitä voidaan käyttää eksplisiittisesti häirinnäksi luettaviin tarkoituksiin. Laajemmassa mittakaavassa voidaan myös tunnistaa yhdenvertaisuuden ja tasa-arvon näkökulmasta ongelmallisia kehityskulkuja. Yksittäisillä sektoreilla tekoälyjärjestelmien laajeneva käyttöönotto saattaa esimerkiksi muuttaa olemassa olevien yritysten liiketoimintamalleja tavoilla, jotka saattavat johtaa eriarvoistumiseen ja heikentää palveluiden saatavuutta tiettyjen ryhmien kohdalla[55].

Euroopan digitaalisten oikeuksien raportissa myös huomautetaan, että tekoälysovellusten vaikutukset voivat kohdistua elämänalueille, joita yhdenvertaisuutta sääntelevät lait eivät suoranaisesti kata, mutta joilla nämä vaikutukset voivat olla tässä suhteessa enemmän tai vähemmän merkittäviä[56]. Yhdenvertaisuuslain ja tasa-arvolain rajoitettu soveltamisala huomioiden voidaan huomauttaa, että tekoälysovellusten käytön vaikutukset voivat olla näilläkin elämänalueilla sosiaalieettisestä näkökulmasta merkityksellisiä.

2.3.2 Mitä on välitön algoritminen syrjintä?

Välitön syrjintä tekoälysovellusten kontekstissa on periaatteessa suhteellisen yksiselitteisesti määriteltävissä ja todennettavissa yksilötasolla[57]. On huomattava kuitenkin, että yhdenvertaisuuslain 11 § ja 12 § asettavat erilaiselle kohtelulle erinäisiä oikeuttamisperusteita ja tällöin tekoälysovelluksen tasolla tehtävä tarkastelu ei ole riittävää todentamaan välitöntä syrjintää joka tapauksessa. Oikeuttamisperusteiden puuttuessa tekoälysovelluksia voidaan pitää *prima facie* välittömästi syrjivinä kahdessa tapauksessa:

1. Malli sisältää kielletyn syrjintäperusteen (esim. muuttujana) tulosteiden tuottamisessa, mikä johtaa ihmisen erilaiseen kohteluun syrjintäperusteeseen perustuen. Tulosteen arvo perustuu tällöin informaatioon kielletystä

55 Center for Data Ethics and Innovation, 2020b, 115

56 Euroopan digitaaliset oikeudet 2021, sivut 54-55.

57 Wachter, Mittelstadt, & Russell, 2021, 16–17.

syrjintäperusteesta ja tulostetta käytetään päätöksenteossa, jolla on oikeudellisia vaikutuksia[58].

2. Tuloste itsessään sisältää eksplisiittistä tietoa kielletystä syrjintäperusteesta (esim. tulosteen arvo on kielletty syrjintäperuste) ja tulostetta käytetään päätöksenteossa, jolla on oikeudellisia vaikutuksia.

Molemmissa tapauksissa algoritmi käsittelee eksplisiittistä dataa kielletyistä syrjintäperusteista: ensimmäisessä tapauksessa tuottaakseen tulosteen, toisessa tapauksessa tuloste itse sisältää tietoa kielletystä syrjintäperusteesta. Mikäli päätöksentekijä kuitenkin hyödyntää tietoa päätöksenteon kohteen henkilökohtaisista ominaisuuksista (esim. sukupuolesta tai iästä) tosiasiallisessa päätöksenteossa ilman oikeudellista perustetta, voi algoritmiavusteinen päätöksenteko olla välittömästi myös silloin, kun ehdot (1-2) eivät täyty.

2.3.3 Mitä on välillinen algoritminen syrjintä?

Tekoälysovelluksia voidaan periaatteessa pitää *prima facie* välillisesti syrjivinä, kun niillä tuotetut vaikutukset asettavat systemaattisesti tietyn ryhmän jäsenet tosiasiallisesti epäsuotuisampaan asemaan, vaikka tulosteen arvo tai sen tarkkuus ei ole riippuvainen eksplisiittisesti informaatiosta koskien kiellettyä syrjintäperustetta. Yhdenvertaisuuslain 13 § mukaan välillisestä syrjinnästä ei ole kyse, kun ”säännöllä, perusteella tai käytännöllä on hyväksyttävä tavoite ja tavoitteen saavuttamiseksi käytetyt keinot ovat asianmukaisia ja tarpeellisia”. Välillisen syrjinnän tarkka ja yksiselitteinen määritteleminen ”algoritmien kielellä” onkin avoin kysymys, sillä ei ole selvää, missä suhteessa tulosteet eivät saa olla riippuvaisia kielletyistä syrjintäperusteista. On esimerkiksi epäselvää, kuinka vahva tilastollinen korrelaatio kielletyn syrjintäperusteen ja tulosteen välillä pitäisi olla (välittömän tai) välillisen syrjinnän todentamiseksi[59].

Joka tapauksessa välillisen syrjinnän todentaminen edellyttää mahdollisten oikeuttamisperusteiden (esim. tavoitteiden hyväksyttävyys, keinojen välttämättömyys ja oikeasuhtaisuus) tarkastelua. On täten epäselvää, joskin epätodennäköistä, että yksinomaan tekoälysovelluksen suorittamaa päätösfunktiota tai tulosteiden tilastollista jakaumaa tarkastelemalla saataisiin ratkaisevia todisteita mahdollisesta välillisestä syrjinnästä[60]. Tämä tekee

58 Ennakkotapauksena tällaisesta tilanteesta voidaan mainita Svea Ekonomi AB:n tapaus, jossa edellä mainitun katsottiin syyllistyneen välittömään moniperusteiseen syrjintään yhdenvertaisuuslain 8 § perusteella. Ks. yhdenvertaisuus ja tasa-arvolautakunnan tapausseloste 216/2017. https://www.yvtltk.fi/material/attachments/ytaltk/tapausselosteet/2SVk-NzOWF/YVTltk-tapausseloste-_21.3.2018-luotto-moniperusteinen_syrjinta-S._L.pdf.

59 Gerards & Xenidis, 2021, 64.

60 Ks. myös Wachter, Mittelstadt, & Russell, 2021.

syrjintäriskien hallinnasta ja ehkäisemistä kontekstuaalista ja tapauskohtaista harkintaa vaativaa. Siinä missä välittömän syrjinnän riskejä on kehittäjien ja käyttäjien näkökulmasta helpompi tunnistaa ja hallita (esim. poistamalla harjoitusdatasta mahdollinen kiellettyjä syrjintäperustoja koskeva harjoitus- tai syötedata tai mallin muuttujat), välillisen syrjinnän riskejä ei voida useimmissa tapauksissa ehkäistä samalla tavalla. Tämä johtuu siitä, että opetusprosessissa järjestelmä voi oppia korvannaismuuttujan (engl. *proxy*) kielletylle syrjintäperusteelle ja tällöin näennäisen neutraali muuttuja kantaa epäsuorasti mukanaan tietoa kielletystä syrjintäperusteesta. Tämä ei ole kuitenkaan aina selvää sovelluksen kehittäjille tai käyttäjille[61].

2.3.4 Moniperusteinen algoritminen syrjintä ja intersektionaalisuus

Lähes puolet tarkastelluista raporteista mainitsee moniperusteisen syrjinnän erillisenä riskinä. Moniperusteista syrjintää – eli syrjintää, joka perustuu useaan kiellettyyn syrjintäperusteeseen – voi periaatteessa esiintyä tekoälyn kontekstissa niin välittömässä kuin välillisessäkin muodossa. Ensimmäisessä tapauksessa algoritmi voi käsitellä eksplisiittistä dataa useasta kielletystä syrjintäperusteesta ja käyttää tietoa tulosteen muodostamisessa. Jälkimmäisessä tapauksessa algoritmi oppii datasta korvannaismuuttujan usealle syrjintäperusteelle. Kasvojentunnistusalgoritmit voivat esimerkiksi tuottaa systemaattisesti epätarkempia ennusteita tummaihoisille naisille kuin muille vertailukelpoisille ryhmille[62]. Osassa raporteista intersektionaalinen syrjintä erotetaan moniperusteisesta syrjinnästä nk. additiivisessa mielessä (vrt. erilainen kohtelu sukupuolen vuoksi ja iän vuoksi). Intersektionaalisuuden viitekehyksessä painotetaan, että vaikka syrjintä voi perustua moneen kiellettyyn syrjintäperusteeseen, on eri ryhmien kokema syrjintä laadullisesti erilaista (vrt. erilainen kohtelu iäkkäänä naisena).

2.3.5 Syrjinnän tahallisuus

Useassa raportissa huomautetaan, että tekoälyn kontekstissa niin välitön kuin välillinenkin syrjintä voi olla joko tahallista tai tahatonta. Tutkimuskirjallisuudessa on esitetty samankaltaisia huomioita[63]. Tahallisen syrjinnän riskeiksi mainitaan muun muassa se, että kehittäjät tai käyttäjät voivat rakentaa algoritmeista tarkoituksellisesti syrjiviä käyttämällä kiellettyä syrjintäperustetta tai sen kanssa korreloivaa korvannaismuuttujaa mallissa[64]. Olennai-

61 Euroopan unionin perusoikeusvirasto, 2020, 71.

62 Buolamwini & Gebru, 2018.

63 Ks. esim. Barocas & Selbst, 2016.

64 Gerards & Xenidis, 2021, 44.

sena riskinä tässä suhteessa pidettiin mahdollisuutta ”piilottaa” tai ”naamioida” syrjivät päätöksentekosäännöt käyttämällä useita korvannaismuuttujia kielletyille syrjintäperusteille, mikä tekee syrjivien vaikutusten tunnistamisesta hankalampaa[65] erityisesti tekoälyjärjestelmien läpinäkyvyyden puutteen vuoksi. Lisäksi yhdeksi riskiksi tunnistetaan, että tekoälyä käyttävät organisaatiot voivat käyttää syrjivää mallia päätöksentekoon mutta esittää auditointia suorittaville osapuolille (esim. tutkimusryhmille tai kolmansille osapuolille) käyttävänsä ns. ”reilua” eli yhdenvertaisuusvaikutuksiltaan neutraalia tai positiivista mallia[66].

2.3.6 Tosiasiallisen yhdenvertaisuuden edistäminen tekoälyn kontekstissa

Kartoituksen perusteella huomionarvoinen löydös on, että keskustelu algoritmisista vinoumista keskittyy lähes yksinomaan syrjinnän ehkäisemiseen heijastaen yhdenvertaisuuslainsäädännön asettamia syrjinnän kieltoja[67]. Syrjinnän ehkäisemisen lisäksi yhdenvertaisuuslaki asettaa kuitenkin viranomaisille, koulutuksentarjoajille ja työnantajille velvollisuuden edistää tosiasiallisen yhdenvertaisuuden toteutumista. Yksi erityinen velvollisuus tässä suhteessa on velvollisuus tehdä kohtuullisia mukautuksia vammaisille tai rajoitteisille henkilöille. Esimerkiksi työnantajat, jotka eivät ennakoi ja minimoi negatiivisia vaikutuksia, joita tekoälyn käytöllä on vammaisiin hakijoihin tai työntekijöihin, voivat täten laiminlyödä velvollisuuttaan tarjota kohtuullisia mukautuksia[68].

Tosiasiallisen yhdenvertaisuuden edistämisen velvollisuudet kattavat myös vaatimuksen edistää yhdenvertaisuutta laajemmin toimin kuin esimerkiksi kohtuullisin mukautuksin. Konkreettisiin toimiin tässä suhteessa lukeutuvat muun muassa yhdenvertaisuusvaikutustenarviointi (YVL 5-7 §), suunnitelmat tosiasiallisen yhdenvertaisuuden edistämiseksi

65 Ks. Barocas & Selbst, 2016.

66 Algo:aware, 2018, 27.

67 Syrjinnän kieltoihin keskittyvä diskurssi voi osaltaan johtua siitä, että niin tutkimuskirjallisuus kuin yleinen keskustelukin ovat keskittyneet Yhdysvaltain kontekstiin, jonka yhdenvertaisuutta koskeva lainsäädäntö eroaa eurooppalaisesta lähestymistavasta. Yhdysvaltojen kontekstissa keskeinen periaatteellinen lähestymistapa kattaa syrjinnän kiellot (engl. *anti-classification*), kun taas Suomessa ja Euroopassa yleisemmin myös velvollisuudet edistää yhdenvertaisuutta ovat keskeisessä roolissa yhdenvertaisuuslainsäädännössä.

68 Institute for the Future of Work, 2020.

(YVL 5-7 §) ja näiden läpinäkyvyys (YVL 6-7 §) sekä sidosryhmien kuuleminen (YVL 6-7 §). Velvollisuuksien laiminlyömistä on jo ilmennyt tekoälyjärjestelmien kontekstissa maailmalla, kuten esimerkiksi Etelä-Walesin poliisin tapauksessa[69].

2.3.7 Kielletyt syrjintäperusteet ja systemaattinen erilainen kohtelu

Keskustelu syrjinnästä koskee usein yleisempiä syrjinnän muotoja, kuten sukupuoleen tai etniseen taustaan perustuvaa syrjintää. Osassa raportteja ilmaistaan huoli, että tekoälysovellusten epäsuotuisat vaikutukset voivat koskea ryhmiä, jotka eivät kuulu lainsäädännöissä nimenomaisesti mainittuihin kiellettyihin syrjintäperusteisiin. Tekoälyjärjestelmät voivat päätellä, löytää tai luoda uusia perusteita ja ryhmiä, jotka eivät suoranaisesti korreloi kiellettyjen syrjintäperusteiden kanssa, erotellakseen ihmisiä. Tämä nostaa esiin kysymyksen siitä, ”luoko algoritmien kasvava käyttö uusia syrjinnän muotoja”[70]. Riskinä nähdään, että tällaisten perusteiden avulla tapahtuva erottelu voi olla yhtäältä moraalisessa mielessä epäreilua ja vahvistaa yhteiskunnallista eriarvoisuutta sekä toisaalta pysyä yhdenvertaisuuslain tavoittamattomissa. Esimerkiksi rekrytointisovelluksia tarkastelevassa raportissa huomautetaan, että ”koneoppimisjärjestelmät voivat uusintaa epätasa-arvoa, jota syrjinnän kapea määritelmä ei tavoita, tunnistaen ryhmiä tavanomaisten piirteiden perusteella, joita ei lueta kielletyiksi syrjintäperusteiksi, tehden päätöksiä, jotka voivat sulkea heidät työmarkkinoilta suuressa mittakaavassa.”[71]

Suomen yhdenvertaisuuslain kontekstissa kyseiset tapaukset eivät näyttäisi lipuvan täysin yhdenvertaisuuslain tavoittamattomiin[72]. Yhdenvertaisuuslain soveltamisalaan kuuluvat useat alueet, jossa yllä kuvatun kaltaista systemaattista erilaista kohtelua voi tapahtua (esim. rekrytointi) ja niin välittömän kuin välillisenkin syrjinnän kielto pätee myös näissä konteksteissa. Keskeisempänä syynä on kuitenkin yhdenvertaisuuslaissa asetetun kiellettyjen syrjintäperusteiden listan avoimuus. YVL 8 §:ssä mainitut kielletyt syrjintäperusteet kattavat sosiaalisten ja demograafisten seikkojen lisäksi ”muut henkilöön liittyvät seikat”. Muut henkilöön liittyvät syyt kattavat paitsi henkilön sisäsyntyiset ominaisuudet myös oikeudellisen aseman ja sosiaalisessa kanssakäymisessä merkitykselliset ominaisuudet (i) sikäli, kun nämä liittyvät henkilöön itseensä eivätkä esimerkiksi hänen toimintaansa ja (ii) sikäli, kun nämä ominaisuudet ovat rinnastettavissa yhdenvertaisuuslaissa

69 Hovioikeus, R (Bridges) v. CC South Wales, [2020] EWCA Civ 1058, 11. Elokuuta 2020. https://www.judiciary.uk/judgments/r-bridges-v-cc-south-wales/.

70 Borgesius, 2018, 228.

71 Institute for the Future of Work, 2020, 18.

72 Gerards & Xenidis, 2021.

nimenomaisesti mainittuihin syrjintäperusteisiin[73]. Erot ihmisten tosiasiallisissa olosuhteissa, toiminnassa tai menettelytavoissa eivät yleensä muodosta tällaista henkilöön liittyvää syytä[74], ja merkitystä on sillä, onko kyseessä ominaisuus tai asema, jolla ihmiset tai ryhmät ovat tosiasiallisesti eroteltavissa toisistaan[75].

Euroopan ihmisoikeustuomioistuimen käytännöstä käy ilmi, että sen arvioiminen, onko tietty seikka siinä määrin henkilön henkilökohtainen tai muulla tavoin tunnistettava ominaisuus tai asema, että kyse on muusta henkilöön liittyvästä syytä, tulee tehdä tapauskohtaisesti ottaen huomioon tapauksen olosuhteet kokonaisuudessaan[76]. Tutkimuskirjallisuudessa on myös painotettu, että tekoälysovellusten auditoimisessa vertailuluokkien valintaa syrjintäriskien arvioimisessa hankaloittaa se, että "vertailuluokat ovat tapauskohtaisia ja ne määritellään vasten kiistanalaista ja riitautettua käytäntöä tai sääntöä" ja tapauksen faktuaalista perustaa[77].

2.4 Tekoälysovellusten syrjintäriskien teknologiset syyt

Tässä osiossa pureudutaan syvemmälle algoritmisen syrjinnän riskeihin tarkastelemalla vinoumien syntymekanismeja teknisellä tasolla – ts. datan, algoritmien ja mallien sekä näiden luomisprosessien tasolla.

- **Kartoituksen pohjalta käy selväksi, että (algoritmiset) vinoumat ovat ennemmin sääntö kuin poikkeus. Kaikki vinoumat eivät toisaalta välttämättä johda syrjintään oikeudellisessa merkityksessä tai ole muutoin moraalisesti ongelmallisia.** Tekoälysovelluksen käyttö ei välttämättä tuota oikeudellisia vaikutuksia tai kyseiset vaikutukset voivat olla oikeutusperiaatteiden valossa oikeutettuja, välttämättömiä ja oikeasuhtaisia. On

73 Judgment on merits and just satisfaction. *Hode and Abdi v. The United Kindgdom*, no. 22341/09, ECHR 2012. Ks. myös "Syrjintäperusteet". *Yhdenvertaisuusvaltuutettu*. syrjinta.fi/syrjintaperusteet.

74 Perustuslakivaliokunta on esimerkiksi katsonut, ettei käräjäoikeudessa käsiteltävien nk. summaaristen velkomusasioiden oikeudenkäyntimaksun suuruuden porrastamisessa asian vireilletulotavan mukaan ollut kysymys eri asemaan asettamisesta henkilöön liittyvän syyn perusteella (PeVL 35/2010 vp, s. 2. https://www.eduskunta.fi/FI/vaski/Lausunto/Documents/pevl_35+2010.pdf.

75 Ks. esim. CLIFT v. THE UNITED KINGDOM 7205/07, kohta 55.

76 Euroopan ihmisoikeustuomioistuin on oikeuskäytännössään pitänyt esimerkiksi entisen KGB-agentin ammattia (SIDABRAS AND DŽIAUTAS v. LITHUANIA 55480/00 59330/00) sekä kalastajan ammattia (ALATULKKILA AND OTHERS v. FINLAND 33538/96 muuna henkilöön liittyvänä syynä.

77 Wachter, Mittelstadt & Russell, 2021, 22.

huomionarvoista, että syrjiviä vaikutuksia voidaan myös ehkäistä rakentamalla tarkoituksellisesti dataan tai algoritmeihin vinoumia.

- **Vinoumat ja erityistapauksissa mahdollinen syrjintä voivat johtua niin virheellisestä kuin huolellisestakin suunnittelusta.** Ensimmäisissä tapauksissa opetusdata saattaa olla epäedustavaa tai mallin muuttujat operationalisoitu tavalla, joka johtaa syrjiviin vaikutukseen. Jälkimmäisessä tapauksessa populaation tarkkakin mallinnus voi johtaa päätöksiin tai suosituksiin, jotka heijastavat olemassa olevia käytäntöjä, rakenteita tai esimerkiksi mallinnettavan populaation preferenssejä. Vinoumia voi syntyä eri tavoin riippuen esimerkiksi sovelluksen käyttökontekstista ja -tarkoituksesta, datan keruun ja esikäsittelyn tavoista, mallintamisesta ja siinä käytetyistä menetelmistä, ihmisen ja koneen vuorovaikutuksesta sekä sovelluksen käytöstä.
- **Vinoumat ovat monivaiheisten ja erilaisia aktiviteetteja sekä toimijoita yhdistävien arvoketjujen tuotosta.** Tekoälyn tuotantoketjun monivaiheisuus, mahdollinen iteratiivisuus, syklisyys sekä taipumus yhdistää useita toimijoita tuottavat haasteita vinoumien tunnistamiselle ja asianmukaiselle ehkäisemiselle.

2.4.1 Vinoumien luokittelusta

Vinoumia voidaan luokitella eri tavoin ja tarkastellussa aineistossa ilmenee useita, joskin useimmiten yhteneviä tapoja luokitella vinoumia. Yksi yleinen tapa luokitella vinoumia on erotella niitä sen perusteella, rakentuvatko ne järjestelmään esimerkiksi opetusdatan vaiko algoritmin myötä. Tutkimuskirjallisuudessa vinoumia on lisäksi luokiteltu esimerkiksi tilastollisiin, oikeudellisiin ja moraalisiin vinoumiin riippuen siitä, viitataanko tarkastellussa tapauksessa mallin poikkeamiseen tilastollisesta (vrt. empiirinen), juridisesta vaiko eettisestä standardista[78]. Taulukossa 5 puolestaan on listattu erilaisia esimerkkejä vinoumista niiden syntymiseen johtavien mekanismien perusteella. Kyseinen listaus on suuntaa antava eikä suinkaan kattava kuvaus kaikista mahdollisista vinoumista.

78 Ks. Danks & London, 2017; Fazelpour & Danks, 2021.

Taulukko 5. Esimerkkejä erilaisista vinoumista ja niiden syistä (Mehrabi ym., 2021).

Datasta algoritmiin *Vinoumat datassa, jotka voivat johtaa vinoumiin tulosteiden tasolla.*
1. *Mittausvinouma.* Vinouma, joka juontaa siitä, miten ominaisuuksia tai ilmiöitä raportoidaan, valikoidaan, hyödynnetään, operationalisoidaan tai mitataan.
2. *Puuttuvan muuttujan vinouma.* Vinouma, joka seuraa merkityksellisen muuttujan jättämisestä pois mallista.
3. *Edustavuusvinouma.* Vinouma, joka juontaa näytteenotosta populaatiosta datankeruun aikana.
4. *Yhdistämis- tai aggregaatiovinouma.* Vinouma, joka juontaa siitä, että populaatiota koskevista havainnoista tehdään päätelmiä yksilöistä. a) Simpsonin paradoksi, jossa kahden muuttujan välinen korrelaatio muuttuu päinvastaiseksi, kun huomioon otetaan kolmas muuttuja. b) Muokattavan alueyksikön ongelma, joka seuraa esimerkiksi geospatiaalisessa analyysissä, kun data sisältää erilaisia spatiaalisia tarkastelutasoja.
5. *Otantavinouma.* Vinouma, joka juontaa epäedustavasta tai ei-sattumanvaraisesta näytteenotosta (vrt. edustavuusvinouma).
6. *Pitkittäisdatan vinouma.* Vinouma, joka voi seurata esimerkiksi heterogeenisten kohorttien analysoinnista poikkileikkausanalyysin menetelmin pitkittäisanalyysin sijaan.
7. *Linkitysvinouma.* Vinouma, joka syntyy, kun toimijoiden yhteyksiä, toimintaa ja vuorovaikutusta tarkastelemalla tuotettu tieto toimijaverkoston (esim. sosiaalisen median alustan) ominaisuuksista ei edusta toimijoiden todellista käyttäytymistä tai antaa siitä väärän kuvan.
Algoritmista tulosteisiin / käyttäjään *Algoritmin käytön myötä syntyvät vinoumat, jotka vaikuttavat käyttäjiin*
8. *Algoritmin vinouma.* Vinouma, joka syntyy algoritmin käyttämisen tuloksena syötedatan mahdollisista vinoumista riippumatta.
9. *Käyttäjäinteraktiovinouma.* Vinouma, jossa ympäristön ja toimijan vuorovaikutuksen syntyviä havaintoja tulkitaan yksinomaan käyttäjän käyttäytymistä edustavana. a) *Näkyvyysvinoumat* syntyvät, kun esimerkiksi verkkopalvelun käyttäjien klikkauksista päätellään heidän suosivan tiettyä sisältöä, vaikka sisällön näkyvyys vaikuttaa siihen, mitä käyttäjät klikkaavat. b) Järjestysvinoumat syntyvät, kun sisällön esittämisen järjestys vaikuttaa käyttäjien klikkauksiin.
10. *Populaarisuusvinouma.* Vinouma, joka perustuu siihen, että suosittu sisältö (esim. verkkokaupoissa) saa enemmän näkyvyyttä.

Algoritmista tulosteisiin / käyttäjään

Algoritmin käytön myötä syntyvät vinoumat, jotka vaikuttavat käyttäjiin

11. *Ilmentyvä (emergentti) vinouma*. Vinouma, joka syntyy käyttäjien tai populaation ja käytetyn järjestelmän vuorovaikutuksessa, ja ilmenee esimerkiksi muutoksina populaation käyttäytymisessä tai sosiaalisissa ja kulttuurisissa arvoissa.
12. *Arviointivinouma*. Vinouma, joka syntyy koneoppimisalgoritmin avulla tuotettua mallia arvioitaessa esimerkiksi soveltumattomin mittarein tai vertailuarvoin.

Käyttäjästä dataan

Käyttäjistä tai mallinnettavasta populaatiosta dataan siirtyvät vinoumat

13. *Historiallinen vinouma*. Vinouma, joka seuraa historiallisten ja olemassa olevien yhteiskunnallisten, rakenteellisten ja/tai sosiaalisten erojen mallintamisesta (jopa silloin, kun mallintamisprosessi on täydellisen tarkka).
14. *Populaation vinouma*. Vinouma, joka seuraa siitä, kun mallinnettavan kohdepopulaation (esim. sovelluksen käyttäjät) tilastolliset, demografiset tai muut piirteet eroavat populaatiosta yleisesti.
15. *Itsevalikoitumisvinouma*. Yhdenlainen edustavuus- tai otantavinouma, joka syntyy, kun otos perustuu yksilöiden itsevalikoitumiseen (esim. Facebook-ryhmään valikoituu homogeeninen joukko yksilöitä).
16. *Sosiaalinen vinouma*. Vinouma, joka syntyy, kun yksilön arviointikykyyn tai toimintaan vaikuttavat havainnot muiden toimijoiden toiminnasta (esim. verkkokaupassa yksilön antamaan arvioon tuotearvioon voivat vaikuttaa muiden antamat arviot).
17. *Käyttäytymisvinouma*. Vinouma, joka seuraa toimijoiden eroavasta käyttäytymisestä eri alustoilla, eri konteksteissa tai eri datajoukoissa.
18. *Ajallinen vinouma*. Vinouma, joka seuraa muutoksista toimijoiden käyttäytymisen ajallisessa tarkastelussa.
19. *Sisällöntuottamisvinouma*. Vinouma, joka juontaa rakenteellisista, leksikaalisista, semanttisista tai syntaktisista eroista toimijoiden tavoissa tuottaa sisältöä (esim. eri ikäryhmien eroavaisuudet kielenkäytössä).

2.4.2 Vinoumien paikantamisesta tekoälyjärjestelmän arvoketjussa

Keskeinen huomio on vinoumien syntymisestä on, että ”[v]inoumat eivät nouse vain datajoukoista tai yksittäisistä aktiviteeteista koneoppimisen tuotantoketjussa (esim. datan käsittely, mallin opettaminen) vaan [näiden] aktiviteettien yhdistelmistä”[79]. Esimerkiksi datassa esiintyviä vinoumia voidaan tapauksesta riippuen joko uusintaa, vahventaa tai korjata riippuen käytettävästä mallinnusmenetelmästä ja oppimisalgoritmista. Täten vinoumien ymmärtäminen tietyssä arvoketjun vaiheessa, kuten datankeruussa, syntyvinä yksittäisinä tapahtumina on jokseenkin harhaanjohtava, joskin eri vinoumien lähteiden ja juurisyiden erottaminen on käytännöllisesti hyödyllistä riskinarvioinnin ja -hallinnan näkökulmasta.

Tuotantoketjut yhdistävät myös usein monia toimijoita, mikä vaikeuttaa vinoumien tunnistamista. Esimerkiksi vinoumia sisältävät harjoitusdatajoukot saattavat levitä nopeasti eri toimijoille ja eri käyttökonteksteihin[80]. Syrjinnän todentamisen näkökulmasta merkityksellisiä ovat kuitenkin ensisijaisesti päätökset, joilla on oikeudellisia vaikutuksia, eivätkä tekoälysovellusten tulosteet sinänsä. Tutkimuskirjallisuudessa onkin osittain näistä syistä kyseenalaistettu ajatus, että syrjintää – etenkään välillistä syrjintää – voitaisiin tunnistaa tarkastelemalla yksinomaan tekoälysovellusta tai sen yksittäisiä komponentteja, kuten koodia tai harjoitusdataa[81].

Syrjinnän riskien tunnistaminen on kuitenkin toivottavaa ja välttämätöntä yhdenvertaisuusperiaatteen toteutumisen näkökulmasta. Tekoälysovellusten kehitys- ja tuotantoketjua tarkastelemalla voidaankin saada hienojakoisempi kuva vinoumien lähteistä. Tässä osiossa kuvatut teknologiset syyt ja riskialueet osoittautuvatkin merkityksellisiksi mahdollisten syrjintäriskien tunnistamisen ja ehkäisemisen näkökulmasta. Syrjintäriskien tunnistaminen ja syrjivien vaikutusten ehkäiseminen vaatii kuitenkin keskittymistä niin yksittäisiin teknologisiin ja teknologian ulkopuolisiin syihin kuin tekoälysovellukseen kokonaisuudessaan ja käyttökontekstiinsa sidottuna.

2.4.3 Vinoumien syntymekanismit

Yksittäisten vinoumien lähteitä voidaan eritellä perustuen arvoketjun eri vaiheisiin. Tässä osiossa keskitymme erityisesti tekoälysovelluksen komponentteihin ja niihin liittyvän kehitystyön keskeisiin vaiheisiin.

79 Euroopan digitaaliset oikeudet 2021, sivu 67.

80 Achiume 2020, 11. Ks. myös. Peng ym. 2021.

81 Wachter, Mittelstadt & Russell, 2020.

2.4.3.1 Tehtävät, algoritmit ja mallit

Sovelluksen tehtävän ja tavoitefunktion (engl. *objective function*) määrittäminen ovat keskeisessä osassa algoritmien ja tekoälyjärjestelmien suunnitteluprosessia, sillä tässä vaiheessa myös määritetään, mitä algoritmin on tarkoitus optimoida. Algoritmit ja dataan sovitettavat mallit voivat tuottaa tulosteiden tasolla ilmeneviä vinoumia, jotka eivät palaudu hyödynnetyn datan mahdollisiin vinoumiin. Käsittelemme näitä alla.

Tekoälyjärjestelmät toteuttavat tiettyä tavoitefunktiota eli laskennallista tehtävää. Tämä tehtävä on tekoälysovellusten tapauksessa useimmiten optimointitehtävä. Kysymys siitä, mihin tarkoitukseen sovellusta yhtäältä optimoidaan – sekä toisaalta, mihin sitä tosiasiallisesti käytetään (ks. alla) – ja onko kyseinen käyttötarkoitus yhteensopiva yksilöiden perusoikeuksien, mukaan lukien syrjinnän kiellon, sekä eettisten standardien kanssa on keskeinen[82]. Joissakin tarkastelluista raporteista huomautetaan, että algoritmin tavoitefunktio tai sovelluksen toteuttama tehtävä voi olla epäoikeudenmukainen tai moraalisesti arveluttava vaikka ihmisiä kohdeltaisiinkin yhdenvertaisesti sen käytössä[83]. Algoritmit voidaan esimerkiksi suunnitella tarkoituksellisesti syrjiviksi (vrt. rekrytointiin käytettävä algoritmi, joka suunnitellaan suodattamaan raskaana olevat naiset pois kandidaattien joukosta).

Algoritmeissa saatetaan hyödyntää tilastollisia estimaattoreita, joilla pyritään korjaamaan datassa esiintyviä vinoumia, lisäämään reliabiliteettia tai robustiutta tai välttämään mallin ylisovittamista kohinaa sisältävään dataan (engl. *noisy data*) tai poikkeaviin havaintoihin (engl. *outliers*). Koneoppimisalgoritmit eivät tässä mielessä ole neutraaleja vaan välttämättä vinoutuneita jossain mielessä, mutta usein nimenomaan algoritmin luotettavuutta ja tarkkuutta palvelevista syistä.[84] Joissakin tapauksissa algoritmit itsessään saattavat tuottaa epätoivottavia vinoumia (laskennalliseen tai päätöksentekoprosessiin), jotka eivät palaudu vinoumiin datassa. Tekoälysovelluksissa hyödynnetyt algoritmit tai mallit saattavat olla eri tavoin soveltumattomia käyttökontekstiin. Ennakoivassa poliisitoiminnassa hyödynnetään esimerkiksi maanjäristysten tunnistamiseen suunniteltuja geospatiaalisia ennustemalleja. Maanjäristysten monitoroiminen on kuitenkin järjestelmällisempää ja vähemmän altista virheille kuin rikostilastointi, ja poliisitoiminnassa mallien on todettu vinoutuneen siten, että korkean poliisiläsnäolon alueilla rikosten esiintyvyys (ja algoritmia käytettäessä ilmaantuvuus) tulee yliarvioiduksi[85]. Asianmukaisen algoritmin ja mallin valinta on täten merkityksellistä niiden toimivuuden mutta myös niiden oikeudellisten vaikutusten näkökulmasta.

82 Institut Montaigne, 2020; The Greenlining Institute, 2020; Euroopan digitaaliset oikeudet, 2021. Ks. myös Passi & Selbst, 2019.

83 Euroopan digitaaliset oikeudet, 2021, 93.

84 Danks & London, 2017, 3.

85 World Economic Forum, 2018, 10.

Koneoppimisalgoritmille valitaan optimointikriteerit, joiden perusteella se maksimoi tiettyä arvoa (esim. osumatarkkuus riskinarviointimalleissa tai relevanssi suosittelujärjestelmissä) ja/tai minimoi toista (esim. virhetasot regressiomalleissa). Optimointikriteerien valinta vaikuttaa siihen, mitä algoritmi oppii datasta ja täten, mitä mahdollisia vinoumia algoritmiin voi rakentua[86]. Yleinen riski syrjiville vaikutuksille onkin, että algoritmi opetetaan maksimoimaan osumatarkkuutta, jolloin se voi oppia toistamaan tai jopa vahvistamaan datassa esiintyviä vinoumia. Huomattakoon, että algoritmi voidaan opettaa myös maksimoimaan jonkinlaista ”reiluutta” tai tasa-arvoisuutta samanaikaisesti esimerkiksi osumatarkkuuden kanssa. Erilaisten reiluuden (ks. alla) tai diversiteetin määritelmien käyttäminen optimointikriteereinä voidaankin ymmärtää tässä mielessä mahdollisena algoritmisten vinoumien lähteenä mallinnusvaiheessa.

2.4.3.2 ”Mustan laatikon ongelma” ja läpinäkyvyyden puute

Huoli tekoälysovelluksien läpinäkymättömyydestä nousee keskeisenä teemana useimmissa raporteissa. Tekoälyjärjestelmien hyödyntämät suuret datamäärät, tilastollisten mallien moniulotteisuus sekä tiedonkäsittelyn kerrostuneisuus ja kompleksisuus osoittautuvat usein esteeksi (i) niiden mallin toiminnan ymmärtämiselle ja/tai (ii) järjestelmän tulosteiden jäljittämiselle ja ymmärtämiselle. Selitettävyyden ongelmaan ja läpinäkyvyyden puutteeseen viitataan usein ”mustan laatikon ongelmana”. Mustan laatikon ongelma on paikoin ymmärretty käsitteellisessä mielessä syrjintäriskeistä erillisenä huolenaiheena, mutta se voi haitata syrjintäriskien hallintaa ja muodostuu merkitykselliseksi objektiivisen oikeutuksen antamisen ja päätöksenteon kohteen oikeuksiin pääsyn näkökulmista. Läpinäkyvyyden puute on kuitenkin käytetyn algoritmin tai mallin ominaisuuksista johtuva riski syrjiville vaikutuksille, minkä takia läpinäkyvyyden puute luetaan algoritmeihin ja malleihin liittyvään riskikategoriaan.

Läpinäkyvyyden puute voi itsessään johtaa päätelmään *prima facie* syrjinnästä, kuten Equinetin raportissa huomautetaan vedoten useisiin oikeudellisiin ennakkotapauksiin[87]. Riskinä on, että algoritmien läpinäkyvyyden puutteesta johtuen yritykset eivät esimerkiksi voi välttämättä tarjota objektiivista oikeutusta päätöksilleen. Tässä suhteessa erityisesti läpinäkyvyys nk. *selitettävyytenä* eli algoritmin tai mallin ymmärrettävyytenä ja tulosteiden jäljitettävyytenä on olennaista. Useammassa raportissa läpinäkyvyyden puute todettiin ongelmalliseksi myös päätöksenteon subjektien näkökulmasta sikäli, kun tiedon puute päätöksenteon perusteita koskien voi estää heitä käyttämästä oikeuksiaan ja riitauttamasta heitä koskevia päätöksiä. Esimerkiksi rekrytointia ja tekoälyä käsittelevässä

86 The Greenlining Institute, 2020, 7; Euroopan digitaaliset oikeudet, 2021.

87 Allen QC & Masters, 2020, 47.

raportissa huomautetaan, että "henkilön, jonka kohdalla tekoälyjärjestelmää on käytetty, voi olla hankala tietää tai osoittaa, tällä käytöllä on ollut heihin ja kuinka tämä on yhteydessä kiellettyyn syrjintäperusteeseen"[88].

2.4.3.3 Opetus-, syöte- ja arviointidata

Käytetyn datan, erityisesti harjoitusdatan, vinoumat nousivat jokaisessa raportissa keskeisimpinä syinä syrjintäriskeille ja mahdollisille syrjiville vaikutuksille. Datan vinoumat voivat johtaa siihen, että algoritmi oppii syrjiviä ennustemuuttujia tai jättää oppimatta piirteitä, jotka ennustavat merkityksellisiä eroja yksilöiden tai ryhmien välillä. Voidaankin tunnistaa kolme yleistä dynamiikkaa, jotka saattavat johtaa negatiivisiin yhdenvertaisuusvaikutuksiin:

I. Opetusdata sisältää kiellettyjä syrjintäperusteita, jolloin algoritmi voi oppia kohtelemaan yksilöitä eri tavalla datasta opitun syrjintäperusteen perusteella.
II. Opetusdata ei koosta edustavaa otosta, jolloin algoritmi ei opi piirteitä, jotka ennustavat osapopulaatioiden merkityksellisiä eroja.
III. Opetusdata heijastaa rakenteellisia eroja populaatiossa, jolloin algoritmi oppii toistamaan kyseisiä eroja.
IV. Opetusdata sisältää sisältöä, joka on loukkaavaa, halventavaa tai syrjiviä stereotypioita heijastavaa.

Kaksi jälkimmäistä dynamiikkaa voivat esiintyä paikoin samanaikaisesti. Painotamme, että opetusdataan liittyvät mahdolliset ongelmat ovat usein merkityksellisiä myös syöte- ja arviointidatan sekä käyttöliittymäsuunnittelun näkökulmasta. Käytännössä keskeisin opetusdataan liittyvä riski on, että opetusdatajoukko ei koosta edustavaa otosta (ks. alla).

Opetusdata sisältää kiellettyjä syrjintäperusteita. Opetusdatan sisältäessä kiellettyjä syrjintäperusteita syntyy riski välittömälle syrjinnälle (vrt. osio 2.3.2). Mikäli kielletyn syrjintäperusteen arvot (esim. sukupuoli "mies", "nainen") korreloivat kohdemuuttujan arvon kanssa, algoritmi saattaa oppia käyttämään syrjintäperustetta päätöksentekokriteerinä. Tekoälyjärjestelmien kehittäjät saattavatkin poistaa kielletyt syrjintäperusteet opetusdatasta, tai jättää keräämättä dataa niistä, välttääkseen negatiivisia yhdenvertaisuusvaikutuksia tai esimerkiksi tietosuojaa koskevista syistä. Tämä strategia – nk. "sokeus

88 Institute of the Future of Work, 2020, 19.

kielletyille syrjintäperusteille" – ei ole suuressa osassa tapauksia riittävä välttämään välillisiä syrjiviä vaikutuksia, sillä algoritmi saattaa oppia datasta korvannaismuuttujia kielletyille syrjintäperusteille.[89]

Attribuuttien aliedustus opetusdatassa. Koneoppimisalgoritmit antavat usein enemmän painoa ennustaville tekijöille, jotka esiintyvät opetusdatan numeerisen enemmistön kohdalla. Toisin sanoen, mikäli algoritmille ei näytetä tarpeeksi harjoitusesimerkkiä jokaisesta relevantista luokasta tai osapopulaatiosta, se ei välttämättä opi tunnistamaan piirteitä, jotka ennustavat kohdemuuttujan arvoa kyseisen luokan tai osapopulaation instanssien kohdalla. Esimerkiksi kasvojentunnistusalgoritmi saattaa tunnistaa heikommin tummaihoisten ihmisten kasvoja, jos kasvoja, joissa on tumma pigmentti, ei löydy tarpeeksi opetusdatasta. Tätä kutsutaan otanta- tai edustavuusvinoumaksi[90]. Edustavuuden puutteille on lukuisia käytännöllisiä syitä tekoälysovellusten kehittämisprosessien näkökulmasta[91].

- Käytetty data voi tulla lähteistä (vrt. palvelut tai palveluekosysteemit), joita ryhmien (i) edustajat eivät käytä yhtä paljon (esim. sosiaalisen median alustat), (ii) joihin heillä ei ole pääsyä tai (iii) joissa kyseisen ryhmän edustajat ovat yliedustettuna (esim. sosiaali- ja terveyspalveluiden data).
- Edustavan datan hankkiminen voi olla järjestelmän kehittäjien tai operoijien perspektiivistä hankalaa, kallista ja sitä voivat rajoittaa yksityisyydensuojaan liittyvät säännökset ja lait[92]. Datasubjektit saattavat esimerkiksi vetää GDPR:n asettamien oikeuksien mukaisesti suostumuksensa pois, jolloin heitä koskeva data täytyy poistaa rekisteristä[93].
- Datan keruuseen käytetyt instrumentit ja sensorit sekä esimerkiksi käyttöliittymäsuunnittelu saattavat muodostua esteeksi edustavan datan hankkimiselle. Esimerkiksi vammaiset henkilöt eivät välttämättä voi antaa vaadittavia biometrisiä tunnisteita, kuten sormenjälkiä tai silmän iiriksen kuvia[94]. Saattaa olla myös, että tekoäly- tai yleisemmin digitaaliset järjestelmät eivät tue kaikkia kieliä, joita sovellusta käyttävässä populaatiossa esiintyy[95].

89 Institut Montaigne, 2020; Centre for Data Ethics and Innovation, 2020a; Euroopan digitaaliset oikeudet, 2021.

90 Mehrabi ym., 2019.

91 Ks. esim. Orwat, 2020, 56; Mehrabi ym., 2019.

92 Ks. esim. Holstein ym., 2019.

93 Ks. esim. Centre for Data Ethics and Innovation, 2020b, 81.

94 Achiume, 2020, 12.

95 Achiume, 2020, 9.

Yllä mainitut haasteet hankaloittavat myös syrjintäriskien arvioimista ja ehkäisemistä, sillä algoritmien testaamiseen ja auditointiin tarvitaan useimmissa tapauksissa tietoa kiellettyihin syrjintäperusteisiin liittyen (ks. alla).

Nimikkeiden valikoitumisvinoumat. Opetusdatan ali- tai yliedustuksen ongelmien lisäksi datan nimikkeisiin (engl. *labels*) liittyvät puutteet saattavat aiheuttaa vinoumia. Menneisyyttä koskevasta opetusdatasta puuttuvat esimerkiksi useimmiten kontrafaktuaaliset lopputulemat, mikä johtaa niin kutsuttuun *nimikkeiden valikoitumisvinoumaan* (engl. *label selection bias*). Tällä tarkoitetaan tilannetta, joissa kohdemuuttujan havainnoidut arvot opetusdatassa valikoituvat menneistä päätöksistä ja osa kohdemuuttujan mahdollisista arvoista jää havainnoimatta[96]. Esimerkiksi yliopiston sisäänottoprosessissa ei pystytä havainnoimaan, kuinka potentiaaliset opiskelijat, joilta tosiasiassa evätään pääsy, olisivat pärjänneet, mikäli he olisivat päässeet sisään.

Systemaattiset virheet opetusdatan nimikkeissä. Nimikevinoumia (engl. *label bias*) voi ilmetä myös, kun harjoitus- tai syötedatan esimerkkejä on nimetty tai luokiteltu väärin, esimerkiksi systemaattisesti syrjivällä tavalla[97]. Tällöin vinouma ei johdu merkityksellisten todellisten arvojen puutteista (vrt. valikoitumisvinouma), vaan esimerkiksi kohdemuuttujan operationalisoinnista, mittauksen tai datankeruumenetelmien systemaattisista vinoumista tai konstruktiovaliditeetin puutteesta. Hypoteettisena esimerkkinä voidaan kuvitella, että rekrytoinnin yhteydessä on esimerkiksi mahdollista, että miehiä arvioidaan suotuisammin kuin tosiasiallisesti yhtä päteviä naisia, jolloin arviointiprosessin tuottamaan dataan rakentuu nimikevinouma.

Opetus- tai syötedata heijastaa rakenteellisia eroja populaatiossa. Käytetty data voi toisaalta olla edustavaa tilastollisessa mielessä, mutta silti ns. rakenteellisesti vinoutunutta siten, että ihmisryhmät eivät ole yhtäläisesti edustettuna kohdemuuttujan arvojen eri luokissa. Toisin sanoen malli voi oppia lähes ”virheettömän” mallin populaatiosta tai ilmiöstä, joka ilmentää yhteiskunnille ja populaatioille ominaista monimuotoisuutta, olemassa olevaa epätasa-arvoa tai rakenteellista, yhteiskunnallista stratifikaatiota. Päätöksentekon kannalta merkitykselliset attribuutit (esim. taudin riskitekijät) tai kohdemuuttujan arvot (esim. riski taudille) saattavat olla esiintyvyydeltään erilaisia eri osapopulaatioissa. Yleisesti ottaen opetusdatassa esiintyvien arvojen jakaumat voivat kuvastaa esimerkiksi eroja ryhmien käyttäytymisessä, kyvykkyyksissä ja taidoissa, intresseissä ja preferensseissä, yksilöiden sosioekonomisessa statuksessa jne. Tällaiset erot näkyvät ryhmien välisinä eroina kohdemuuttujan arvoissa ja voivat ilmetä mahdollisena välillisenä syrjintänä mallin käytössä.

96 Fazelpour & Danks, 2021.

97 Barocas & Selbst, 2016.

Keskeinen riski on, että rakenteellisesti vinoutuneella datalla opetetussa mallissa kohdemuuttujan arvojen jakaumat heijastavat historiallista tai olemassa olevaa epätasa-arvoista kohtelua tai hyödykkeiden ja mahdollisuuksien epätasaista jakautumista. Erot voivat olla historiallisesti tuotettuja, mukaan lukien syrjivien käytäntöjen tuotosta. Esimerkiksi ennustavassa poliisitoiminnassa käytetty data saattaa sisältää rikostilastoja, jotka heijastavat tiettyjen asuinalueiden tai etnisten ryhmien suurempaa poliisivalvontaa[98]. Samankaltainen dynamiikka tunnistettiin Yhdysvalloissa terveydenhuollossa käytetystä algoritmista, joka ennusti potilaiden tarvetta terveydenhuollolle käyttämällä hoitomaksuja korvikemuuttujana. Järjestelmä asetti mustat potilaat systemaattisesti epäsuotuisaan asemaan johtuen ryhmien välisistä eroavaisuuksista hoitomaksujen keskimääräisissä suuruuksissa, aliarvioiden mustien potilaiden tarvetta hoidolle.[99]

Opetusdata sisältää syrjivää, halventavaa tai loukkavaa sisältöä. Joissakin raporteissa huomautetaan, että algoritmin opetusdata ja/tai tulosteet voivat olla ns. ”sisällöllisesti” syrjiviä sikäli, kun niiden ilmaisema merkitys yhdistyy esimerkiksi sosiaalisesti vallitseviin normatiivisiin käsityksiin, stereotyyppeihin tai halventaviin käsityksiin. Esimerkiksi luonnollisen kielen prosessoinnissa opetusdatajoukot saattavat sisältää loukkaavaa kieltä tai vihapuhetta, kuten ”slurreja”[100]. Tällöin syrjinnän ja yhdenvertaisuuden näkökulmasta merkityksellinen riski ei liity suoranaisesti esimerkiksi tiettyjen ominaisuuksien todennäköisyysjakaumiin käytetyssä datassa vaan sisältöön, jota algoritmi tuottaa kulloistenkin sosiokulttuuristen merkitysten ja käytänteiden valossa. Riskinä on pikemminkin, että algoritmin tulosteet saattavat heijastaa ja ilmaista joko systemaattisesti tai yksittäistapauksissa syrjiviä tai stigmatisoivia merkityksiä. Esimerkkinä tällaisesta toimii Google Photos -sovelluksen kuvanluokittelualgoritmi, joka luokitteli tummaihoisten ihmisten kuvia nimikkeellä ”gorilla”[101].

Yhdenvertaisuuslain mukaan ”[h]enkilön ihmisarvoa tarkoituksellisesti tai tosiasiallisesti loukkaava käyttäytyminen on häirintää, jos loukkaava käyttäytyminen liittyy 8 §:n 1 momentissa tarkoitettuun syyhyn ja käyttäytymisellä luodaan mainitun syyn vuoksi henkilöä halventava tai nöyryyttävä taikka häntä kohtaan uhkaava, vihamielinen tai hyökkäävä ilmapiiri” (YVL 14 §). Häirinnäksi luokiteltava syrjintä ei ole riippuvaista häirintään syyllistyneen toimijan intentioista, vaan päätelmä häirinnästä voidaan tehdä tarkastelemalla tosiasiallisia vaikutuksia niin kauan kuin häirintä liittyy kiellettyyn syrjintäperusteeseen. Mikäli uhkaaviksi, vihamielisiksi tai hyökkääviksi ilmapiireiksi lasketaan

98 Richardson, Schultz & Crawford, 2019.

99 Obermeyer ym., 2019.

100 Ks. Shearer ym., 2019.

101 Google apologises for Photos app’s racist blunder.” (1.7.2015). BBC News. https://www.bbc.com/news/technology-33347866. Viitattu 1.12.2021.

myös digitaaliset ympäristöt esimerkiksi julkisten palveluiden tai rekrytointipalveluiden yhteydessä, nousee kysymys automatisoidusti tuotetun loukkaava sisällön ja häirinnän suhteesta.

Arviointidatan edustavuus. Koneoppimisjärjestelmiä arvioidaan usein datajoukoilla, joka on erotettu opetusdatajoukosta ennen mallin opettamista tai sovittamista. Arviointidata voi yhtä lailla sisältää yllä käsiteltyjä edustavuus-, otanta- ja nimikevinoumia. Mikäli opetus- ja arviointidata sisältävät samat vinoumat, kyseiset vinoumat saattavat jäädä huomaamatta, kun mallia testataan arviointidatalla. Euroopan digitaaliset oikeudet -edunvalvontaryhmän raportissa todetaankin, että algoritmien auditoimiseen tai testaamiseen käytettävän datajoukon tulisi olla tarpeeksi edustava ja että jokaisen järjestelmän vaikutuksen alaan kuuluvan osapopulaation tulisi olla riittävästi edustettuna koko monimuotoisuudessaan, jotta harhaanjohtavat reiluusarviot voidaan välttää[102].

2.5 Teknologian ulkopuoliset syyt syrjintäriskeille

Tekoälysovellusten teknisellä tasolla tunnistettavat mahdolliset riskit realisoituvat vasta järjestelmän tosiasiallisessa käytöksessä. Tässä osiossa tarkastellaan algoritmisten vinoumien sosioteknisiä syitä sekä välittäviä tekijöitä, jotka vaikuttavat tekoälyn syrjivien vaikutusten syntymiseen, luonteeseen ja skaalaan.

- **Sosiotekninen käyttökonteksti vaikuttaa syrjintäriskien realisoitumiseen ja mahdollisten vaikutusten luonteeseen.** Tekoälyjärjestelmien tosiasialliset vaikutukset ovat riippuvaisia useista kontekstuaalisista sosio-teknisistä tekijöistä, kuten ihmiskäyttäjien arviointikyvystä, kompetensseista ja kognitiivisista vinoumista; sovelluksen käytön skaalasta sekä kohdepopulaatiosta ja sen demografisesta koostumuksesta. Vaihtelu edellä mainituissa tekijöissä voi tarkoittaa, että saman tekoälytuotteen tai -palvelun käytöllä voi olla erilaisia vaikutuksia riippuen kontekstista. Tämä huomio korostaa tarvetta tapauskohtaiselle ja faktuaaliset olosuhteet huomioivalle tarkastelulle mahdollisten syrjivien vaikutusten arvioimisessa.
- **Tekoälyavusteisessa päätöksenteossa järjestelmän käyttäjän (tai operoijan) rooliin tulee kiinnittää huomiota.** Käyttäjien kognitiot ja tulkinta korostuvat syrjivien vaikutusten lähteinä, kun tekoälyjärjestelmän tuottamia tulosteita hyödynnetään päätöksenteossa. Luottamus siihen, että ihmisarvioija pystyy jokaisessa tapauksessa tunnistamaan syrjivät tulosteet

102 Euroopan digitaaliset oikeudet, 2021, 70.

päätöksentekoprosessissa ei ole perusteltua. Avoin kysymys koskee myös sitä, tulisiko vinoutuneet tekoälyn tulosteet tulkita oikeudellisesti käskynä tai ohjeena syrjiä. Väärinkäytön riskeihin (esim. tekoälyn käyttö häirintään tai vihapuheen levittämiseen) tulee kiinnittää huomiota silloinkin, kun järjestelmä itsessään ei sisällä potentiaalisesti syrjiviä vinoumia.

- **Pitkäjänteinen tosiasiallisen ja rakenteellisen eriarvoisuuden vähentäminen edellyttää useiden toimijoiden koordinoitua toimintaa.** Algoritmiset interventiot, joilla pyritään eriarvoisuuden vähentämiseen, saattavat johtaa epäsuotuisiin tai päinvastaisiin lopputulemiin, mikäli näitä interventioita ei suoriteta koordinoidusti eri toimijat, sidosryhmät ja käyttökontekstille ominaiset dynamiikat huomioiden. Esimerkiksi tekoälyjärjestelmien takaisinkytkennät voivat johtaa negatiivisiin spiraaleihin, joissa yksilöiden asema jopa huononee lyhytkatseisten interventioiden johdosta.

2.5.1 Käyttäjät (tai järjestelmän operoijat)

Tekoälysovelluksia käytetään usein ihmisen päätöksenteon tukena, minkä takia on merkityksellistä huomioida välitön ihmiskäyttäjä – esim. päätöksentekijä, joka hyödyntää tekoälyjärjestelmää – mahdollisena syrjivän vinouman lähteenä.

2.5.1.1 Käyttäjän kognitiot ja tulkinnat

Tapauksissa, joissa tekoälysovellus tukee ihmisen päätöksentekoa, käyttäjän kognitiot ja tulkinnat ovat välittävä tekijä mahdollisten syrjivien vaikutusten realisoitumisessa. Tässä suhteessa esimerkiksi käyttäjien kompetenssi järjestelmän käytössä, tulosteiden realisoitumista moduloivat kognitiot ja tulkinnat sekä myös järjestelmän käyttöliittymäominaisuudet osoittautuvat merkityksellisiksi. Tutkimuskirjallisuudessa on esitetty, että ihmisen suorittaman valvonnan ja validoinnin vaatimus (nk. ”human-in-the-loop”) yksinään ei ole kaikissa tapauksissa riittävä keino syrjinnän riskien ehkäisemiseksi[103].

Ihmisen tulkinta algoritmin tulosteesta – tulkinta siitä, mitä ennuste edustaa ja mitä siitä voidaan päätellä – on olennainen mahdollisten syrjivien vaikutusten näkökulmasta. Tutkimuskirjallisuudessa *tulkintavinoumaksi* on kutsuttu ”epäsuhtaa (i) algoritmin tuottaman informaation ja (ii) käyttäjän tai tulostetta hyödyntävän järjestelmän informaatiovaatimusten välillä”[104]. Esimerkkinä voidaan mainita esimerkiksi Yhdysvalloissa joissakin

103 Green, 2021b.

104 Danks & London, 2017, 4.

tuomioistuimissa rikoksenuusimisen riskinarviointiin käytettävä COMPAS-algoritmi, johon sisältyy riski, että järjestelmän tulosteita tulkitaan väärin. Järjestelmässä rikoksenuusimisen riski on operationalisoitu uudelleenpidätyksen todennäköisyytenä. Täten on periaatteessa mahdollista, että käyttäjä voi tulkita tulosteen edustavan todennäköisyyttä, jolla arvioitu yksilö uusii rikoksen, vaikka malli todella ennustaa yksilön todennäköisyyttä tulla pidätetyksi vapauttamisen jälkeen[105]. Empiirisessä tutkimuksessa on myös esitetty, että järjestelmien operoijat ottavat eri tavoilla tulosteita huomioon tekoälyavusteisessa päätöksenteossa[106].

Osassa raportteja korostetaan nk. *automaatioharhan* ongelmallisuutta tulosteiden tulkinnan näkökulmasta: ihmiset ovat taipuvaisia luottamaan kyseenalaistamatta automaattisesti tuotettuihin ennusteisiin ja suosituksiin. Tutkimuskirjallisuus tukee näkemystä: sovellusten ihmiskäyttäjät voivat olla taipuvaisia ottamaan algoritmien ennusteet objektiivisina ja neutraaleina[107]. Empiirisessä tutkimuksessa on esitetty, että järjestelmien operoijat ottavat myös eri tavoilla tulosteita huomioon tekoälyavusteisessa päätöksenteossa[108]. Luottamus algoritmin ennusteisiin saattaa myös olla suurempaa silloin, kun ennusteet sopivat yhteen käyttäjän mahdollisesti syrjivien stereotypioiden, ennakkouskomusten ja -asenteiden kanssa[109]. Mikäli käytetyssä järjestelmässä esiintyy esimerkiksi vallitsevia syrjiviä stereotypioita mukailevia vinoumia, käyttäjän kompetenssi ja kyky arvioida tulosteita yhdenvertaisuusnäkökulmasta muodostuu erityisen tärkeäksi.

2.5.1.2 Käskyt ja ohjeet syrjiä

Yhdenvertaisuuslain mukaan käskyt tai ohjeet syrjiä tulee ymmärtää syrjintänä (YVL 8 §). On avoin kysymys, missä määrin tekoälysovelluksen tuottama ennuste tulisi ymmärtää tällaisena ohjeena tai käskynä (niissä tapauksissa, joissa tulosteet johtavat syrjiviin vaikutuksiin). Yhdessä raportissa huomautetaan, että "vaikka [syrjintään käskemisen tai ohjeistamisen] käsitettä ei ole määritelty Euroopan unionin lainsäädännössä eikä ole vielä tulkittu tuomioistuimissa, innovatiivinen tulkinta 'ohjeistuksesta syrjiä' voisi auttaa välttämään sisällöllisiä ja proseduraalisia esteitä, joita algoritmisen syrjinnän kontekstissa kohdataan"[110].

105 Dressel, & Farid, 2018, 3.
106 Green & Chen, 2019; Green & Chen, 2021.
107 Green, 2021b, luvut 4.1.2. ja 4.1.3.
108 Green & Chen, 2019; Green & Chen, 2021.
109 Green & Chen, 2019.
110 Gerards & Xenidis, 2021, 143.

Mikäli tekoälyjärjestelmien ennusteita voidaan tulkita ohjeina tai käskyinä, merkityksellisiksi oletettavasti muodostuvat seuraavat tekijät: (1) onko tekoälyjärjestelmän tulosteilla potentiaalisesti syrjiviä vaikutuksia, (2) onko järjestelmän kehittäjän tai käyttäjän mahdollista ymmärtää ja/tai selittää järjestelmän toimintaa, (3) minkälainen ohjeistus järjestelmän käyttäjälle on annettu ja (4) kuinka paljon tosiasiallista harkinnanvaraisuutta järjestelmän ennusteiden hyödyntämiseen osana päätöksentekoa sisältyy?

2.5.2 Tekoälysovelluksen käyttötarkoitus ja -konteksti, skaala ja (kohde)populaatio

Tekoälysovelluksia voidaan tyypillisesti käyttää eri kohderyhmiin ja eri tarkoituksiin, ja ne tyypillisesti suunnitellaan laajamittaista käyttöönottoa varten. Käyttötarkoitus-, konteksti, käytön skaala ja kohdepopulaation koostumus vaikuttavat yhdenvertaisuusvaikutusten realisoitumiseen.

2.5.2.1 Sovelluksen käyttötarkoitus ja tuetut toimenpiteet

Yllä todettiin, että tekoälyn tehtävän määrittelyn tavalla on merkitystä yhdenvertaisuusvaikutusten näkökulmasta. Esimerkiksi julkisen sektorin päätöksenteossa niillä voidaan tukea niin ennaltaehkäiseviä toimenpiteitä ja interventioita kuin myös promotiivisia toimenpiteitä. Yhdellä ja samalla järjestelmällä voikin olla negatiivisia tai positiivisia vaikutuksia riippuen niiden tosiasiallisesta käytöstä[111]. Esimerkiksi poliisitoimintaan käytettävissä sovelluksissa mahdolliset syrjivät vaikutukset saattavat olla erilaisia riippuen siitä, pyritäänkö tunnistamaan potentiaalisia rikoksen tekijöitä vaiko uhreja[112]. Maantieteellisten alueiden rikosprevalenssia kuvaavaa järjestelmää voidaan periaatteessa käyttää ennakoivasti rikoksien estämiseen tai tiedonlähteenä poliittisissa toimenpiteissä, kuten korkean rikollisuuden alueiden elinympäristön ja palveluiden kohentamisen.

2.5.2.2 Käytön skaala ja kohdepopulaatio

Tekoälysovelluksia saatetaan käyttää hyvin erilaisissa konteksteissa, jopa kansainvälisesti, jolloin syrjintäriskien tunnistaminen ja ehkäiseminen hankaloituu[113]. Skaalalla ja kohdepopulaatiolla on usein merkitystä sovellusten vaikutusten ja niiden realisoitumisen näkökulmasta (vrt. yllä). Mikäli esimerkiksi (kohde)populaation koostumus vaihtelee käyttökontekstien välillä, syrjintäriskit saattavat realisoitua eri tavalla ja eri mittakaavassa.

111 Ks. esim. Presidentin toimeenpanovirasto, 2016.

112 Centre for Data Ethics and Innovation, 2020b, 41.

113 Selbst ym. 2019.

Minimaaliseltakin vaikuttava riski, kuten kasvojentunnistusteknologian matalaksi luettava virhetaso, voi johtaa useisiin ongelmallisiin oikeudellisiin vaikutuksiin, kun teknologiaa käytetään suurella skaalalla (esim. tuhansiin ihmisiin päivässä)[114]. Jos virhetaso on suurempi esimerkiksi tummaihoisten ihmisten kohdalla, tekee sovellus enemmän virheitä (absoluuttisesti vs. suhteellisesti), kun sitä käytetään alueilla, joissa liikkuu enemmän tummaihoisia ihmisiä. Nämä huomiot ovat merkityksellisiä erityisesti välillisen (mutta myös välittömän) syrjinnän näkökulmasta, koska välillistä syrjintää arvioitaessa tarkastellaan toimenpiteen tai käytännön (vrt. algoritmi) käytön ryhmäkohtaisia vaikutuksia. Tällöin vaikutuksenalaan kuuluvan populaation demografinen koostumus sekä käytön skaala ovat olennaisia välillisesti syrjivien vaikutusten todentamiseksi.

2.5.2.3 Tekoälysovellusten väärinkäyttö

Erityinen riski tässä suhteessa koskee teknologian väärinkäyttöä, jonka mahdolliset syrjivät vaikutukset voivat olla riippumattomia mahdollisista tilastollisista vinoumista datassa, algoritmissa tai mallissa. Esimerkiksi näennäisen harmitonta viihde- ja vapaa-ajan käyttöön tarkoitettua tekoälyteknologiaa voidaan käyttää yhdenvertaisuuden ja tasa-arvon näkökulmista ongelmallisilla tavoilla. Muun muassa syvävärennös-teknologialla (engl. *deepfake*) voidaan tuottaa realistisia mutta keinotekoisia ääni-, kuva- ja videomateriaaleja, joita saatetaan käyttää häirintään[115]. Teknologian käytön ongelmalliset vaikutukset eivät näissä tapauksissa ole suoraan riippuvaisia tekoälysovelluksen mahdollisista vinoumista.

Mahdollisesti vinoutuneita koneoppimisalgoritmeja saatetaan hyödyntää myös sovelluksissa, alustoissa tai digitaalisessa infrastruktuurissa, joita itsessään saatetaan käyttää – mahdollisesti väärin tai muutoin kiistanalaisesti – julkisten toimijoiden tehtävissä. Esimerkiksi Yhdysvaltain kansalaisuus- ja maahanmuuttovirasto USCIS on käyttänyt konekääntämissovelluksia pakolaisten sosiaalisen median tilien sisällön tulkkaamisessa tarkistusprosesseissa, vaikka konekääntämissovelluksissa on tunnistettu merkittäviä vinoumia[116]. Kasvojentunnistussovellus Clearview AI:n kiistanalaista käyttöä poliisin toimesta

114 Euroopan unionin perusoikeusvirasto, 2020, 28.

115 Allen QC & Masters, 2020, 41.

116 "Google Says Google Translate Can't Replace Human Translators. Immigration Officials Have Used It to Vet Refugees". (26.9.2019). ProPublica. https://www.propublica.org/article/google-says-google-translate-cant-replace-human-translators-immigration-officials-have-used-it-to-vet-refugees. [Viitattu 23.11.2021]

on ilmennyt niin kansainvälisellä tasolla[117] kuin Suomessakin[118], vaikka kasvojentunnistusteknologian on yleisesti tunnistettu sisältävän merkittäviä syrjintäriskejä sekä tieto- ja yksityisyydensuojaongelmia[119].

Syrjintäriskien arvioinnissa tulisi ottaa huomioon yllä kuvattujen kaltaiset riskit teknologian väärinkäytölle ”alajuoksussa” käyttäjien toimesta. Joissakin raporteissa huomioitiin esimerkiksi mahdollisuus hyödyntää dokumentaatiomenetelmiä (ks. alla) sekä erilaisia vakuutuksia ja sertifikaatteja[120] vastuullisen käytön edistämiseksi alajuoksussa.

Yhdenvertaisuuslain ja tasa-arvolain soveltamisala huomioiden tässä raportissa todetaan, että laajemmat sosiaalieettiset ja rakenteelliset yhdenvertaisuus- ja tasa-arvonäkökulmat on myös huomioitava tekoälyn mahdollisten epäsuotuisten vaikutusten kontekstissa. Tekoälysovellustenkäyttö ja sosiaalieettisessä mielessä merkitykselliset yhdenvertaisuus- ja tasa-arvovaikutukset eivät rajoitu yhdenvertaisuuslain ja tasa-arvolain soveltamisalueille vaan strukturoivat yksityis- ja perhe-elämää sekä uskonnonharjoittamista sikäli, kun näiden elämänalueiden toimintaa harjoitetaan digitaalisissa ympäristöissä ja/tai tekoälysovellusten avulla.

2.5.3 Algoritmiset interventiot ja kontekstuaaliset ja ajalliset dynamiikat

Yhdenvertaisuusvaikutuksia voidaan tarkastella myös ajallisesta ja dynaamisesta näkökulmasta huomioimalla, että tekoälyjärjestelmiä käytetään tyypillisesti jatkuvissa päätöksentekoprosesseissa ja dynaamisesti muuttuvissa käyttökonteksteissa ja populaatioissa. Erotamme tässä suhteessa kolme eri dynamiikkaa, jotka ovat merkityksellisiä tässä suhteessa. Näiden kolmen dynamiikan kohdalla merkityksellisen eron voi tehdä se, hyödyntääkö käytetty tekoälyjärjestelmä jatkuvaa oppimista vaiko staattista mallia.

---

117 “The NYPD used a controversial facial recognition tool. Here’s what you need to know.” (9.4.2021). *MIT Technology Review*. https://www.technologyreview.com/2021/04/09/1022240/clearview-ai-nypd-emails/. [Viitattu 23.11.2021].

118 ”KRP:ssä tehdystä kasvontunnistusohjelman testikäytöstä tehty ilmoitus tietosuojavaltuutetulle”. (9.4.2021). *Poliisi*. https://poliisi.fi/-/krp-ssa-tehdysta-kasvontunnistusohjelman-testikaytosta-tehty-ilmoitus-tietosuojavaltuutetulle. [Viitattu 23.11.2021]; ”Amerikkalaismedia varoitti Suomen poliisia kiistanalaisen kasvojentunnistusohjelman käytöstä – KRP kompuroi vastauksessaan”. (23.4.2021). *Yle Uutiset*. https://yle.fi/uutiset/3-11898702. [Viitattu 23.11.2021].

119 Ks. esim. Buolamwini & Gebru, 2018; Data for Black Lives & Demos, 2021; Centre for Data Ethics and Innovation, 2020b; Euroopan komissio, 2020.

120 Ks. esim. Centre for Data Ethics and Innovation, 2020a; UNESCO, 2020; ks. myös Euroopan unionin perusoikeusvirasto, 2020.

2.5.3.1 Jatkuvasti oppivat mallit

Erityisesti tutkimuskirjallisuudesta mutta myös osasta aineistoa nouseva huomio tekoälysovellusten yhdenvertaisuusvaikutusten realisoitumista koskien liittyy niiden käyttöympäristöjen ja kohdepopulaatioiden dynaamisiin muutoksiin. Jatkuvaa oppimista (engl. *continuous learning* tai *online learning*) hyödyntävien tekoälysovellusten mallit muuttuvat dynaamisesti uuden syötedatan pohjalta ja tekoälysovelluksia myös päivitetään iteratiivisesti. Jatkuvassa oppimisessa malli muokkautuu käytössä dynaamisesti siihen syötetyn syötedatan pohjalta alustavien opetuskierrosten jälkeenkin ja tällöin sen soveltamat suosittelu- tai päätöksentekomallit voivat muuttua dynaamisesti. Yhdellä ajanhetkellä tarkasteltu malli saattaa täten sisältää erilaisia vinoumia kuin toisena ajanhetkenä tarkasteltuna, jolloin ”suhde algoritmin ennusteiden ja kiellettyjen syrjintäperusteiden välillä sekä niiden korrelaation vahvuus saattaa muuttua ajassa algoritmin mukana”[121]. Dynaamiset muutokset tekoälyn käyttöympäristössä ja kohdepopulaatiossa korostavat tarvetta syrjintäriskien arvioimisen jatkuvuudelle ja/tai säännöllisyydelle jatkuvasti oppivien algoritmien kohdalla.

2.5.3.2 Takaisinkytkentä

Jatkuvissa päätöksentekoprosesseissa (vrt. esim. luottokelpoisuuden arviointi) voi syntyä yhdenvertaisuus- ja tasa-arvonäkökulmasta ongelmallisia takaisinkytkentöjä, joita on tutkimuskirjallisuudessa kutsuttu ”negatiivisiksi spiraaleiksi”[122]. Yksilö voi saada ensin päätöksen (joka itsessään voi olla perusteltu tai syrjivä), joka vaikuttaa negatiivisesti esimerkiksi hänen taloudelliseen tilanteeseensa. Nämä vaikutukset voivat toimia perusteena kohdella samaa yksilöä epäsuotuisasti joko saman päätöksentekokontekstin toisella kierroksella (esimerkiksi, kun yksilö hakee pidennystä lainaansa) tai toisessa päätöksentekokontekstissa (esimerkiksi, kun yksilö hakee vuokra-asuntoa). Negatiivisia spiraaleja, joissa huono-osaisten henkilöiden tilanne huonontuu jatkuvasti takaisinkytkennän vuoksi, voi muodostua niin staattisten mallien kuin jatkuvasti oppivien algoritmienkin yhteydessä. Esimerkiksi malleja päivitettäessä “vinoumat voivat vahvistua takaisinkytkentöjen myötä, kun malleja opetetaan vähitellen uudelleen datalla, joka on tuotettu täysin tai osittain mallin aikaisempien versioiden käytössä osana päätöksentekoa”[123].

121 Gerards & Xenidis, 2021, 66.

122 Citron & Pasquale, 2014; Zarsky, 2014.

123 Centre for Data Ethics and Innovation, 2020a, 20.

Euroopan komission ehdotuksessa visioidaan velvollisuutta takaisinkytkennän riskien arvioimiseen ja ehkäisemiseen ennen käyttöönottoa:

> Suuririskiset tekoälyjärjestelmät, jotka jatkavat oppimista markkinoille saattamisen tai käyttöönoton jälkeen, on kehitettävä siten, että varmistetaan, että mahdollisesti vinoutuneisiin tuloksiin, jotka johtuvat tulosten käytöstä tulevien toimintojen syöttötietona […] puututaan asianmukaisin korjaavin toimenpitein.[124]

Takaisinkytkennät ja nk. negatiivisten spiraalien riskit korostavat kuitenkin yhdenvertaisuusvaikutustenarvioinnin jatkuvuuden ja säännöllisyyden merkitystä[125].

2.5.3.3 Käyttökontekstien ja kohdepopulaatioiden dynaamiset muutokset

Osassa raportteja huomioitiin, että muutokset populaatiossa ja käyttökontekstissa voivat olla merkityksellisiä mahdollisten syrjivien vaikutusten syntymisen suhteen[126]. Tässä voidaan erottaa erilaisia muutoksia riippuen siitä, muuttuuko riippumattoman vaiko riippuvan muuttujan esiintyvyys populaatiossa vaiko näiden suhde:

I. Riippumattoman muuttujan ajallinen muutos ilmenee syötedatan attribuuttien jakauman muutoksena. Tätä kutsutaan datan kulkeutumiseksi (engl. *data drift*) tai kovariaattien muutokseksi.
II. Riippuvan muuttujan ajallinen muutos ilmenee kohdemuuttujan tosiasiallisen jakauman muutoksena. Tätä kutsutaan ennakko- tai apriori-todennäköisyyden muutokseksi (engl. *prior probability shift*).
III. Konstruktin tai käsitteen ajallinen muutos ilmenee, kun tosiasiallinen suhde riippumattoman ja riippuvan muuttujan – ts. syöteattribuuttien ja kohdemuuttujan – välillä muuttuu. Tätä kutsutaan konstruktin tai käsitteen muutokseksi (engl. *concept drift*).

Kaikista näistä muutoksista voi seurata heikentyminen mallin tosiasiallisessa ennustevoimassa. Mahdollisten syrjivien vaikutusten näkökulmasta ne ovat kuitenkin merkityksellisiä erityisesti, mikäli kyseiset muutokset koskevat vain tiettyjä ihmisryhmiä, koska tämä saattaa vaikuttaa käytetyn järjestelmän ryhmäkohtaisiin vaikutuksiin.

---

124 TEKOÄLYÄ KOSKEVISTA YHDENMUKAISTETUISTA SÄÄNNÖISTÄ (TEKOÄLYSÄÄDÖS) JA TIETTYJEN UNIONIN SÄÄDÖSTEN MUUTTAMISESTA, artikla 15, kohta 5.

125 Euroopan unionin perusoikeusvirasto, 2020, 99.

126 Ks. esim. Algo:aware, 2018; Euroopan digitaaliset oikeudet, 2021.

Korostamme, että nämä muutokset voivat seurata jopa näennäisen ”reilujen” mallien käytössä, mikä tuottaa haasteita myös syrjivien vaikutusten ehkäisemiseen pyrkivien menetelmien käytössä (ks. alla).

2.6 Tekoälysovellusten yhdenvertaisuusvaikutusten arvioiminen ja hallinta

Tässä osiossa siirrytään tarkastelemaan löydöksiä, jotka liittyvät mahdollisten syrjivien vaikutusten tunnistamiseen ja ehkäisemiseen liittyviin keinoihin. Tässä osiossa esitellyt löydökset koodattiin analyysivaiheessa kartoitusta ohjanneen kolmannen tutkimuskysymyksen (*mitä keinoja* [*algoritmisen syrjinnän*] *riskien ehkäisemiseksi on tunnistettu?*) alle.

- **Opetusdatan vinoumien tunnistamiseksi on suositeltu muun muassa datan laadunarviointiin keskittyviä toimenpiteitä ja prosesseja.** Tekoälyjärjestelmiä kehittävien ja/tai käyttävien organisaatioiden suositellaan tarkastelevan datan edustavuutta ja muita laatutekijöitä mahdollisten syrjivien vaikutusten tunnistamiseksi. Dokumentaatiokeinot, kuten nk. datakortit, nähdään osaltaan vastuullista ja läpinäkyvää tekoälyjärjestelmien käyttöä edesauttavina keinoina.
- **Keskeisin aineistossa tunnistettu keino syrjintäriskien arvioimiseksi on algoritmien testaaminen ja auditointi ongelmallisten vinoumien varalta.** Tekoälyjärjestelmien auditoimiseen kehitettyjen tilastollisten testien ja nk. oikomismenetelmien, joita voidaan hyödyntää järjestelmien arvoketjun eri vaiheissa tasapainottamaan mallin tulosteiden jakaumaa, odotetaan muodostuvan standardimenetelmiksi.
- **Teknisellä tasolla syrjintäriskien tunnistamista ja hallintaa voivat edesauttaa myös tekoälyn selitettävyyttä ja ymmärrettävyyttä lisäävät menetelmät – nk. selitysmenetelmät.** Monimutkaisten ”musta laatikko” -mallien selitettävyyden lisäämisen odotetaan mahdollistavan objektiivisen oikeutuksen antamista ja parantavan päätöksenteon kohteena olevien ihmisten pääsyä oikeuksiinsa.
- **Aineistosta nousee esille myös tarve laajemmille proseduraalisille, organisaatiotason ja lainsäädännön tason keinoille.** Teknologisen kehyksen ulkopuolella toteuttaviksi keinoiksi tunnistettiin muun muassa laajemmat riskin- ja vaikutuksenarviointiprosessit; heuristiikkojen ja tarkistuslistojen hyödyntäminen; työvoiman kouluttaminen ja poikkitieteellisyyden, diversiteetin ja inklusiivisuuden lisääminen; datan ja mallien dokumentaatio; vaihtoehtoisten menettelytapojen käyttöönotto sekä päätöksenteon lopputulemien riitauttamisen ja hyvityksen saamisen mahdollistaminen.

2.6.1 Datan laadunarviointi

Opetusdatan arvioiminen vinoumien suhteen sekä laajempi dataan keskittyvä auditointi koettiin useissa raporteissa merkityksellisesti sikäli, kun datan vinoumat nostettiin esille suurimpana riskinä negatiivisille yhdenvertaisuusvaikutuksille. Datan auditoinnit voivat esimerkiksi kolmansien osapuolien suorittamana auttaa ”kehittäjiä varmistamaan, että heidän algoritminsa toimii ja että se on soveltuvien lakien ja eettisten normien mukainen”[127].

Keskeisiä opetusdatan suhteen arvioitavia laadullisia ominaisuuksia nähtiin olevan datan tarkkuus ja edustavuus[128]. Datan validiteetti ja reliabiliteetti (ml. esim. mittausvirheiden ja vanhentuneiden datapisteiden mahdollisuus), epätäydellisyys ja puutteet sekä esimerkiksi vähemmistö- ja marginalisoitujen ryhmien edustus opetusdatassa muodostuvat näistä näkökulmista olennaisiksi.

Datan laadunvalvontaan suositeltiin niin mallintamista edeltäviä tarkistuksia kuin jatkuvia valvontamenetelmiä ja -prosessejakin[129]. Raporteissa painotettiin myös yhteisten standardien kehittämisen tärkeyttä ja hyödyntää alan asiantuntijoiden konsultoimista datan keruun ja laadunarvioinnin suhteen[130]. Tutkimuskirjallisuudesta kumpuavat dokumentointimenetelmät, kuten data- ja mallikortit, nähtiin myös tapana edesauttaa vastuullista ja läpinäkyvää datan käyttöä[131]. Datankeruun suorittanut taho voisi esimerkiksi toimittaa mallin kehittäjille tarvittavaa, dokumentoitua informaatiota opetusdatasta ja sen ominaisuuksista, mukaan lukien laadusta, mahdollisista puutteista ja soveltuvista käyttökohteista.

2.6.2 Algoritmien auditointi ja teknisen tason menetelmät

Keskeisin menetelmä syrjintäriskien arvioimisessa, joka useissa tarkasteltuun aineistoon sisältyvissä raporteissa mainittiin, on tekoälysovellusten testaaminen vinoumien varalta. Datan laadun ja edustavuuden arviointi esitettiin merkityksellisenä tässä suhteessa ja useissa raporteissa erityisenä teknisenä menetelmänä mainittiin algoritmien auditointi sekä mallien testaaminen nk. reiluuden suhteen. Auditointimenetelmiksi mainittiin muun

127 The Greenlining Institute, 2020, 27.

128 Ks. esim. Euroopan unionin perusoikeusvirasto, 2018.

129 Esim. UNESCO, 2020.

130 UNESCO 2020, 21.

131 Euroopan unionin perusoikeusvirasto, 2019, 14. Ks. myös Gebru ym., 2021; Holland, Hosny & Newman, 2020; Stoyanovich & Howe, 2019.

muassa keinotekoisten profiilien avulla suoritettavat kontrafaktuaaliset testit ja kehitetyt tilastolliset menetelmät, kuten reiluusmetriikoiden avulla suoritettavat tilastolliset testit (ks. alla).

2.6.2.1 Reiluusmetriikat

Reiluusmetriikat (engl. *fairness metrics*) ovat testejä ja mittareita, joita käytetään tunnistamaan vinoumia malleissa. Reiluusmetriikoita on esitetty useita kymmeniä tutkimuskirjallisuudessa[132] ja niihin lukeutuu erilaisia tilastollisia testejä sekä menetelmiä, joilla tekoälysovellusten laskennallisen prosesseja muuttujien välisiä suhteita tai tulosteiden jakaumia voidaan arvioida yhdenvertaisuuden tai tasa-arvon näkökulmasta. Näiden lähestymistapojen taustamotivaationa on löytää tilastollisia tai mallin suhteen kausaalisia riippuvuussuhteita kiellettyjen syrjintäperustojen ja mallin ennusteiden tai siitä laskettavien erinäisten tilastollisten suhdelukujen välillä. Erilaiset lähestymistavat ”algoritmiseen reiluuteen” voidaan luokitella karkeasti kolmeen luokkaan[133]:

I. Yksilöreiluuden testeihin, joilla arvioidaan, kohteleeko algoritmi samankaltaisia mutta sensitiivisiltä ominaisuuksiltaan, kuten sukupuolelta, eroavia yksilöitä samalla tavalla.
II. Ryhmäreiluuden tilastollisiin testeihin, joilla tarkastellaan ryhmien välisiä eroja tilastollisissa mitoissa (esim. väärän positiivisen ennusteen todennäköisyys).
III. Kausaalisiin ja kontrafaktuaalisiin lähestymistapoihin, joilla pyritään arvioimaan, kuinka paljon kielletyt syrjintäperusteet vaikuttavat tulosteen arvoon käytetyssä mallissa (riippumatta kyseisten perusteiden ennustevoimasta).

Reiluuden mittareita käytetään myös algoritmien (uudelleen)optimointikriteereinä mallin opetusvaiheessa ja täten niitä voidaan ajatella myös eräänlaisina formalisoituina määritelminä reiluudelle (engl. *fairness definition*). Tutkimuskirjallisuudessa esitettyjen mittareiden taustalla onkin erilaisia oikeudellisia ja filosofisia näkemyksiä syrjinnästä ja tasa-arvosta (vrt. esim. Yhdysvaltain lainsäädännön 4/5-sääntö välillisen syrjinnän arvioimisessa). Koneoppimiskirjallisuudessa esitellyt määritelmät ovat kuitenkin osa laajempaa teoreettisen tutkimuksen perinnettä, jossa on kehitetty tilastollisia menetelmiä vinoumien ja syrjinnän tunnistamiseksi[134].

132 Ks. Verma & Rubin, 2018.
133 Fazelpour & Danks, 2021.
134 Hutchinson & Mitchell, 2019.

Suurin osa reiluusmetriikoista on kehitetty luokittelu- ja regressioalgoritmien kontekstiin[135] (vrt. yllä). Kirjallisuudessa on esitelty myös suosittelujärjestelmien[136] ja klusterointialgoritmien[137] reiluuden arvioimiseen kehitettyjä mittareita. Pelkästään luokittelu- ja regressioalgoritmeille on olemassa yli 20 reiluuden mittaria[138] ja esimerkiksi klusterointialgoritmeille yli 15[139]. Luokittelu- ja regressioalgoritmien kontekstissa käytetyt tilastolliset mittarit laskevat jonkin tilastollisen suhdeluvun eroja valittujen vertailuluokkien välillä (esim. miehet ja naiset). Mitatut tilastolliset suhdeluvut saadaan sekoitusmatriisista (Taulukko 6).

Taulukko 6. Sekoitusmatriisi

	Todellinen arvo (posiitivinen)	**Todellinen arvo (negatiivinen)**
Ennusteen arvo (positiivinen)	*Tosi positiivinen* (TP) *Positiivinen ennustearvo* (PPV): TP/(TP+FP) *Sensitiivisyys* (TPR): TP/(TP+FN)	*Väärä positiivinen* (FP) *Väärien löydösten osuus* (FDR): FP/(TP+FP) *Väärien positiivisten osuus* (FPR): FP/(FP+TN)
Ennusteen arvo (negatiivinen)	*Väärä negatiivinen* (FN) *Väärien omissioiden osuus* (FOR): FN/(TN+FN) *Väärien negatiivisten osuus* (FNR): FN/(TP+FN)	*Tosi negatiivinen* (TN) *Negatiivinen ennustarvo* (NPV): TN/(TN+FN) *Spesifisyys* (TNR): TN/(TN+FP)

Reiluusmetriikoiden luonteen kuvaukseksi tässä yhteydessä voidaan mainita muutamia tutkimuskirjallisuudesta nousevia luokittelu- ja regressioalgoritmien metriikoita, kuten *(ehdollinen) tilastollinen pariteetti, kalibraatio, virhetasojen yhtäläisyys* ja *yksilöreiluus*. Taulukko 7 sisältää nämä mittarit ja lyhyet kuvaukset.

135 Ibid.; ks. myös Mitchell ym., 2021.

136 Ks. esim. Elahi ym., 2021.

137 Chhabra, Masalkovaitė & Mohapatra, 2021.

138 Verma & Rubin, 2018; Mitchell ym., 2021.

139 Chhabra, Masalkovaitė & Mohapatra, 2021.

Taulukko 7. Reiluusmetriikat.

Määritelmä	Kuvaus	Mittari
Muodollisesti samanlainen kohtelu (vrt. ”sokeus kielletyille syrjintäperusteille”)	Kiellettyjä syrjintäperusteita ei käytetä mallin muuttujina	$G \notin V$
Tilastollinen pariteetti (Dwork et al. 2012)	Vertailtavat ryhmät ovat yhtä todennäköisiä saamaan positiivisen ennusteen. *Esim. positiivisen ennusteen todennäköisyys tulee olla riippumaton henkilön ikäryhmästä.*	$P(D = 1 \mid g_a) = P(D = 1 \mid g_b)$
Ehdollinen tilastollinen pariteetti (Dwork et al., 2012)	Kiellettyihin syrjintäperustoihin kuuluvat ryhmät ovat yhtä todennäköisiä saamaan positiivisen ennusteen, mutta tämä on ehdollisena tiettyjen muuttujien arvoille. *Vrt. tilastollinen pariteetti, kun vaatimus riippumattomuudelle iästä on ehdollinen tiettyjen muuttujien arvoille.*	$P(D = 1 \mid L, g_a) = P(D = 1 \mid L, g_b)$
Kalibraatio (Chouldechova, 2017)	Todennäköisyyslaskelmien tulee heijastaa populaation todellista todennäköisyysjakaumaa. *Esim. tulosteen edustama todennäköisyysarvio on osumatarkkuudeltaan yhtäläinen eri etnisen taustan omaavien henkilöiden välillä.*	$P(Y = 1 \mid S, g_a) = P(Y = 1 \mid S, g_b)$
Virhetasojen yhtäläisyys (Hardt et al. 2016)	Kiellettyihin syrjintäperustoihin kuuluvilla ryhmillä tulee olla yhtäläiset virhetasot. *Esim. todennäköisyys väärälle negatiiviselle ja väärälle positiiviselle ennusteelle tulee olla yhtä suuri riippumatta henkilön etnisestä taustasta.*	$P(D = 1, \mid Y, g_a) = P(D = 1 \mid Y, g_b)$
Yksilöreiluus (Dwork et al. 2012)	Jokaiselle parille yksilöitä pätee mallissa, että heidän ennusteensa eroaa vain siinä määrin, mikä on laskettu ero heidän välillään silloin, kun laskettuun eroon ei vaikuta kielletty syrjintäperuste. *Esim. pätevyydeltään samankaltaisia yksilöitä pitäisi kohdella samalla tavalla riippumatta heidän etnisestä taustastaan.*	[Formalisointi edellyttää mittaria samankaltaisuudelle, kuten Euklidiaanista etäisyyttä.]

Termien selitteet: Todennäköisyyttä merkataan termillä *P*. Binääriarvoista ennustetta merkataan $D = \{0, 1\}$ ja todellista arvoa $Y = \{0, 1\}$. Todennäköisyyslukema, joka edustaa mallin pohjalta tehtyä todennäköisyyslaskelmaa, merkataan *s* siten, että $s \in S$ ja siten, että $S = \{0, 1\}$. *S* määrää mallissa *D*:n arvon yksilölle $i \in I$ riippuen mallin oppimasta tai päätöksentekijän asettamasta kynnysarvosta. Mallin muuttujia merkataan termillä *V*. Tämän joukon osajoukko on syrjintäperusteet $G \in V$, ja jokaiselle yksilölle *i* pätee, että hänellä on tietyn syrjintäperusteen arvo *g* siten, että tämä arvo on binäärinen, kategorinen tai jatkuva. Esimerkiksi yksilö voi olla siviilisäädyltään naimisissa tai naimaton $g_{\text{siviilisääty}} = \{a, b\}$ tai kuulua tiettyyn etniseen ryhmään $g_{\text{etnisyys}} = \{a, b, \ldots, n\}$. *V*:n osajoukoksi voidaan lukea myös tietyissä tapauksissa oikeutettavuusperusteet *L*, jotka saavat tuottaa riippuvaisuuksia *G*:n ja *D*:n välille.

Huomattakoon, että Taulukossa 7 kuvatut mittarit mittaavat syrjinnän kiellon näkökulmasta erilaisia asioita tekoälysovelluksen toimintaan liittyen. Esimerkiksi *yksilöreiluuden* mittari tarkastelee näitä vaikutuksia vertailemalla yksilöitä ja yksittäisiä ennusteita, kun taas muut mittarit vertailevat ryhmiä. *(Ehdollisen) tilastollisen pariteetin* mittari puolestaan tarkastelee ennusteiden (vrt. päätösten) jakautumista, kun taas *kalibraatio* ja *virhetasojen yhtäläisyys* vertailevat arvioitujen todennäköisyyksien ja ennusteiden tarkkuutta ja virheiden määriä ryhmien välillä. Tutkimuskirjallisuudessa on huomautettu, että tästä syystä sopivan mittarin valinta ei ole eettisesti, poliittisesti tai oikeudellisesti neutraali valinta[140].

Reiluusmetriikat eivät myöskään ole suoraan sovellettavissa läpi koneoppimismenetelmien ja algoritmityyppien. Esimerkiksi luokittelu- ja regressioalgoritmit antavat ennusteen, jonka arvo saattaa olla binäärinen (esim. ”kyllä”/”ei”) tai jatkuva (esim. ”90 % todennäköisyys). Suosittelujärjestelmät taas usein esittävät käyttäjälle suositeltavaa sisältöä soveltuvuuden mukaisesti järjestetyissä listoissa, jolloin käyttäjän näkemä tuloste voi koostua yhdestä tai useammasta ennusteesta (esim. järjestetty lista suositeltuja tuotteita). Tällöin voi olla esimerkiksi tarpeen tarkastella ”suosituslistojen” reiluutta yksittäisten ennusteiden sijaan. Luokittelu- ja regressioalgoritmeja myös opetetaan usein ohjatulla oppimisella ja niiden reiluutta voidaan arvioida vasten todellisia lopputulemia eli havaittua pohjatotuutta. Ohjaamattomassa oppimisessa, kuten klusteroinnissa, tällaista pohjatotuutta ei kuitenkaan välttämättä ole saatavilla. Kuten mainittu yllä, pohjatotuuden saatavuus on rajattua myös kontrafaktuaalisten lopputulemien havainnoimisen haasteellisuudesta johtuen (vrt. nimikkeiden valikoitumisvinouma).

EU:n yhdenvertaisuuslainsäädännön kontekstiin keskittyvässä tutkimuksessa on esitetty, että *ehdollinen tilastollinen pariteetti* (ks. Taulukko 7) mukailee parhaiten Euroopan unionin tuomioistuimen ”kultaista standardia” syrjintätapauksia koskevassa oikeuskäytännössä. Yksinkertaisimmillaan testi tarkoittaisi esimerkiksi yliopistovalintojen kontekstissa seuraavaa: Mikäli valintaperusteena toimisi todistuksen keskiarvo, testin mukaan meidän pitäisi olettaa, että mikäli 70 % miehistä, joilla on 9.0 keskiarvo pääsee sisään, myös 70 % naisista, joilla on 9.0 keskiarvo pitäisi päästä sisään. Ehdollisen tilastollisen pariteetin laskelma perustuu siis kielletyn syrjintäperusteen alle kuuluvien ryhmien jäsenten ehdolliseen todennäköisyyteen saada positiivinen ennuste siten, että laskelmassa sallitaan tiettyjen muuttujien (esim. ryhmäkohtaisen hakuprosentin) tuottaa eroja ryhmienväliseen edustukseen hyväksyttyjen hakijoiden luokassa. Olennainen kysymys onkin, mitkä muuttujat voivat kussakin tapauksessa tuottaa objektiivisesti oikeutettavia eroja tässä suhteessa, jotta päätöksentekoprosessi ei olisi välillisesti syrjivä.

140 Narayanan, 2018; Wachter, Mittelstadt & Russell, 2020; Wachter, Mittelstadt & Russell, 2021.

Niin eurooppalaista kuin Yhdysvaltojen yhdenvertaisuuslainsäädäntöä käsittelevässä tutkimuksessa on kuitenkin pidetty kyseenalaisena ajatusta, jonka mukaan reiluusmetriikat olisivat yksinään riittäviä yhdenvertaisuusvaikutusten arvioimiseen ja syrjinnän todentamiseen – ts., että olisi yksi ainoa syrjintäriskien arviointiin soveltuva reiluusmittari[141]. Osassa tarkastelluista raporteista ilmaistaan sama huoli[142]. Tässä raportissa huomautamme myös, että tilastolliset testit algoritmien reiluudelle eivät välttämättä ole riittäviä ehkäisemään esimerkiksi loukkaavaan tai halventavaan sisältöön ja algoritmien läpinäkyvyyteen ja/tai selitettävyyteen liittyviä oikeudellisesti ja eettisesti merkityksellisiä vaikutuksia.

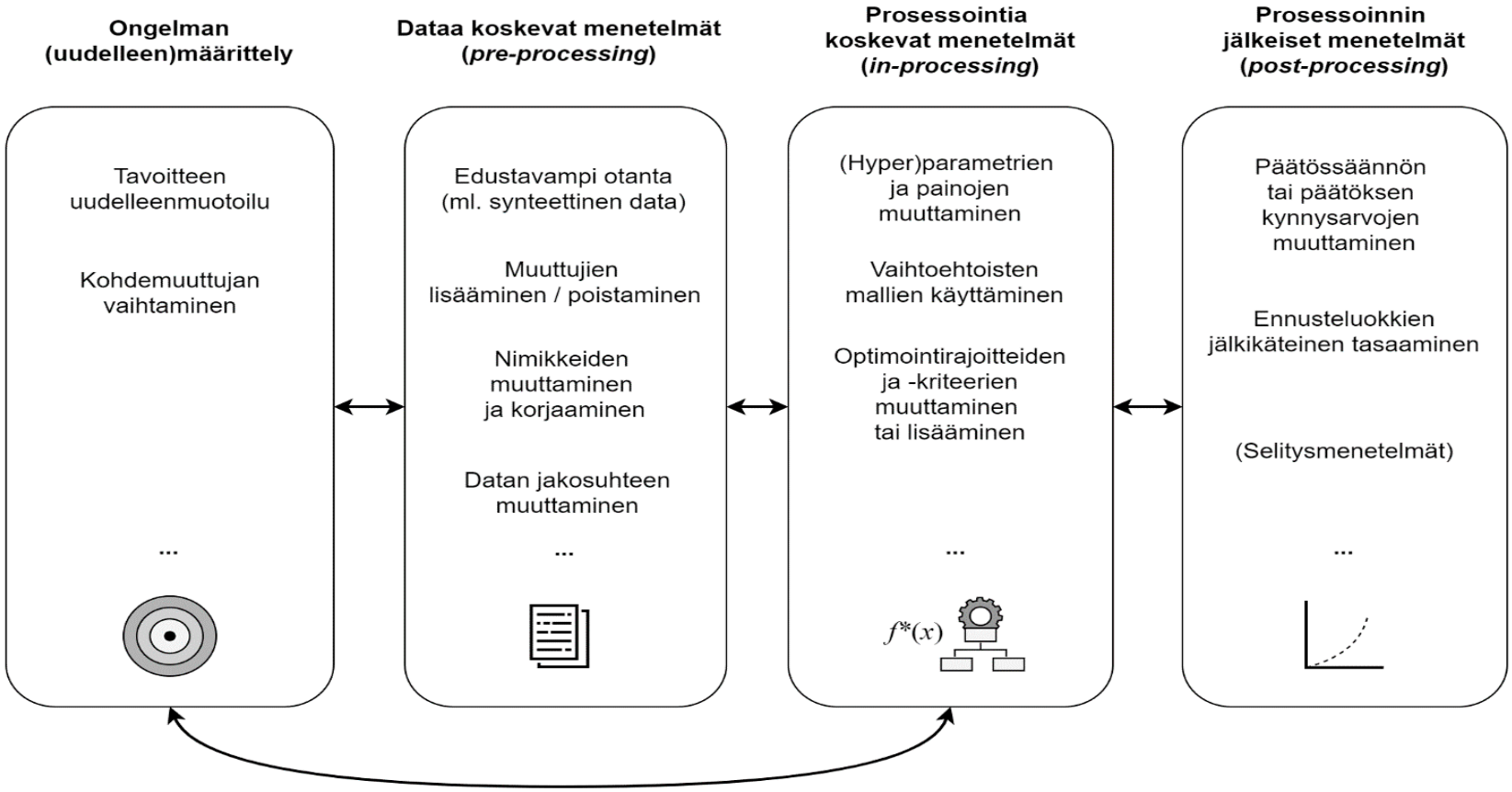
2.6.2.2 Vinoumien ”oikomismenetelmät”

Tunnistettujen vinoumien korjaamiseksi on esitetty useita teknisiä menetelmiä tutkimuskirjallisuudessa, ja näitä menetelmiä suositellaan käytettäväksi myös osassa tarkastelluista raporteissa. Tässä raportissa näitä menetelmiä kutsutaan ”vinoumien oikomismenetelmiksi” (engl. *de-biasing methods* tai *bias mitigation methods*). Oikomismenetelmillä pystytään kuitenkin osaltaan vastaamaan tunnistettuihin syrjintäriskeihin jo sovelluksen kehittämisen aikana muokkaamalla sen komponentteja, ja niitä voidaan käyttää eri vaiheissa koneoppimissovelluksen kehitysprosessia (ks. Kuva 2 alla). Korostamme kuitenkin, että vinoumat sinänsä tulee ymmärtää aina jonkin evaluatiivisen standardin kontekstissa – ts. niiden eliminoiminen täysin ei ole mahdollista sosioeettisesti ”neutraalin” mallin saavuttamisen mielessä.

141 Wachter, Mittelstadt & Russell, 2020; Xiang & Raji, 2019.

142 Ks. esim. Euroopan digitaaliset oikeudet (2021), jossa esitetään kattava kuvaus algoritmisen reiluuden ongelmista tässä suhteessa.

Kuvio 2. Vinoumien ehkäiseminen ja oikominen teknisin menetelmin.



Prosessointia edeltävillä menetelmillä (engl. *pre-processing methods*) pyritään parantamaan datan laatua ja/tai edustavuutta ennen mallin opettamista esimerkiksi esiprosessoimalla opetusdataa tai parantamalla aliedustettujen ryhmien edustusta siinä[143]. Osassa tarkasteltuja raportteja myös esitettiin synteettisen datan käyttämistä opetusdatan vinoumien oikomisessa[144].

Opetusprosessia voidaan kontrolloida muun muassa asettamalla sille laskennallisia rajoituksia tai muokkaamalla tavoitefunktiota (engl. *in-processing methods*)[145]. Edellisissä tapauksissa reiluusmetriikoita voidaan käyttää rajoittamaan optimointiprosessia siten, että algoritmi maksimoi mallin tarkkuutta niin kauan, kun ryhmien edustus esimerkiksi väärien negatiivisten luokassa ei eroa.

Jo opetetun mallin (nk. ”musta laatikko -mallin”) ennusteita voidaan myös muokata jälkikäteen kajoamatta harjoitusdataan tai mallin ominaisuuksiin (engl. *post-processing methods*)[146]. Epäsuotuisaa kohtelua kohtaavien ryhmien tai vähemmistöryhmien edustusta positiivisessa ennusteluokassa voidaan parantaa esimerkiksi vaihtamalla tulosteen tai ennusteen arvoa, mikäli se on lähellä päätössäännön kynnysarvoa (vrt. esim. sisäänpääsyyn vaadittava pisteraja yliopistoissa).

Oikomismenetelmät edellyttävät ainakin implisiittisesti reiluusmetriikoiden käyttämistä, sillä oikomismenetelmän tehokkuutta arvioidaan usein yhtä tai useaa mittaria ja sitä heijastavaa evaluatiivista standardia vasten (ml. mallin osumatarkkuuden). Soveltuvan mittarin ja optimointikriteeristön valinta on erityisen tärkeää vinoumien teknisessä oikomisessa ja mallin parantamisessa.

Reiluusmetriikkoja ja oikomismenetelmiä sisältäviä sovelluspaketteja on tarjolla eri alustoilla, muun muassa avoimena lähdekoodina. Esimerkkejä ovat IBM:n *AI Fairness 360*[147] ja Aequitasin *Bias and Fairness Audit*[148], mutta menetelmiä pystytään hyödyntämään ilmankin näitä sovelluspaketteja.

143 Ks. Mehrabi ym., 2021.
144 Ks. esim. Institut Montaigne, 2020, 61; Centre for Data Ethics and Innovation, 2020, 69.
145 Ks. Mehrabi ym., 2021.
146 Ibid.
147 AI Fairness 360. https://aif360.mybluemix.net/.
148 Aequitas Bias and Fairness Audit. http://aequitas.dssg.io/.

2.6.2.3 Selitysmenetelmät

Tekoälysovellusten läpinäkyvyyden, selitettävyyden ja ymmärrettävyyden lisäämisen katsottiin osassa aineistoa mahdollisesti ehkäisevän keskeisiä negatiivisia yhdenvertaisuusvaikutuksia sekä läpinäkyvyyden puutteeseen liittyviä ongelmia, joita on käsitelty yllä. Mikäli esimerkiksi päätöksentekoprosessin selitettävyyden puute voi johtaa päätelmään syrjinnästä[149], on selitettävyysmetodien käyttäminen olennaista hyvän käytännön ja menettelyn lisäksi myös yhdenvertaisen kohtelun näkökulmasta.

Selitysmenetelmät eivät usein itsessään vaikuta mallin ennusteisiin tai niiden jakaumiin oikomalla esimerkiksi vinoumia vaikkakin tutkimuskirjallisuudessa on huomautettu, että joidenkin menetelmien hyödyntäminen voi heikentää mallin tarkkuutta[150]. Selitettävyys voi kuitenkin parhaimmillaan parantaa sovellusten turvallista ja kompetenttia käyttöä, helpottaa mahdollisten vinoumien tunnistamista ja antaa tietoa tiedonkäsittelyn luonteesta tai päätöksenteon perusteista[151], täten parantaen päätöksenteon kohteiden pääsyä oikeuksiinsa (mukaan lukien syrjimättömyyteen). Selitysmenetelmiä voidaan esimerkiksi ”käyttää virheenkorjaukseen ja mallien käyttäytymisen analysoimiseen niiden antaessa vääriä tai odottamattomia ennusteita”[152] etsimällä esimerkiksi muuttujia, joiden arvojen muuttaminen johtaa ennusteiden tarkkuuden tai luottamustason vaihteluun. Lisäksi esimerkiksi kontrafaktuaalisten selitysmenetelmien avulla voidaan tuottaa tietoa siitä, onko kielletty syrjintäperuste vaikuttanut järjestelmän tuottamaan tulosteeseen[153]. Tässä mielessä selitettävyys ja reiluus toimivat parhaiten yhdessä teknisinä ratkaisumalleina syrjintäriskien löytämisessä ja ehkäisemisessä.

Tekoälyjärjestelmien tulosteiden ja prosessointilogiikan ymmärrettävyyden parantamiseksi on kehitetty useita teknisiä menetelmiä, joita voidaan jaotella eri tavoin[154]. *Mallispesifit menetelmät* soveltuvat yksittäisten koneoppimismenetelmien ja algoritmityyppien selitettävyyden lisäämiseen, kun taas mallien suhteen agnostiset menetelmät soveltuvat kaikkiin koneoppimismenetelmiin ja algoritmityyppeihin. *Lokaaleilla menetelmillä* tuotetaan selityksiä yksittäisille tulosteille, kun taas *globaaleilla menetelmillä* pyritään lisäämään mallin ymmärrettävyyttä. *Ante hoc* -selitettävyys viittaa malleihin, jotka ovat rakenteellisesti ymmärrettäviä tai tulkittavia, kun taas *post hoc* -selitettävyys viittaa tulosteiden tai mallin selitykseen ns. jälkikäteen. Menetelmiä voidaan jaotella myös sen mukaan,

149 Allen QC & Masters, 2020, 47.
150 Ks. Rudin, 2019.
151 Ks. esim. Du, Liu & Hu, 2019; Linardatos, Papastefanopoulos & Kotsiantis, 2021.
152 Du, Liu & Hu, 2019, 75.
153 Wachter, Mittelstadt & Russell, 2017.
154 Ks. esim. Du, Liu & Hu, 2019.

minkälaisiin datatyyppeihin ne soveltuvat (esim. tabulaarinen data, teksti, graafit, kuvat). Erilaisia ratkaisuja ja selitettävyystyökaluja on saatavilla tekoälyjärjestelmien kehittäjille tai operoijille ilmaiseksi tai maksua vastaan[155].

Selitysmenetelmillä on kuitenkin omat vahvuutensa ja heikkoutensa (ks. alla), jotka saattavat vaikuttaa niiden sovellettavuuteen syrjivien vaikutusten ehkäisemisessä.

2.6.3 Päätöksentekoprosesseihin liittyvät organisaatiotason keinot

Aineistossa tunnistettiin myös useita laajempia keinoja, joilla pyrittiin tunnistamaan ja ehkäisemään syrjiviä vaikutuksia. Nämä keinot kohdistuvat esimerkiksi tekoälyjärjestelmien tutkimus-, kehitys- ja käyttöprosesseihin, organisaatiotason toimintaan sekä lainsäädännöllisiin muutoksiin.

2.6.3.1 Riski- ja vaikutustenarviointiprosessit

Rekrytointijärjestelmiä tarkastelevassa Institute for the Future of Workin raportissa huomautetaan, että "mikäli auditointityökalut eivät keskity merkityksellisiin tasa-arvokysymyksiin, ja mikäli niitä ei integroida osaksi laajempaa tasa-arvovaikutustenarviointia, niiden hyöty tasa-arvon edistämisessä on rajallinen"[156]. Osa raporteista suositteleekin teknisen auditoinnin integroimista osaksi laajempia riski- ja vaikutustenarviointeja, kuten algoritmien vaikutustenarviointia (engl. *algorithmic impact assessment*, "AIA"). Kyseiset prosessit sisältäisivät esimerkiksi perustavampaa tarkastelua, kuten perusoikeudellisten vaikutusten ja lainmukaisuuden arviointia, sekä sidosryhmien ja tekoälyjärjestelmän vaikutuksenalaan kuuluvien ryhmien konsultointia[157]. Institute for the Future of Work on ehdottanut laajempaa tasa-arvovaikutusten arviointikehikkoa tekoälyjärjestelmille[158], joka palvelisi työnantajien suorittamia vapaaehtoisia vaikutuksenarviointeja. Euroopan unionin perusoikeusviraston raportissa ehdotetaan kansainvälisten ja kansalaisjärjestöjen sekä tutkijoiden suosituksia mukaillen pakollista ihmisoikeusvaikutustenarviointia tekoälyjärjestelmille, joka toteutettaisiin ennen järjestelmän käyttöönottoa[159].

155 Ks. esim. SHAP, LIME, Alibi, Random Forest Explainer (RFEX2.0), TreeInterpreter, IBM:n AI Explainability 360, Googlen Explainable AI, Microsoftin InterpretML, DeepLift, ja DeepExplain.

156 Institute for the Future of Work, 2020, 3.

157 The Greenlining Institute, 2020; ks. myös Calvo, Peters & Cave, 2020; Kaminski & Malgieri, 2020.

158 Institute for the Future of Work, 2020.

159 Euroopan unionin perusoikeusvirasto, 2020, 8.

Suomen kontekstissa on huomattava, että viranomaisilla, koulutuksentarjoajilla ja työnantajilla voidaan olettaa olevan yhdenvertaisuuslain pykälien 5-7 perusteella velvollisuus arvioida tekoälyjärjestelmän käyttöönoton yhdenvertaisuusvaikutuksia. Suomen tietosuojavaltuutettu on myös painottanut, että rekisterinpitäjien on aktiivisin toimin pidettävä huolta, että käsiteltävät tiedot ja käytetyt algoritmit eivät johda syrjivään tietojenkäsittelyyn[160]. Suomen kontekstissa yhdenvertaisuusvaikutustenarvioinnin voidaankin katsoa olevan olennainen osa yhdenvertaisuuslain asettamia velvollisuuksia mutta myös GDPR:n asettamia vaatimuksia, kuten tietosuojavaikutustenarviointia (DPIA).

2.6.3.2 Dokumentaatio

Aineiston raporteissa paikoin esitetään tai käsitellään tekoälyjärjestelmiä koskevia pakollisia dokumentaatiovaatimuksia ja niitä koskevia läpinäkyvyyttä tukevia toimenpiteitä. Yllä käsitellyt dokumentaatiomenetelmät, kuten tutkimuskirjallisuudessa standardimenetelmiksi esitetyt käytettyä dataa koskevat menetelmät, mutta myös esimerkiksi malleja ja algoritmien toimintaa koskeva dokumentaatio[161], olisivat sovellettavia tässä suhteessa. Suomalainen yritys Saidot on myös kehittänyt alustaa mallien metadatan hallinnoimiseen, jonka avulla pyritään helpottamaan organisaatioiden ja ekosysteemien läpinäkyvää tekoälyjärjestelmien rekisterinpitoa ja vastuullista hallinnoimista[162].

Datan, mallien ja algoritmien dokumentaation nähdään palvelevan niin (i) sovellusten vastuullista käyttöä kuin (ii) syrjinnän kohteeksi joutuneiden pääsyä oikeuksiinsa. Dokumentoitua tietoa voitaisiin jakaa useita toimijoita yhdistävissä arvoketjuissa esimerkiksi opetusdataa keränneen ja sitä hyödyntävän toimijoiden välillä, mutta myös julkisesti tai rajatusti asianmukaisten osapuolien välillä läpinäkyvyyden lisäämiseksi. Syrjintäepäilyjen tapauksissa dokumentoituun tietoon olisi pääsy asianmukaisilla tahoilla, kuten tasa-arvo- ja yhdenvertaisuusvaltuutetulla tai muilla auditointia suorittavilla kolmansilla osapuolilla, ja tätä pääsyä voitaisiin säännellä lailla näissä tapauksissa[163].

2.6.3.3 Heuristiikat

Osassa raportteja esitettiin syrjintäriskien arvioinnin avuksi heuristiikkoja, kuten tarkistuslistoja (engl. *checklist*), kehotteita ja kysymyspatteristoja sekä itsearviointityökaluja. Esimerkiksi Equinetin raportissa esitellään tarkistuslista, joka auttaa tuote- ja

160 Automaattinen päätöksenteko ja profilointi, Tietosuojavaltuutetun toimisto. https://tietosuoja.fi/automaattinen-paatoksenteko-profilointi.

161 Mitchell ym., 2019.

162 https://www.saidot.ai/.

163 Orwat, 2020; World Economic Forum, 2018.

palvelukehittäjiä arvioimaan syrjintään liittyviä kysymyksiä sovelluksensa kontekstissa[164], ja Euroopan unionin perusoikeusvirasto suosittelee, että EU:n jäsenvaltioissa hyödynnetään olemassa olevia heuristiikkoja itsesääntelyn tukena[165].

2.6.3.4 Vaihtoehtoisten menettelytapojen käyttöönotto

Yhdenvertaisuusperiaatteen toteutuminen ei typisty eri ihmisten identtiseen kohteluun vaan edellyttää myös esimerkiksi henkilöiden erityistarpeiden tunnistamista. Negatiivisten yhdenvertaisuusvaikutusten ehkäiseminen, kuten aineistosta nousee esiin, voi tässä mielessä edellyttää myös vaihtoehtoisten menettelytapojen mahdollistamista ja/tai käyttöönottoa tekoälyjärjestelmän käytön ohella. Esimerkiksi rekrytointisovelluksia käsittelevässä raportissa suositeltiin, että työnantajat käyttäisivät "vaihtoehtoisia testaus- ja seulontamenetelmiä huomioidakseen vammaiset hakijat tavalla, joka ei estä heidän mahdollisuuksiaan osoittaa taitojaan"[166]. Päätöksentekoprosessien yhdenvertaisuusvaikutuksia voidaan paikoin parantaa käyttämällä eri tilastollisia malleja eri ihmisryhmien kohdalla[167]. On kuitenkin vielä epäselvää, onko tämä yhteensopivaa yhdenvertaisuuslain kanssa, ja jos on, milloin[168].

2.6.3.5 Riitauttamisen mahdollistaminen ja hyvitys

Yhdenvertaisuuslain yksi tarkoitus on tehostaa syrjinnän kohteiksi joutuneiden oikeusturvaa (YVL 1 §). Mahdollisuus tekoälyavusteisten päätöksentekoprosessien riitauttamiseen nähdäänkin suuressa osassa aineistoa olennaiseksi, jotta voidaan taata päätöksenteon kohteiden pääsy oikeuksiinsa, mukaan lukien syrjimättömyyteen[169].

Tekoälyn selitettävyys on olennaista riitauttamisen ja hyvityksen saamisen mahdollisuudelle, sillä yksilöiden täytyy saada tietää, millä perusteilla päätöksiä tehdään, jotta he voivat riitauttaa niitä ja kerätä mahdollista todistusaineistoa syrjiviä vaikutuksia koskien. Datasubjekteille GDPR:ssä asetetut oikeudet palvelevat osaltaan riitauttamisen mahdollisuutta. Julkisten ja kollektiivisten täytäntöönpanomekanismien, kuten EU:n

164 Equinet, 2020.

165 Euroopan unionin perusoikeusvirasto, 2020, 8.

166 Center for Democracy & Technology, 2020, 16.

167 Australian ihmisoikeuskomissio, 2020, 50–51.

168 Hoch ym. (2021) mukaan EU:n yhdenvertaisuus- ja tietosuojalainsäädäntö sallisivat esimerkiksi erillisten kynnysarvojen käyttämisen tekoälyavusteisessa päätöksentekoprosessissa eri ryhmille, kuten miehille ja naisille. Tällainen käytäntö katsottaisiin heidän mukaansa positiiviseksi erityiskohteluksi, mikäli se tasapainottaisi mallin sensitiivisyyden ryhmien välillä.

169 Center for Democracy & Technology, 2020; Gerards & Xenidis, 2021; Euroopan digitaaliset oikeudet, 2021.

tasa-arvoelinten, on myös visioitu helpottavan näiden mahdollisuuksien toteuttamisessa[170]. Lisäksi joissakin raporteissa nostetaan esille joukkokanteiden nostamisen mahdollistaminen algoritmisen syrjinnän tapauksissa[171].

2.7 Haasteita syrjintäriskien hallinnalle

Tässä osiossa tarkastellaan tekoälyjärjestelmien syrjintäriskien arvioimisen ja ehkäisemisen keinoihin liittyviä haasteita, joita on tunnistettu niin tarkastellussa aineistossa kuin tutkimuskirjallisuudessakin. Tämä osio vastaa kysymykseen: *mitä haasteita liittyy syrjintäriskien arviointia ja ehkäisemistä tukevien menetelmien käyttöön?* Näitä haasteita tarkastelemalla voidaan tuottaa tietoa siitä, mitä tunnistettujen ratkaisujen asianmukainen ja vaikuttava käyttö tosiasiallisissa suunnittelu- ja käyttökonteksteissa edellyttää.

- **Kartoitus antaa syytä epäillä, että algoritmien auditointi ja teknisten oikomismenetelmien käyttö olisivat itsessään riittäviä syrjinnän ehkäisemisen keinoina ja yleisesti yhdenvertaisuuslain tarkoituksien toteuttamisen välineinä.** Mikäli tekniset oikomismenetelmät toimivat itsesääntelyn apuvälineinä ja ilman asianmukaisia standardeja ja vastuumekanismeja, tämä saattaa johtaa väärinkäyttöön sekä erheellisiin tai jopa vahingollisiin interventioihin. Tekniset menetelmät voivat asianmukaisesti käytettynä kuitenkin edesauttaa yhdenvertaisuuden toteutumista.
- **Koneoppimismallien arviointiin käytettävien mittareiden sekä dataan, malliin tai ennusteisiin kohdistuvien oikomismenetelmien käyttäminen ei ole riittävää yhdenvertaisuuslaissa asetettujen velvoitteiden täyttämisen näkökulmasta.** Yhdenvertaisuuslain tavoitteet saattavat olla ristiriidassa reiluusmetriikoiden ja oikomismenetelmien perustavan oletuksen kanssa, jonka mukaan yhdenvertaisuus voitaisiin typistää yhdeksi universaalisti sovellettavaksi säännöksi (vrt. esim. kohtuulliset mukautukset).
- **Standardilähestymistavat saattavat keskittää valtaa yksityisen sektorin tekoälykehittäjille ja epäonnistua edistämään yhdenvertaisuutta dynaamisissa sosioteknisissä ekosysteemeissä.** Teknisellä tasolla toteutetut, koordinoimattomat ja lyhytjänteiset interventiot yhdenvertaisuuden ja tasa-arvon edistämiseksi saattavat tuottaa vastakkaisia vaikutuksia pitkässä juoksussa. Yksittäisiä toimijoita tulee ohjeistaa ja kannustaa suorittamaan syrjintäriskien arviointia ja ehkäisemistä oikomisen

170 Gerards & Xenidis, 2021, 11.

171 Ks. esim. Orwat, 2020.

keinoin, mutta tämän ei tule antaa sivuuttaa ”radikaalimpia keinoja vastata laajoihin, rakenteellisiin haittavaikutuksiin, joita tekoäly synnyttää”[172].

- **Teknologiateollisuuden ja yksittäisten organisaatioiden tasolla esiintyvä diversiteetin puute muodostuu esteeksi yhdenvertaisuusnäkökulmat huomioivalle tekoälyn suunnittelulle sekä asianmukaiselle ja vaikuttavalle syrjivien vaikutusten tunnistamiselle ja ehkäisemiselle.** Monimuotoisuuden ja eri väestöryhmien edustuksen puutteen teknologiateollisuudessa ja teknologiayrityksissä nähdään olevan yhteydessä muun muassa aliedustettujen ryhmien vaatimusten, intressien ja oikeuksien sivuuttamiseen teknologiasuunnittelussa sekä merkityksellisten näkökulmien laiminlyömiseen yhdenvertaisuus- ja tasa-arvovaikutusten arvioinnissa käyttöönotossa.
- **Käytännön kehittämistyön ja organisaatioiden toiminnan tasolla esiintyy epätietoisuutta ja epäselvyyttä lainmukaisuusvaatimuksista, kuten yhdenvertaisuusvaikutusten arvioimisen sekä ehkäisemisen vaatimuksista ja mahdollisuuksista.** Toimijoilta saattaa puuttua tietoa, jota vaaditaan asianmukaiseen riskinarviointiin ja -hallintaan tekoälysovelluksien kontekstissa. Sen lisäksi, että organisaatioilta uupuu tieto siitä, ”kuinka oikeat menetelmät tulisi valita” esimerkiksi mallien arvioimisen suhteen, niiltä saattaa myös puuttua näkemys siitä, kuinka arviointi ”tulisi rakentaa osaksi kehitys- ja operatiivisia prosesseja”[173]. Vastuiden epäselvyys muodostuu ongelmaksi monimutkaisissa ja useita toimijoita yhdistävissä arvoketjuissa.

2.7.1 Teknisiin menetelmiin liittyvät haasteet ja ongelmat

Tekoälysovellusten syrjivien vaikutusten arviointiin ja ehkäisemiseen tarkoitettujen teknisten menetelmien valintaan ja hyödyntämiseen liittyy lukuisia haasteita. Osa esittelemistämme huomioista nousee esiin aineistoon sisältyneistä raporteista, mutta täydennämme haasteiden kokonaiskuvaa myös tutkimuskirjallisuudesta esiin nousevilla teemoilla.

172 Euroopan digitaaliset oikeudet, 2021, 9.

173 Centre for Data Ethics and Innovation, 2020a, 9.

2.7.1.1 Voiko syrjiviä vaikutuksia todentaa mallin tasolla?

Reiluusmetriikoiden ja tilastollisten testien perustava oletus on, että syrjiviä vaikutuksia voitaisiin todentaa tarkastelemalla järjestelmää mallin (ts. datan, algoritmien ja tulosteiden) tasolla. On kuitenkin jokseenkin avoin kysymys, voiko tällä tavalla tuotettu kvantitatiivinen tai tilastollinen informaatio antaa todisteita mahdollisesta syrjinnästä. Voidaankin oikeastaan erottaa kolme erillistä kysymystä

I. Tuleeko mallin oppima tilastollinen korrelaatio, joka on olennaisesti ryhmä- tai populaatiotason päätelmä, ymmärtää henkilöä (tai jotakuta toista) koskevana oletuksena?
II. Kuinka vahva korrelaatio kielletyn syrjintäperusteen ja tietyn muuttujan välillä tulee olla, jotta kyse on syrjinnästä?
III. Minkä kanssa kielletyn syrjintäperusteen pitäisi korreloida, jotta kyse olisi syrjinnästä – esim. tulosteen (vrt. päätöksen) vaiko sen tarkkuuden?

Käsittelemme näitä kysymyksiä seuraavaksi.

Tilastollinen diskriminaatio ja henkilöä koskeva oletus. Yhdenvertaisuuslain mukaan ”[s]yrjintä on kielletty riippumatta siitä, perustuuko se henkilöä itseään vai jotakuta toista koskevaan tosiseikkaan tai oletukseen” (YVL 8 §). Tekoälyjärjestelmät tai rajatummin koneoppimisjärjestelmät, joiden avulla tehdään ihmistä koskevia päätöksiä, suorittavat nk. tilastollista diskriminaatiota eli kohtelun erottelua tilastolliseen dataan pohjautuen[174]. Mikäli henkilöä koskeva oletus voidaan samaistaa tekoälyn hyödyntämän mallin ilmaisemaan todennäköisyyslaskelmaan, joka on välttämättä opetusdatan pohjalta päätelty tilastollinen yleistys, voitaisiin mallia tarkastelemalla periaatteessa saada etukäteistä tietoa ainakin syrjinnän riskistä.

Toisaalta näyttäisi siltä, että ”tosiasiallisten lopputulemien tarkasteleminen [mallin tulosteiden sijaan] on merkityksellisempää, kun on kyse järjestelmien aiheuttamasta mahdollisesta syrjinnästä.”[175] Päätelmä syrjinnästä edellyttää oikeudellisten vaikutusten aiheuttamista, joten tässä suhteessa tekoälyjärjestelmän käyttökontekstilla on olennaisesti merkitystä. Tutkimuskirjallisuudessa on todettu, että mahdollisesti syrjivä tai muutoin epäreilu

174 Ks. Binns, 2018. Määritelmällisesti tilastollinen diskriminaatio voi olla niin välitöntä kuin välillistäkin. Se on välitöntä, kun kielletyn syrjintäperusteen pohjalta tehdään päätelmä tai ennuste kohdemuuttujan arvon esiintymisestä henkilön kohdalla (esim. raskaana olevan ihmisen päätellään ottavan äitiyslomaa). Se on välillistä, kun päätelmää tai ennustetta ei tehdä kielletyn syrjintäperusteen pohjalta, mutta päätelmä johtaa erotteluun sen perusteella (hypoteettisesti esim. pidettyjen vapaiden määrä vaikuttaa työssä ylenemispäätökseen naisten pitäessä keskimäärin enemmän vapaita).

175 Euroopan digitaaliset oikeudet 2021, sivu 24.

kohtelu todetaankin usein vasta, kun sovellus on käytössä[176]. Vahvatkaan indikaatiot mahdollisesta syrjinnästä mallin tasolla eivät tarkoita, että riskit realisoituisivat, mikäli oikeudellisia vaikutuksia ei *de facto* synny tekoälysovelluksen käytössä. Ennakoivaa vaikutuksenarviointia voidaan toki tehdä, mutta vaikutuksenarviointi näyttäisi olevan oikeudellisessa mielessä tarpeellista vain silloin, kun oikeudellisia vaikutuksia voidaan olettaa syntyvän. Huomautettakoon, että vaikutuksenarviointi voi olla tarpeellista moraalisista syistä muissakin tapauksissa.

Oikeudellisten vaikutusten syntyessäkin on kuitenkin epäselvää, vaikuttaako mahdollisen syrjinnän kohteeksi joutunutta yksilöä koskevan syötedatan tarkkuus (vrt. ”henkilöä itseä koskeva tosiseikka”) päätelmään *prima facie* syrjinnästä, ottaen huomioon, että päätelmä syrjinnästä voidaan tehdä, kun syrjivä päätös perustuu *oletukseen*, joka voi koskea muuta kuin henkilöä itseänsä (vrt. yllä).

Korvannaismuuttujat ja korrelaatiot kiellettyjen syrjintäperusteiden kanssa. Kuinka vahva suhde kielletyn syrjintäperusteen ja mallin ennusteiden välillä pitäisi olla, jotta voidaan tehdä päätelmä *prima facie* (välittömästä tai) välillisestä syrjinnästä? Eräässä raportissa esitetään, että

> [a]lgoritmisen syrjinnän esittämä haaste on löytää, milloin korvannaismuuttujien (yhdistelmien) ja affiniteettien voidaan tosiasiassa katsoa olevan niin päällekkäisiä vastaavan kielletyn syrjintäperusteen kanssa, että ne voitaisiin ymmärtää samaksi asiaksi. Tässä suhteessa saattaa nousta kysymys, vaaditaanko todella (lähes) 100 % päällekkäisyys vai riittääkö, että osoitetaan tilastollisesti tietyn muuttujan (tai muuttujien yhdistelmän) olevan 90 % tai 80 % yhteneväinen tietyn kielletyn syrjintäperusteen kanssa.[177]

Keskeinen kysymys reiluusmetriikoiden käyttämisen näkökulmasta on täten asianmukaisen kynnysarvon asettaminen sille, kuinka vahvasti mallin muuttujat saavat korreloida kielletyn syrjintäperusteen kanssa. Yhdysvaltain oikeudellisessa kontekstissa nk. 4/5-sääntö antaa suuntaa tässä suhteessa, mutta samankaltaista heuristiikkaa ei ole käytössä Euroopassa.

176 Holstein ym., 2019.

177 Gerards & Xenidis, 2021, 64.

Miten yhdenvertaisuusvaikutuksia tulisi arvioida reiluusmetriikoiden avulla? Mikäli syrjiviä vaikutuksia voidaan arvioida kvantitatiivisesti mallin tasolla, nousee kysymys: mikä tai mitkä mittarit ovat soveltuvia tähän tarkoitukseen ja missä konteksteissa? Tulisiko arvioijan tarkastella esimerkiksi ennusteiden (vrt. päätösten) vaiko virheellisten ennusteiden (vrt. erheelliset päätökset) jakautumista ryhmien välillä?

Aineistosta ilmenee epätietoisuus ja erimielisyys näiden kysymysten suhteen. Tämä johtunee osittain siitä, että ”yhdenvertaisuuslainsäädännöllä ei ole juurikaan sanottavaa tekoälysovellusten epätarkoista ennusteista (vääristä positiivisista ja vääristä negatiivisista)”[178]. Kuitenkin esimerkiksi kasvojentunnistusteknologian käytössä väärien positiivisten määrä on tunnistettu merkittäväksi mittariksi tässä mielessä[179] ja Alankomaiden SyRI-tapauksen yhteydessäkin ennusteiden todenpitävyys nähtiin merkityksellisenä syrjivien vaikutusten arvioimisessa[180]. Keskeinen ongelma reiluusmetriikoiden käyttämisen suhteen (erityisesti itsesääntelyn apuvälineenä) onkin, että soveltuvan mittarin valinta on kontekstisidonnaista ja hankalaa, eikä valinnalle ei ole olemassa standardikäytäntöä. Tämä puolestaan juontaa osaltaan siitä, että oikeuskäytännössä syrjintätapausten arvioiminen pohjautuu tapauskohtaisesti kunkin tapauksen tosiseikkoihin ja olosuhteisiin, täten horjuttaen oletusta yhtenäisestä ja kaavamaisen selkeästä lähestymistavasta mahdollisten syrjivien vaikutusten arviointiin. Pelkästään niissäkin tapauksissa, joissa tekoälyjärjestelmillä tehdään ihmisiä koskevia päätöksiä ja joissa niillä voi olla merkittäviä oikeudellisia vaikutuksia, riskit eroavat esimerkiksi sektoreiden välillä.

Soveltuvan mittarin valinta osoittautuu haasteelliseksi osittain myös mittareiden heijastamien evaluatiivisten standardien eroista johtuen. Reiluusmetriikoiden merkityksellisiä eroja moraalifilosofisessa, poliittisessa ja oikeudellisessa mielessä on tuotu esille tutkimuskirjallisuudessa[181]. Yllä mainittu ero yksilö- ja ryhmäreiluusmittareiden välillä on yksi tällainen merkittävä ero. Lisäksi eurooppalaiseen yhdenvertaisuuslainsäädäntöön keskittyvässä tutkimuksessa on huomautettu, että niin kutsutut ”vinoumia korjaavat” mittarit (engl. *bias-transforming metrics*) näyttäisivät olevan yhteneviä kyseisen lainsäädännön tavoitteiden ja hengen kanssa[182]. Tällaiset mittarit (mm. *tilastollinen pariteetti* ja *ehdollinen tilastollinen pariteetti*) eivät ota huomioon opetusdatan pohjatotuutta ja ne tulkitsevat täydellisen tarkankin mallin vinoutuneeksi, mikäli se heijastaa rakenteellista epätasa-arvoa lähtötilanteessa. Täten myös mallin parantaminen vinoumia korjaavien mittareiden valossa

178 Borgesius, 2018, 36.

179 Allen QC & Masters, 2020, 43.

180 Haagin käräjäoikeus C-09-550982-HA ZA 18-388, kohta 6.91.

181 Ks. esim. Narayanan, 2018; Wachter, Mittelstadt & Russell, 2021; Lee, Floridi & Singh, 2021; Sahlgren, 2021.

182 Wachter, Mittelstadt & Russell, 2021.

palvelee tosiasiallisen tasa-arvon (engl. *substantive equality*) edistämistä. ”Vinoumia säilyttävät” mittarit (esim. *virhetasojen yhtäläisyys*) puolestaan tarkastelevat ennusteiden tarkkuuden ja virhetasojen jakautumista vertailtavien ryhmien välillä. Jälkimmäisten mittarien on tulkittu edustavan nk. muodollista tasa-arvon käsitystä (engl. *formal equality*).[183]

2.7.1.2 Reiluustavoitteiden yhteensopimattomuus

Yksi keskeisimmistä reiluusmetriikoiden käyttöön liittyvistä haasteista koskee nk. ”valintatilanteita”, ”vaihtokauppoja” tai kompromisseja (engl. *trade-offs*) ja eri reiluustavoitteiden yhteensopimattomuutta. Tutkimuskirjallisuudessa on tunnistettu, että ennustemallien optimoiminen jonkinlaisen reiluuden suhteen merkitsee useimmiten tinkimistä mallin tarkkuudesta ja lähes kaikissa tapauksissa erilaisia reiluusmääritelmiä ei voida saavuttaa yhtäaikaisesti. Esimerkiksi yksilö- ja ryhmäreiluutta ei voida useimmiten saavuttaa samanaikaisesti johtuen rakenteellisista eroista kohdemuuttujan esiintyvyydessä populaation sisällä[184]. Sama pätee kalibraation ja yhtäläisten virhetasojen kohdalla[185].

Reiluuteen liittyviä mahdottomuusteoreemoja käsittelevässä kirjallisuudessa on kuitenkin todettu, että matemaattisella tasolla esiintyvät konfliktit johtuvat rakenteellisen eriarvoisuuden ja stratifikaation olemassaolosta[186]. Algoritmien reiluusmääritelmiä pystytään periaatteessa täydellisesti saavuttamaan yhtäaikaisesti vasta, kun tosiasialliset esteet yhdenvertaisuudelle ja tasa-arvolle on poistettu. Standardikäytäntöjen ja tekniikoiden puuttuessa tekoälysovelluksia kehittävät ja käyttävät toimijat joutuvat kuitenkin käytännössä valitsemaan sopivan määritelmän ja tekemään parhaaksi näkemänsä ratkaisun mallin tarkkuuden ja reiluuden suhteen[187]. On huomattava, että reiluusmetriikat voivat auttaa kuitenkin löytämään matemaattisesti optimaalisia tapoja tasapainottaa esimerkiksi mallien tarkkuutta ja pariteettia kompromisseja vaativien tilanteiden ilmetessä. Käytännössä esimerkiksi malli, joka on Pareto-optimaalinen eri mittareiden suhteen[188] on lähtökohtaisesti parempi kuin malli, jossa reiluutta ei ole otettu lainkaan huomioon. Päätösten tekeminen mallien reiluutta ja yleistä tarkkuutta koskien edellyttää kuitenkin, että mallin käytöllä on hyväksyttävät tavoitteet ja että valitut metriikat ovat soveltuvia käyttökontekstiin, ja vaatii oikeudellisten perusteiden ja sosiaalisen kontekstin huomioonottamista.

183 Ibid.

184 Dwork ym., 2012; Binns, 2020.

185 Chouldechova, 2017; Kleinberg ym., 2017.

186 Binns, 2020; Green, 2021a.

187 Centre for Data Ethics and Innovation, 2020a, 76.

188 Pareto-optimaalisuus tarkoittaa tässä tapauksessa sitä, että mallia ei voida enää parantaa yleisen tarkkuuden tai käytettyjen reiluuden mittareiden suhteen samanaikaisesti huonontamatta mallia jossakin näistä suhteista.

2.7.1.3 Reiluusmetriikoiden väärinkäyttö ja harhaanjohtaminen

Itsesääntelykeinona reiluusmetriikoiden käyttöön liittyy riski, että arviointia suorittava toimija voi esittää harhaanjohtavaa informaatiota mallista tai manipuloida koeasetelmaa siten, että mittarit osoittavat mallin ”reiluksi” sen ollessa syrjivä. Euroopan digitaaliset oikeudet -edunvalvontaryhmän raportissa huomautetaan, että “palveluntarjoajat voivat helposti suorittaa harhaanjohtavia toimia järjestelmää auditoidessaan saadakseen sen tulosteet näyttämään tasapuolisilta”[189]. Toisaalta algoritmien auditoinnin ulkoinenkin sääntely ja arviointi on myös haastavaa tekoälysovellusten käyttötapausten monipuolisuudesta johtuen – auditoinnin edellyttämä arviointidata saattaa olla esimerkiksi olematonta tai hankalasti saatavilla[190].

2.7.1.4 Vertailuluokkien valitseminen

Oikomismenetelmien ongelmia käsittelevässä raportissa käsiteltiin kysymystä siitä, miten vertailuluokat tulisi ja olisi mahdollista valita, jotta yhdenvertaisuusvaikutusten arviointi olisi onnistunutta? (Yllä mainittu haaste koskien intersektionaalisiin ja sosiaalisesti vähemmän merkityksellisiin ominaisuuksiin tai ominaisuuksien joukkoon perustuvaan systemaattiseen erilaiseen kohteluun liittyen on olennainen tässä kontekstissa.) Yksinkertainen vastaus kysymykseen on, että vertailtavien luokkien tulisi kattaa yhdenvertaisuuslaissa nimetyt kielletyt syrjintäperusteet (YVL 8 §). Tämä on kuitenkin käytännöllisesti haasteellista useasta syystä.

Ensimmäinen syy on, että yhdenvertaisuusvaikutusten arvioiminen kiellettyjen syrjintäperusteiden suhteen voi olla haastavaa sensitiivisen datan saatavuudesta johtuen. Esimerkiksi julkisen sektorin toimijoilla saattaa olla yleisen tietosuoja-asetuksen (GDPR) puitteissa pääsy sensitiiviseen dataan, kuten sukupuoleen tai ikään, mutta tämä ei välttämättä päde yksityisen sektorin toimijoiden kohdalla (esim. rekrytoinnissa). Datan saatavuuden ongelmat on nostettu esille tutkimuskirjallisuudessa, minkä lisäksi sensitiivisen datan keräämiseen ja säilömiseen liittyviä yksityisyyden- ja tietosuojaa koskevia riskejä on myös korostettu[191].

Tekoälysäädöksen odotetaan kuitenkin osaltaan selventävän ja edesauttavan sensitiivisen datan käyttömahdollisuuksia tässä suhteessa, yksityisyyttä ja tietosuojaa koskevat riskit huomioon ottaen[192]. Ehdotuksessa sanotaan:

---

189 Euroopan digitaaliset oikeudet, 2021, 76.

190 Ibid.

191 Ks. esim. Holstein ym., 2019.

192 TEKOÄLYÄ KOSKEVISTA YHDENMUKAISTETUISTA SÄÄNNÖISTÄ (TEKOÄLYSÄÄDÖS) JA TIETTYJEN UNIONIN SÄÄDÖSTEN MUUTTAMISESTA, Artikla 10, kohta 5.

”Jotta voitaisiin suojella muita tekoälyjärjestelmissä esiintyvästä vinoutumisesta mahdollisesti johtuvalta syrjinnältä, tarjoajien olisi voitava käsitellä myös erityisiä henkilötietoryhmiä tärkeän yleisen edun nimissä, jotta voidaan varmistaa suuririskisiin tekoälyjärjestelmiin liittyvien vinoutumien seuranta, havaitseminen ja korjaaminen”[193].

Silloinkin, kun auditoijalla on pääsy kiellettyjä syrjintäperusteita koskevaan dataan, merkityksellisten vertailuluokkien valinta *a priori* tai *ex ante* (ennen epäsuotuisten vaikutusten ilmenemistä) on kuitenkin haastavaa. Esimerkiksi yksilöreiluuden arvioimisessa joudutaan formalisoimaan samankaltaisuuden mitta matemaattisella tasolla ja täten määrittämään, mitkä ominaisuudet ovat merkityksellisiä yksilöiden ”samankaltaisuuden” kannalta. Ottaen huomioon, että mahdollinen syrjintä saattaa seurata välillisesti korvannaismuuttujien vaikutuksen kautta, kehittäjien voi olla haastava arvioida, mitkä kaikki muuttujat (vrt. attribuutit tai ryhmäjäsenyydet) tulee ottaa huomioon samankaltaisuutta arvioitaessa, jotta saataisiin asianmukainen arvio yhdenvertaisesta kohtelusta.[194] Samat haasteet koskevat ryhmäreiluuden arvioimista, sillä tässäkin kontekstissa joudutaan priorisoimaan vertailuja tiettyjen ryhmien välillä.

Vertailuluokkien valintaa syrjintäriskien arvioimisessa hankaloittaakin se, että tosiasiallisesti ”vertailuluokat ovat tapauskohtaisia ja ne määritellään [oikeuskäytännössä] vasten riitautettua käytäntöä tai sääntöä” sekä tapauksen faktuaalista perustaa[195]. Mikäli on odotettavissa, että tiettyyn populaatioon tai käyttötarkoitukseen käytettävä järjestelmä todennäköisesti suosii tiettyä ryhmää tai kohtelee tätä ryhmää epäsuotuisasti, järjestelmää tulisi oletettavasti testata kyseisen ja vertailukelpoisten ryhmien kohdalla vinoumien varalta[196]. Merkityksellisiä vertailuluokkia saattaa olla kuitenkin useita, mikä tekee niiden valinnasta käytännöllisesti haastavaa, jopa kun tiettyjä vertailuluokkia (esim. sukupuoli, etnisyys) voidaan pitää ennalta huomionarvoisina. Tutkimuksessa on myös huomautettu, että järjestelmän kehittäjän tai operoijan näkökulmasta syrjiviä vaikutuksia on usein hankala tunnistaa ennen kuin ne ilmenevät käytössä[197]. Erityisesti merkityksellisten ”muiden henkilöön liittyvien syiden” tunnistaminen etukäteisesti on erityisen kontekstisidonnaista. Vertailuryhmien valinta saattaakin edellyttää erityisesti käyttökontekstiin liittyvää erityisasiantuntijuutta ja tietoa, esimerkiksi teknologiakehittäjiltä saattaa puuttua käytännöllisistä syistä ja olosuhteista johtuen[198].

193 Ibid., huomio (44), s. 31.
194 Ks. esim. Green & Hu, 2018; Euroopan digitaaliset oikeudet, 2021.
195 Wachter, Mittelstadt & Russell, 2021, 22.
196 Wachter, Mittelstadt & Russell, 2020, 61.
197 Ks. Holstein ym., 2019; Euroopan digitaaliset oikeudet, 2021.
198 Holstein ym., 2019.

Moniperusteinen syrjintä esiintyy erityisenä haasteena vertailuryhmien valinnan suhteen tekoälyn kontekstissa. Moniperusteisen syrjinnän todentaminen on usein oikeudellisessa mielessä hankalaa, mutta tekoälyn kontekstissa myös teknisesti hankalaa (tai jopa mahdotonta). Oikeuskäytäntöjen puutteellisuus moniperusteisen syrjinnän käsittelemisessä heijastuukin tässä suhteessa automatisoitujen järjestelmien kontekstiin[199]. Esimerkiksi teknisestä näkökulmasta moniperusteisen syrjinnän riskien tunnistaminen on haastavaa johtuen mahdollisten kompleksisten ryhmien (esim. ”nainen” + ”70-vuotias”) määrästä[200]. Korostettakoon, että tutkimuskirjallisuudessa on esitetty teknisiä menetelmiä, joilla pystytään arvioimaan myös kompleksisten ryhmien suhteellista edustusta opetusdatassa[201] sekä ennusteiden jakaumissa[202]. Näiden menetelmien ei voida sosiaalieettisestä näkökulmasta kuitenkaan odottaa tekevän mahdollisesti oikeudellisesti merkityksellistä eroa moniperusteisen syrjinnän additiivisten ja intersektionaalisten muotojen välillä. Oikeudellisesta näkökulmasta saattaa olla esimerkiksi merkityksellistä, kohdellaanko yksilöä syrjivästi naissukupuolensa vuoksi ja ikänsä vuoksi (additiivinen merkitys) vai kohdellaanko häntä epäsuotuisasti 70-vuotiaana naisena (intersektionaalinen merkitys).

2.7.1.5 Oikomismenetelmien soveltuvuus ja riittävyys

Suuressa osassa kartoitettuja raportteja oikomismenetelmistä sekä teknisellä tasolla tapahtuvista interventioista visioidaan keskeistä lähestymistapaa syrjivien vaikutusten ehkäisemisessä. Kriittisiäkin huomioita on kuitenkin löydettävissä, jotka horjuttavat oletusta näiden menetelmien soveltuvuudesta tai riittävyydestä. Euroopan digitaaliset oikeudet -edunvalvontaryhmän raportissa huomautetaan seuraavaa:

> Oikomismenetelmät yksinkertaistavat liiallisesti monimutkaisia epäoikeudenmukaisuuden ongelmia tai jopa politiikkaa. […] [Niitä] voi olla hankala soveltaa tai ne voivat olla riittämättömiä tekoälysovellusten syrjivien vaikutusten auditoimisessa. Vinoumien viitekehys on potentiaalisesti hyödyllinen post-hoc -työkaluna tunnistettavissa olevien syrjivien vaikutusten tunnistamisessa. […] [N]äiden menetelmien soveltaminen on kuitenkin kaukana takeesta, että algoritmit tai järjestelmät, joihin ne on integroitu, ovat ”vapaita syrjivistä vaikutuksista”.[203]

199 Wachter, Mittelstadt & Russell, 2020, 20–21.

200 Kearns ym., 2018.

201 Jin ym., 2020.

202 Kearns ym., 2019.

203 Euroopan digitaaliset oikeudet, 2021, 113.

Tässä osiossa tarkastellut haasteet syrjivien vaikutusten tunnistamiselle antavat tukea tälle väitteelle, mutta tutkimuskirjallisuudessa on myös huomautettu, että oikomismenetelmien käyttöön ei ole yleistä tai vertailun mahdollistavaa ohjeistusta. Riskinä on, että ”näitä työkaluja käytetään niille soveltumattomiin käyttötarkoituksiin, tulkitaan väärin huolimatta käytettyjen menetelmien ennakko-oletuksista ja rajoitteista, ja/tai käytetään (tarkoituksellisesti tai muutoin) virheellisenä sertifikaattina algoritmin reiluudesta”[204].

Milloin oikomisesta tulee positiivista erityiskohtelua?

Kysymys, joka ei juurikaan noussut esille tarkastellussa aineistossa, mutta joka on saanut huomiota tutkimuskirjallisuudessa, on seuraava: tuleeko tekoälyn hyödyntämän mallin muokkaamista pitää oikeasuhtaisena ”positiivisena erityiskohteluna” silloin, kun kyseiset muokkaukset johtavat heikommassa asemassa olevan ryhmän parantamiseen[205]?

Yhdenvertaisuuslaissa positiivisen erityiskohtelun oikeutettavuuden ehtoihin kuuluvat muun muassa suunnitelmallisuus ja oikeasuhtaisuus (ts. velvollisuus päättää kohtelusta etukäteen eikä vasta päätöstilanteessa ja mitoittaa toiminta niin, että se on oikeassa suhteessa tavoiteltuun päämäärään eli tosiasiallisen yhdenvertaisuuden edistämiseen). Lisäksi positiivinen erityiskohtelu ei voi tarkoittaa ehdottoman etusijan antamista aliedustettuun ryhmään kuuluvalle henkilölle automaattisesti, vaan esimerkiksi rekrytoinnissa hakijoiden välillä on suoritettava tasapuolinen vertailu ja kyseiseen ryhmään kuuluva henkilö voidaan valita suunnilleen yhtä pätevien hakijoiden joukosta.

Suunnitelmallisuuden ehdon voidaan kenties katsoa täyttyvän algoritmisen päätöksenteon kontekstissa, mikäli jonkinlainen positiivisen erityiskohtelun mekanismi rakennetaan järjestelmään jo lähtökohtaisesti. Oikeasuhtaisuudelle ei kuitenkaan ole olemassa standardisoitua tai yksiselitteistä tilastollista sääntöä, jota voitaisiin hyödyntää algoritmien reiluutta koskevissa tarkasteluissa (mikä vaikeuttaa esimerkiksi rajanvetoa välillisen syrjinnän ehkäisemisen ja positiivisen erityiskohtelun välillä). EU:n oikeudellisessa kontekstissa kysymystä käsittelevässä tutkimuskirjallisuudessa on esitetty, että mallien reiluuden parantaminen yhtäläisen sensitiivisyyden (vrt. tosien positiivisten määrä) varmistamiseksi ryhmien välillä käyttämällä eri kynnysarvoja päätöksenteossa voidaan ymmärtää oikeutettuna positiivisena erityiskohteluna[206].

204 Lee & Singh, 2021.

205 Ks. esim. Dwork ym., 2012.

206 Hoch ym., 2021.

Oikomismenetelmien käyttö dynaamisissa konteksteissa

Yllä tarkastellut käyttökontekstin ja kohdepopulaatioiden dynaamiset muutokset ovat huomionarvoisia myös oikomismenetelmien käytön kontekstissa. Käyttöympäristön dynaamiset muutokset – ml. riippumattoman tai riippuvan muuttujan muutokset, tai konstruktion muutokset – saattavat nimittäin vaikuttaa tosiasiallista yhdenvertaisuutta edistämään pyrkivien oikomismenetelmien käytön tehokkuuteen, kun tekoälysovelluksen käyttöä ja vaikutuksia tarkastellaan pitemmällä aikavälillä. On huomattava myös, että edellä mainitut muutokset voivat johtua itse tekoälyjärjestelmän käyttöönotosta. Käyttöönotto saattaa muuttaa kohdepopulaation käyttäytymistä sekä mahdollisesti myös sosiaalisia normeja tai arvoja, mistä voi seurata, että interventiot (vrt. oikominen) eivät ole vaikuttavia tai tuottavat jopa negatiivisia vaikutuksia[207].

Oikomismenetelmät keskittyvät yhden toimijan paikalliseen näkökulmaan, mutta interventioiden vaikutukset realisoituvat usein monien toimijoiden vuorovaikutuksen tuloksena. Esimerkiksi mikäli vain osa sosio-teknisen ekosysteemin toimijoista tekee ”reiluutta” parantavia interventioita algoritmeihinsa (esim. osa tietyllä alueella rekrytoivista yrityksistä), voi kyseiseen ekosysteemiin muodostua kannustinrakenteita, joilla on negatiivisia yhdenvertaisuusvaikutuksia pitkällä tähtäimellä[208]. Oikomismenetelmien käytöllä voi olla myös pitkässä juoksussa tavoitellulle lopputulemalle päinvastainen vaikutus[209]. Laajempien yhdenvertaisuusvaikutusten ja niiden kestävyyden arvioiminen näyttäisikin edellyttävän jatkuvaa arviointia, joka ottaa huomioon algoritmien vaikutukset, kun niitä käytetään ns. useammalla päätöksenteon kierroksella.

Yllä esitetyt huomiot ympäristön dynamiikan vaikutuksista ovat merkityksellisiä syrjivien vaikutusten ehkäisemisen kannalta. Ne ovat kuitenkin merkityksellisiä myös yhteiskunnallisesta ja sosiaalieettisestä tasa-arvon ja yhdenvertaisuuden pitkäjänteisen edistämisen näkökulmasta. Dynaamisten muutosten huomiotta jättäminen ei välttämättä yksittäisten toimijoiden tapauksessa tarkoita, että kyseiset toimijat olisivat laiminlyöneet velvollisuuttaan edistää tosiasiallista yhdenvertaisuutta. Kuitenkin rakenteellisten yhdenvertaisuuden ja tasa-arvon esteiden poistaminen saattaa edellyttää koordinoitua toimintaa, jossa useiden ekosysteemissä käytettyjen tekoälysovellusten yhteisvaikutukset ja ympäristön sekä populaation muutokset otetaan huomioon.

207 Selbst ym., 2019.

208 Ks. Fazelpour, Lipton & Danks, 2021.

209 Liu ym., 2018.

2.7.1.6 Selitysmenetelmiin liittyvät haasteet

Tekoälyjärjestelmien läpinäkyvyyden ja/tai selitettävyyden puute tunnistettiin aineistossa yhdeksi riskialueeksi yhdenvertaisuuden näkökulmasta, sillä se voi muodostua haasteeksi objektiivisen oikeutuksen antamisen ja päätöksenteon kohteen oikeuksiinsa pääsemisen näkökulmasta. Selitysmenetelmillä voidaan osaltaan vastata näihin haasteisiin tuottamalla selityksiä järjestelmän toiminnasta tai tulosteiden taustalla olevista syistä. Kyseisten menetelmien valintaan ja käyttöön liittyy kuitenkin haasteita, jotka vertautuvat soveltuvien reiluusmittareiden ja oikomismenetelmien valintaan. Yleiset ongelmat liittyvät muun muassa siihen, että:

- datasta opitut mallit sisältävät tilastollista epävarmuutta ja tämä epävarmuus koituu haasteeksi myös, kun mallin pohjalta tuotetaan selityksiä;
- selitysmenetelmillä tuotetut selitykset kuvaavat mallia eivätkä mallinnetun ympäristön kausaalisia rakenteita;
- selitysmenetelmät operoivat olettamuksella, että mallin muuttujat tai opitut piirteet eivät korreloi keskenään[210].

Yksittäisiin selitysmenetelmiin liittyy myös niille ominaisia rajoitteita ja haittapuolia. Mallien suhteen agnostisilla menetelmillä voidaan tuottaa tietoa mallista tai tulosteista esimerkiksi manipuloimalla syötedataa, mallin muuttujia tai piirteiden painoja. Joissakin tapauksissa tuotetut selitykset (esim. visuaaliset mallit) saattavat olla kuitenkin hankalia tulkita moniulotteisuudestaan johtuen, ja osa menetelmistä perustuu olettamukseen, että mallin muuttujat tai opitut piirteet eivät korreloi keskenään[211]. Tämä saattaa olla ongelmallista erityisesti silloin, kun pyritään tuottamaan tietoa siitä, esiintyykö järjestelmän käytössä välillisesti syrjiviä vaikutuksia. Kontrafaktuaalisiin selityksiin, joiden hyödyt syrjivien vaikutusten tunnistamisessa on tunnistettu tutkimuskirjallisuudessa[212], puolestaan liittyy haasteita silloin, kun tulosteelle on useita vaihtoehtoisia kontrafaktuaalisia selityksiä[213].

Selitysmenetelmien käyttö voi myös edellyttää kompromisseja (vrt. reiluustavoitteiden yhteensopimattomuus). Moniulotteisista ”musta laatikko” -malleista voidaan esimerkiksi tuottaa yksinkertaisempia, tulkittavia malleja opettamalla ne edellä mainitun syötteillä ja tulosteilla, mutta tämä lähestymistapa voi tuottaa ymmärrettävämmän mallin usein osumatarkkuuden heikkenemisen kustannuksella[214]. Lisäksi selitysmenetelmien käyttö voi asettaa järjestelmät alttiimmaksi nk. ”häiriöesimerkeillä (engl.

210 Molnar, Casalicchio & Bischl, 2020.
211 Ibid.
212 Wachter, Mittelstadt & Russell, 2017.
213 Tätä ilmiötä kutsutaan ”Rashomon-efektiksi”.
214 Ks. esim. Rudin, 2019.

adversarial example) tehtäville ”häiriöhyökkäyksille” (engl. *adversarial attack*). Esimerkiksi sisällönmoderointialgoritmeja voidaan periaatteessa huijata luokittelemaan vihapuhetta soveliaaksi puheeksi, jolloin algoritmi ei suodata haitallista sisältöä pois[215]. Selitysmenetelmiä hyödyntävät tekoälyjärjestelmät ovat yleisesti ottaen alttiimpia tällaiselle manipulaatiolle, sillä päätöksenteon kohde saattaa kyetä esimerkiksi päättelemään mallin toimintaa ohjaavia periaatteita sitä koskevien selitysten perusteella.

2.7.2 Käyttäjä- ja organisaatiotason haasteet

Kartoituksen pohjalta voidaan eritellä myös yleisemmällä tasolla esiintyviä haasteita, jotka liittyvät muun muassa diversiteetin ja (vähemmistö)ryhmien edustuksen puutteeseen teknologiaa kehittävissä organisaatioissa ja epätietoisuuteen lainmukaisuusvaatimuksista (tai näiden uupumiseen).

2.7.2.1 Diversiteetin ja edustuksen puute

Useassa raportissa nostettiin esille diversiteetin ja edustuksen puute teknologiaa kehittävissä organisaatioissa erityisesti naisten, etnisten vähemmistöjen ja vammaisten henkilöiden suhteen[216]. Diversiteetin puute ja tiettyjen ihmisryhmien aliedustus voidaan nähdä osaltaan sekä riskinä syrjivien vinoumien syntymiselle että haasteena yhdenvertaisuusvaikutusten arvioimiselle ja ehkäisemiselle. Aineistosta nousee esille, että monimuotoisuuden ja edustuksen puute koskee niin teknologia-alaa yleisesti kuin tekoälyä kehittäviä organisaatioita, ammattiryhmiä sekä yksittäisiä kehitystiimejä. Paikoin korostetaan, että pelkkä edustus esimerkiksi kehitystyölle periferaalisissa rooleissa ei riitä, vaan tarvitaan keinoja purkaa myös sukupuolittuneita normeja ja käytäntöjä organisaatioiden ja tiimien sisällä[217]. Korostettakoon myös, että yhdessä raportissa Suomi mainitaan esimerkkinä maasta, jossa naisten edustus teknologiateollisuudessa on puutteellista ja jossa koulutusvalinnat ovat erityisen sukupuolittuneita[218].

Diversiteetin ja edustuksen puutteen nähdään olevan yhteydessä muun muassa aliedustettujen ryhmien vaatimusten, intressien ja oikeuksien sivuuttamiseen teknologiasuunnittelussa sekä merkityksellisten näkökulmien laiminlyömiseen yhdenvertaisuus- ja tasa-arvovaikutusten arvioinnissa käyttöönotossa. Tutkija Joy Buolamwini onkin kutsunut

215 Algo:aware, 2018, 28.

216 Ks. esim. Algo:aware, 2018; Centre for Data Ethics and Innovation, 2020a; Gerards & Xenidis, 2021; UNESCO, 2020; The Greenlining Institute, 2020.

217 UNESCO, 2020, 27.

218 Gerards & Xenidis, 2021, 89.

”etuoikeutetuksi tietämättömyydeksi” ilmiötä, jossa ”suuri osa [tekoäly]tutkijoista, -kehittäjistä ja -koulutuksentarjoajista on suojattu tekoälyjärjestelmien tuottamilta vahingoilta”[219]. Tämä puolestaan johtaa aliedustettujen ryhmien ”ongelmien aliarvioimiseen, toissijaistamiseen ja tiedostamattomuuteen” ja kontekstille sokeiden tekoälyratkaisujen luomiseen[220]. Diversiteetin ja edustuksen vaatimusten ei tulisi täten ymmärtää koskevan vain koulutukseen ja työmarkkinoihin liittyviä tasa-arvonäkökulmia vaan myös tiedollisia näkökulmia ja asiantuntijuutta, joita tarvitaan ihmisten yhdenvertaisuutta kunnioittavien tekoälyteknologioiden luomiseen.

2.7.2.2 Epätietoisuus lainmukaisuusvaatimuksista

Nimenomaisesti tekoälyn vinoumia koskevan sääntelyn sekä auditointia tai yleisempää vaikutuksenarviointia koskevien standardien puute – tai epätietoisuus olemassa olevasta sääntelystä ja standardeista – näyttäisi muodostavan haasteen vinoumien asianmukaisen tunnistamisen ja ehkäisyn kohdalla[221]. Epäselvyyteen ja epätietoisuuteen liittyvät ongelmat limittyvät yllä tunnistettujen haasteiden kanssa, kuten soveltuvien reiluusmetriikoiden valitsemiseen (ks. yllä) ja sensitiivisen datan käyttöön liittyvien avoimien kysymysten kanssa. Joissakin raporteissa nostettiin esille, että teknologian kehittäjillä ja käyttäjillä ei useinkaan ole tietoa siitä, milloin, miten ja missä määrin kiellettyjä syrjintäperusteita koskevaa dataa voidaan hyödyntää syrjintäriskien ehkäisemisessä ja arvioimisessa[222]. Yhdessä tarkastelluista raporteista huomautettiin lisäksi, että mikäli ”[tekoälyjärjestelmien] ei anneta ottaa kiellettyjä syrjintäperusteita, kuten rotua tai etnisyyttä, huomioon ne voivat tehdä jäsenvaltioiden yhtäläisiin työmahdollisuuksiin tähtääviä erityisiä työllistämistoimenpiteitä tai positiivista erityiskohtelua tyhjäksi”[223].

Kommunikaatiokatkokset organisaatioissa

Kommunikaatiokatkokset organisaatioissa voivat muodostua haasteeksi syrjivien vaikutusten tunnistamiselle ja ehkäisemiselle. Tekoälysovellusten kehittämisessä ”lainmukaisuuden arviointi on paikoin irrallaan teknisestä puolesta”[224], mikä vaikeuttaa oikeudellisten vaikutusten arviointia. Organisaatiotason pirstoutuneisuus ja yhteisten käytäntöjen puute syrjintäriskien tunnistamista ja ehkäisemistä koskien voidaan täten nähdä organisaatioiden

219 UNESCO, 2020, 23.
220 Ibid.
221 Centre for Data Ethics and Innovation, 2020b.
222 Ks. esim. Centre for Data Ethics and Innovation, 2020a.
223 Achiume, 2020, 9.
224 Euroopan unionin perusoikeusvirasto, 2020, 94.

toimintaan, kulttuuriin ja kommunikaatioon liittyvänä erityishaasteena[225]. Tutkimuskirjallisuus tukee huomiota siitä, että tekoälysovellusten tutkimus- ja kehittämisprosesseissa epätietoisuus lainmukaisuusvaatimuksista sekä kommunikaation puute kehittävän tahon tiimien välillä saattavat koitua esteiksi asianmukaiselle riskinarvioinnille[226].

Epäselvyys vastuukysymyksistä

Syrjintäriskien ja yhdenvertaisuusvaikutusten arvioinnit vaikuttaisivat olevan vielä harvinaisia kentällä, joskin tietoisuus aiheesta näyttäisi kehittyvän. Tekoälyjärjestelmien monimutkaiset tuotanto- ja palveluketjut kuitenkin hankaloittavat syrjintäriskien arvioimista, mikä näkyy epätietoisuutena vastuukysymyksistä[227]. Keskeinen kysymys onkin, kenen tulisi suorittaa algoritmien auditointia ja missä vaiheessa arvoketjua.

Sovellusten kehittäjät saattavat ajatella perusoikeusvaikutuksienarvioinnin kuuluvan asiakkaille, jotka käyttävät heidän tuotteitaan[228]. Kenellä on vastuu järjestelmän vinoumien ehkäisemisestä on epäselvää erityisesti, kun tuotanto- ja palveluketjut yhdistävät useita toimijoita ja koska eri tahot saattavat kehittää lopullisen järjestelmän eri komponentteja[229].

Vastuukysymysten epäselvyys on erityisen ongelmallista sikäli, kun esimerkiksi EU:n poliittisissa asiakirjoissa teknisiä oikomismenetelmiä on nostettu esiin itsesääntelylähestymistapaan sopivina apuvälineinä. Tämä lähestymistapa tekoälysovellusten voi kuitenkin antaa palveluntarjoajille merkittävää valtaa syrjintää ja eriarvoisuutta koskevissa kysymyksissä[230]. Lähestymistavan soveltuvuus esimerkiksi API-muotoisten tekoälyjärjestelmien kontekstissa on myös kyseenalaista, sillä API-muotoisia tekoälyjärjestelmiä tai -toiminnallisuuksia voidaan ottaa käyttöön eri toimijoiden kehittämissä laajemmissa sovelluksissa tai palveluissa. Mahdollisten syrjivien vaikutusten populaatio- ja kontekstisidonnaisuus huomioiden API-tekoälytoiminnallisuuksia hyödyntävien sovellusten asianmukainen auditoiminen voi edellyttää toimia palveluntarjoajan mutta erityisesti lopullista sovellusta käyttävän tahon toimesta.

225 Centre for Data Ethics and Innovation, 2020b, 60.

226 Holstein ym., 2019.

227 Ks. esim. Centre for Data Ethics and Innovation, 2020a; Euroopan digitaaliset oikeudet, 2021.

228 Euroopan unionin perusoikeusvirasto, 2020, 9.

229 Euroopan digitaaliset oikeudet, 2021, 77.

230 Euroopan digitaaliset oikeudet 2021, 125.

Tutkijoita ja tutkimusyhteisöjä sekä muita kolmannen sektorin toimijoita on myös visioitu tekoälysovellusten auditointia suorittaviksi tahoiksi. Teknologian kehittäjän tai palveluntarjoajan suostumukseen pohjautuva auditointi kolmannen osapuolen toimesta voi olla ongelmallista, sillä edellä mainituilla voi olla sekä intressit että valta estää mahdollisten syrjivien vaikutusten esiintuominen. Esimerkiksi Facebook on estänyt tutkijoita käyttämästä alustaa tutkimustarkoituksiin[231]. Käyttäjille ja/tai muille yksilöille ja yhteisöille voitaisiin myös antaa valta vaatia tai käynnistää riippumaton tekoälysovellusten auditointiprosessi[232]. Periaatteessa auditointi voitaisiin käynnistää ennen järjestelmän käyttöönottoa tai sen jälkeen.

2.7.2.3 Itsesääntelyyn perustuvan lähestymistavan haasteet

Osassa tarkastelluista raporteista ilmaistiin huolta itsesääntelyyn perustuvien lähestymistapojen riittävyydestä mahdollisten syrjivien vaikutusten tunnistamisessa ja ehkäisemisessä. Merkityksellisiä haasteita tässä suhteessa ovat kannustimien puute sekä vallan keskittyminen yksityisen sektorin toimijoille.

Osassa raporteista epäillään, onko teknologiakehittäjillä ja palveluntarjoajilla kannusteita tehdä asianmukaista yhdenvertaisuusvaikutusten arviointia ja ehkäistä syrjinnän riskejä. Esimerkiksi Euroopan digitaaliset oikeudet -edunvalvontaryhmän raportissa huomautetaan, että "oletus, että palveluntarjoajilla olisi välttämättömät ja riittävät kannustimet vinoumien tunnistamiseen ja oikomismenetelmien käyttämiseen, ei pidä aina paikkaansa" erityisesti, kun otetaan huomioon, että nämä toimenpiteet saattavat edellyttää mallin tarkkuudesta tinkimistä, uuden datan keräämistä sekä järjestelmän käyttöönoton viivyttämistä[233].

Paikoin aineistosta nousi myös huoli siitä, että standardien puute – tai epäselvyys olemassa olevia standardeja koskien – suo yksittäisten sektoreiden toimijoille suhteettomasti valtaa oikeudenmukaisuutta koskevissa kysymyksissä[234]. Esimerkiksi finanssialalla esiintyvä epätietoisuus standardeja koskien[235] osoittautuu tässä suhteessa ongelmalliseksi:

231 Facebook bans academics who researched ad transparency and misinformation on Facebook. (4.8.2021). The Verge. Verkossa: thev-erge.com/2021/8/4/22609020/facebook-bans-academic-researchers-ad-transparency-misinformation-nyu-ad-observatory-plug-in.

232 Euroopan digitaaliset oikeudet, 2021.

233 Euroopan digitaaliset oikeudet, 2021, 77.

234 Euroopan digitaaliset oikeudet, 2021.

235 Ks. myös Euroopan unionin perusoikeusvirasto, 2020.

”Jos konsensusta reiluuden suhteen ei löydy, [finanssiala] tulee päättämään siitä. Ilman yhteistä näkemystä ja selkeitä ohjeistuksia datan reilusta käytöstä, teollisuus tulee asettamaan *de facto* standardit itse.”[236]

2.8 Osion yhteenveto

Toisen osatehtävän löydökset osoittavat, että tekoälyn kohdalla yhdenvertaisuusperiaatteen toteutumisen arvioiminen on tärkeää läpi sektorien ja toimialojen. Välillisen syrjinnän riskit vaativat erityistä huomiota, sillä koneoppimisalgoritmit voivat oppia opetusdatasta muuttujia, jotka korreloivat (mahdollisesti usean) kielletyn syrjintäperusteen kanssa. Välittömän syrjinnän riskejä esiintyy erityisesti, kun käsitellään erityisiä henkilötietoja, kuten syrjintäperusteita koskevaa dataa. Koneoppimiseen pohjautuvat mallintamismetodit voivat myös mahdollistaa tahallisen syrjinnän naamioimisen näennäisen neutraalien, mutta välillisesti syrjivien muuttujien ja päätöksentekosääntöjen käyttämisen avulla. Teknologisesta näkökulmasta käytetty data, algoritmit, mallit, optimointikriteerit sekä laskentaprosessien monimutkaisuus osoittautuvat merkityksellisiksi, mutta myös sosioteknisetekijät, kuten käyttäjien rooli päätöksenteossa ja vaikutuksenalaan kuuluvien ihmisten demografiset tekijät, ovat kaikki merkittäviä syrjivien vaikutusten tunnistamisen ja tehokkaan ehkäisemisen näkökulmasta.

Mahdollisuudet parantaa päätöksentekoa yhdenvertaisuuden näkökulmasta käyttämällä tekoälyä tunnistetaan sinänsä laajalti, mutta keskustelu näistä mahdollisuuksista on suurilta osin puutteellista tai keskittyy hyvin kapeaan näkökulmaan, kuten päätösten tarkkuuden tai objektiivisuuden lisäämiseen. Tekoälyn esimerkiksi ajatellaan parantavan päätöksiä näissä suhteissa, vaikka mallien osumatarkkuuden maksimoiminen historiallista dataa vasten saattaa nimenomaan johtaa vinoumien uusintamiseen. Laajempi näkökulma tosiasiallisen yhdenvertaisuuden edistämiseen tekoälyn kontekstissa tarvitaan, jotta vastuullisten toimijoiden velvollisuudet, kuten yhdenvertaisuusvaikutusten ennakkoarviointi, kohtuullisten mukautusten tekeminen ja sidosryhmien osallistaminen, voidaan täyttää. Toisin sanoen, yhdenvertaisuusperiaatteen toteutumista sen täydessä merkityksessä ei redusoida yhteen kriteeriin tai menettelytapaan ja täten automatisoida täysin, vaan yhdenvertaisuuden toteutuminen edellyttää kontekstuaalista ja tapauskohtaista harkintaa sekä paikoin tekoälyavusteiselle tai -vetoiselle päätöksenteolle vaihtoehtoisia tai täydentäviä menettelytapoja.

236 Centre for Data Ethics and Innovation, 2020b, 58.

Laadun- ja vaikutuksenarviointiprosesseilla, läpinäkyvyyttä parantavilla menetelmillä sekä opetusdataan tai järjestelmälle opetettuun malliin kohdistetuilla tilastollisilla testeillä voidaan kuitenkin saada indikaatioita mahdollisista syrjivistä vaikutuksista ja ehkäistä niitä. Yhdenvertaisuusperiaatteen toteutumisen arviointi edellyttää kaikissa tapauksissa myös tapauskohtaista arviointia, joka ottaa huomioon laajemman käyttökontekstin ja proseduraaliset näkökulmat, kuten päätöksenteon läpinäkyvyyden. Erityisesti tilastollisten testien ja reiluusmetriikoiden käyttö voi olla useassa tapauksessa merkittävä tai jopa välttämätön osa riskinarviointia, joskaan ei riittävä yhdenvertaisuusperiaatteen toteutumisen asianmukaiselle arvioinnille. Keskeiseksi puutteeksi tässä suhteessa nousee tilastollisten testien käyttöä koskeva standardien puute ja syrjinnän kieltoon sisäänrakennettu faktuaalisen perustan tarkastelun edellytys sekä siitä seuraava tapauskohtaisuus. Mikäli mallien tasolla tapahtuvasta yhdenvertaisuusvaikutusten tarkastelusta (tai algoritmien auditoinnista) halutaan asianmukainen ja vaikuttava keino syrjivien vaikutusten arviointia ja ehkäisemistä varten, niiden toteuttamiseen tarvitaan yhteisiä ohjaavia standardeja siltä osin kuin ne ovat soveltuvia tähän tehtävään, mutta myös tietoa siitä, milloin ja missä tapauksissa ne eivät ole soveltuvia ja riittäviä.

Kokoamme alla muutamia keskeisiä johtopäätöksiä ja suosituksia vaikuttavan ja tehokkaan arviointikehikon suunnittelun näkökulmasta.

Syrjivien vinoumien tunnistamisesta ja ehkäisemisestä on tehtävä oikeudellisesti velvoittavaa.

Kartoituksen löydökset tukevat näkemystä, että mahdollisesti syrjivien vinoumien arvioimisen ja ehkäisemisen tulisi olla oikeudellisesti velvoittavaa. On huomionarvoista, että yhdenvertaisuuslaki asettaa jo nykyisellään tässä suhteessa merkittäviä velvoitteita, kuten viranomaisten velvollisuuden arvioida yhdenvertaisuusvaikutuksia, ja näiden velvoitteiden voidaan odottaa olevan merkityksellisiä myös tekoälyjärjestelmien kontekstissa. Vinoumien ennakkoarviointia ja korjaamista (vrt. ”oikomista”) koskeva velvollisuus muodostuu oletettavasti tulevaisuudessa osaksi EU-lainsäädäntöä[237].

237 TEKOÄLYÄ KOSKEVISTA YHDENMUKAISTETUISTA SÄÄNNÖISTÄ (TEKOÄLYSÄÄDÖS) JA TIETTYJEN UNIONIN SÄÄDÖSTEN MUUTTAMISESTA, artikla 10, kohdat 2 ja 5.

Kartoituksen löydökset antavat kuitenkin perusteita painottaa seuraavia seikkoja arvioinnin, monitoroinnin ja korjaamistoimenpiteiden asianmukaisuuden ja vaikuttavuuden takaamiseksi:

- Kyseisten toimenpiteiden tulee olla jatkuvia ja/tai säännöllisiä. Tämä pätee erityisesti tapauksissa, joissa käytetty järjestelmä hyödyntää jatkuvaa oppimista tai jonka päätöksentekosääntöjä tai mallia päivitetään ajan kuluessa.
- Toimenpiteiden tulee olla tekoälyjärjestelmän käyttötarkoitukseen ja odotettavien vaikutusten suhteen oikeasuhtaisia.
- Kyseisten toimenpiteiden tulee olla kussakin käyttökontekstissa sovellettavien oikeudellisten normien ja standardien (mukaan lukien hyvän hallinnon periaatteiden ja ammatillisten eettisten standardien) ohjaamia. Tämä edellyttää huomion kiinnittämistä esimerkiksi sektoriin ja toimialaan, jossa tekoälyjärjestelmää operoidaan, maantieteelliseen sijaintiin sekä kohdepopulaation demografisiin tekijöihin.
- Toimenpiteiden tulee olla tutkitun tiedon perusteella vaikuttavia ja kestäviä.

Ihanteellisessa tilanteessa arviointi, monitorointi ja korjaamistoimenpiteet olisivat myös eri toimijoiden kesken ekosysteemitasolla koordinoidusti toteutettuja tosiasiallisen yhdenvertaisuuden ja tasa-arvon edistämisen edesauttamiseksi laajemmin yhteiskunnassa.

Lainmukaisuusvaatimuksien, vastuiden ja velvollisuuksien täytyy olla selviä organisaatiotasolla.

Tekoälyn vinoumien kontekstissa päällekkäisyys yhdenvertaisuus- ja tasa-arvolain, tietosuojalain sekä sektorikohtaisten säännösten välillä muodostuu etenevässä määrin tärkeämmäksi[238], minkä takia lainmukaisuusvaatimusten selkeyttämistä voidaan pitää ensimmäisenä keinona organisaatiotason esteiden poistamisessa. Erityisesti edellä mainittujen lainmukaisuusvaatimusten yksityiskohtaiseen tuntemukseen organisaatioissa, vastuukysymyksiin sekä vaikuttavaa riskinarviointi- ja hallintaa edesauttavan informaation kulkemiseen toimijoiden välillä täytyy kiinnittää erityistä huomiota. Tekoälysovellusten tuotanto- ja käyttöketjut sisältävät useita eri toimijoita, minkä takia on tärkeää selventää, mitkä tuotteet, komponentit tai muut osat näitä ketjuja ovat kunkin toimijan vastuulla ja miten esimerkiksi tulisi vastata tilanteisiin, joissa (kausaalinen) vastuu tekoälyjärjestelmien vaikutuksista näyttäisi jakautuvan eri toimijoiden välillä.

238 Centre for Data Ethics and Innovation, 2020, 11.

Lainmukaisuusvaatimusten selventäminen voi tapahtua niin organisaatioiden kuin viranomaisten toimesta esimerkiksi kompetenssien kasvattamisen muodossa tai kehittämällä tiiviimpää yhteistyötä tekoälyä kehittävien tahojen sekä tietosuoja- ja yhdenvertaisuusvaltuutettujen välillä. Edellä mainitut ovat suositeltavia ratkaisuja Suomen kontekstissa, joskin tekoälysäädöksen oletetaan osaltaan vastaavan esimerkiksi vastuun jakautumista ja vinoumien arviointia ja korjaamista koskeviin kysymyksiin myös EU:n tasolla. Lain tulkintaan liittyvien avoimien kysymysten – esim. syrjivän ohjeen tai käskyn tulkitseminen tekoälyavusteisen päätöksenteon kontekstissa – selventäminen voi osaltaan edesauttaa myös lainmukaisuuden edistämisen toimia organisaatiotasolla.

Riskinarviointi ja auditointi edellyttävät selkeitä standardeja.

Koneoppimisalgoritmien ja tilastollisten mallien käyttö päätöksenteossa tarjoaa periaatteessa tehokkaita keinoja arvioida ja parantaa yhdenvertaisuusvaikutuksia, mutta tämä edellyttää ymmärrystä käyttökontekstista, kohdepopulaatiosta sekä lain asettamista velvollisuuksista ja mahdollisuuksista. Mikäli algoritmien auditoinnista (vrt. esim. reiluuden arviointi) ja oikomismenetelmien käytöstä halutaan tehokas itse- tai ulkoisen sääntelyn keino yhdenvertaisuusvaikutusten parantamiseen, niiden käyttöä tulisivat ohjata tutkittu tieto ja selkeät standardit. Erityiset kysymykset koskevat esimerkiksi erityisten henkilötietojen käsittelyä erinäisissä tarkoituksissa (esim. onko tietojen käsittely testausvaiheessa ja/tai algoritmin tai mallin parantamiseksi yleisen edun mukaista?) sekä soveltuvien lähestymistapojen valintaa (vrt. kuinka valitaan reiluusmetriikat ja vertailuluokat?). Toisin sanoen standardien tulisi asettaa oikeellisuuden kriteerit sille, että toimijoiden voidaan katsoa noudattaneen algoritmien auditointia koskevia velvollisuuksia.

Auditointia koskevien standardien ja velvollisuuksien tulisi oikeellisuuden kriteerien lisäksi myös kattaa ohjeet ja/tai vaatimukset läpinäkyvyyden ja dokumentaation suhteen sekä testaamista tai arviointia suorittavien tahojen suhteen. Kannustin- ja resurssiongelmien ratkaiseminen saattaa edellyttää auditointia tukevien kompetenssien ja alustojen kehittämistä (vrt. tekoälysäädöksen ”testaus- ja kokeilulaitokset”) sekä mahdollisten seurauksien, joita standardien ja velvollisuuksien noudattamatta jättämisestä seuraisi, määrittämistä.

Oikeuksien toteutumista tulee turvata ja aktiivisesti edistää.

Ihmisoikeuksien, kuten yhdenvertaisuuden, on tarkoitus olla konkreettisia ja tehokkaita eikä teoreettisia ja abstrakteja oikeuksia[239]. Sikäli, kun perusoikeuksien tulee yhtä lailla turvata oikeudenmukaisuuden toteutumista, on yksilöiden pääsy heidän oikeuksiinsa turvattava. Oikeuksien tosiasiallinen konkreettisuus ja tehokkuus näyttäisivätkin etenevässä

239 Ks. esim. CLIFT v. THE UNITED KINGDOM 7205/07, kohta 60.

määrin tärkeämmiltä digitalisoituvassa yhteiskunnassa, ja kansalaisyhteiskunnan järjestöt ovatkin kritisoineet muun muassa tekoälysäädöksen ehdotusta sen riskilähtöisestä lähestymistavasta tekoälyn sääntelyyn perusoikeuksista lähtevän sääntelyn sijaan[240].

Yhdenvertaisuusperiaatteen toteutumista arvioitaessa algoritmivetoisia ja -avusteisia päätöksentekoprosesseja ei tule arvioida esimerkiksi pelkästään päätösten mahdollisten syrjivien vaikutusten suhteen. On arvioitava myös, mitkä ovat päätöksenteon kohteiden ja niiden vaikutuksenalaan kuuluvien ihmisten ja ryhmien tosiasialliset mahdollisuudet riitauttaa syrjiviksi havaittuja päätöksiä ja mikäli tarvitaan vaihtoehtoisia menettelytapoja tietyille ryhmille sikäli, kun he ovat niihin oikeutettuja. Yleisemmin, asianmukaisen ja vaikuttavan algoritmien auditoinnin tulisi heijastaa vaatimuksia ja velvollisuuksia, joita perus- ja ihmisoikeudet konkreettisina ja tehokkaina oikeuksina edellyttävät. Tässä mielessä näyttäisi tärkeältä, että esimerkiksi auditoinnista ei muodostu pinnallisen itsesääntelyn instrumentti, jota voidaan väärinkäyttää tai manipuloida tai joka keskittää valtaa yhdenvertaisuutta koskevien kysymysten saralla teknologia-alan toimijoille[241].

240 Ks. "The EU needs an Artificial Intelligence Act that protects fundamental rights". (30.11.2021). AccessNow. https://www.accessnow.org/eu-artificial-intelligence-act-fundamental-rights/. [Viitattu 10.12.2021]

241 Euroopan digitaaliset oikeudet, 2021.

3 Arviointikehikko syrjimättömille tekoälyjärjestelmille

3.1 Johdanto

Raportin kolmas osa esittelee hankkeessa luodun arviointikehikon syrjimättömille tekoälyjärjestelmille. Kehikko auttaa tunnistamaan ja hallitsemaan erityisesti julkisen sektorin tekoälyjärjestelmiin liittyviä syrjintäriskejä sekä edistämään yhdenvertaisuutta tekoälyn käytössä. Arviointikehikossa on painotettu sekä Suomen yhdenvertaisuuslakia että julkisen hallinnon tarpeita.

Arviointikehikko perustuu aiemmassa kahdessa osiossa esitetylle tilannekuvalle siitä, miten tekoälyä sovelletaan Suomessa eri sektorit ja toimijat huomioiden. Se painottaa myös oikeudelliseen syrjinnän näkökulmaa, ja syventää näin jo olemassa olevia ohjeistuksia[242] tekoälyn eettiseen hyödyntämiseen. Arviointikehikon tavoitteena on lisätä julkisen sektorin ja viranhaltijoiden kyvykkyyksiä arvioida tekoälyjärjestelmien mahdollisia syrjiviä vaikutuksia eri vaiheissa: järjestelmän hyödyntämisen suunnittelussa, hankinnassa, kehittämisessä tai käyttöönotossa. Kehikon tavoitteena on myös auttaa yksityisiä tekoälyn ja palveluiden kehittäjiä omien sovellustensa ja kehitysmenetelmiensä syrjintäriskien arvioinnissa erityisesti silloin, kun niitä kehitetään julkiselle sektorille. Kehikon muotoilussa on pyritty huomioimaan tarve yksityiskohtaiselle ohjeistukselle sekä samanaikaisesti huomioida syrjinnän tilannekohtaisuus ja tekoälyjärjestelmien verrattain nopea kehitys.

Kehikko pyrkii säilyttämään tasapainon Suomeen liittyvien erityiskysymysten ja eettiseen tekoälyn soveltamiseen keskittyvän kansainvälisen keskustelun välillä. Alati kehittyvä tutkimustieto, uudet tekoälyn sovellusalat sekä kehittyvä sääntely vaikuttavat siihen, miltä arviointikehikon perusteet näyttävät esimerkiksi viiden vuoden kuluttua. Tästä syystä kehikon lähtöoletukset, tausta sekä tilanneanalyysi on dokumentoitu mahdollisimman läpinäkyvästi. Tämä on keskeistä kehikon hyödynnettävyyden kannalta, jotta tilannekuvan muuttuessa ja sovellusten yleistyessä kehikko säilyisi käyttökelpoisena. Tästä johtuen myös tuoreiden lainsäädännöllisten hankkeiden, kuten kansallisen yleislainsäädännön automaattiselle päätöksenteolle ja EU:n tekoälyasetuksen vaikutuksia on pyritty kehikossa ennakoiden huomioimaan.

242 Katso liite 4.

Tämä osio kuvaa seuraavaksi lyhyesti kansainvälistä keskustelua algoritmisesta vaikutustenarvioinnista ja peilaa sitä Suomen julkisen hallinnon kontekstiin. Luvussa 3.2. esitetään arviointikehikon yhteiskehittämisprosessi ja sen muodostamiseksi tehty vertaileva analyysi olemassa olevista vastaavista kehikoista. Varsinainen arviointikehikko ja ohjeet sen käyttämiselle esitetään luvussa 3.3. Osion viimeinen luku, eli 3.4. esittelee koko hankkeen tuloksiin tukeutuvat politiikkasuositukset, jotka tukevat kehikon käyttöönottamista.

3.1.1 Algoritminen vaikutustenarviointi

Algoritminen vaikutusarviointi on noussut kansainvälisesti esiin yhtenä ratkaisuna tekoälyn luotettavaan ja eettisiin käyttöön, erityisesti julkisella sektorilla[243]. Algoritmisen vaikutustenarvioinnin tarkoituksena on saada tekoälyjärjestelmiä kehittävät ja tilaavat toimijat arvioimaan ja dokumentoimaan järjestelmien mahdollisia haitallisia yhteiskunnallisia vaikutuksia jo ennen käyttöönottoa niihin varautumiseksi ja vastuullisten tahojen tunnistamiseksi. Algoritmisen vaikutustenarvioinnin potentiaali piilee siinä, että se yhdistää järjestelmien haittojen näkyväksi tekemisen ja toimijoiden asettamisen vastuuseen niiden korjaamisesta käytännössä.[244] Se toimii siis riskienhallinnan ja vastuiden tunnistamisen menetelmänä algoritmisten järjestelmien suunnittelun ja käyttöönoton eri vaiheissa.

Algoritminen vaikutustenarviointi perustuu olemassa oleviin ihmisoikeuksia, tietosuojaa, yksityisyyttä ja ympäristövaikutuksia koskeviin vaikutustenarviointeihin. Se sisältää usein muun muassa algoritmien läpinäkyvyyden ja lainmukaisuuden arviointia sekä yhteiskunnallisten sidosryhmien konsultointia tekoälyjärjestelmän vaikutuksista. Vaikutustenarvioinnin vaatimiseen ja laukaisemiseen vaikuttavat tekoälyn suunniteltu käyttökohde ja riskialttius. Vaikka algoritminen vaikutustenarviointi suositellaan yleensä toteutettavaksi ennakkoon, se voidaan suorittaa eri vaiheissa tekoälyn elinkaarta: esim. ennen julkista hankintaa, kehitystyön aikana tai säännöllisin väliajoin käyttöönoton aikana. Tavoitteena on ennakoivasti tunnistaa tekoälyjärjestelmän käyttöönottoon liittyvät haitalliset riskit ja toimenpiteet niiden lieventämiseksi.

Algoritmista vaikutustenarvointia on kritisoitu siitä, että se keskittää liikaa valtaa algoritmien yksityisille kehittäjille, koska se perustuu asiantuntemukseen ja tietoon, johon vain järjestelmän kehittäjillä on pääsy.[245] Tästä johtuen on korostettu, että algoritmisen vaikutustenarvioinnin ei tulisi olla ainoastaan vapaaehtoinen itsearvioinnin työkalu eikä

243 Reisman ym., 2018.
244 Moss ym. 2021.
245 Selbst, 2021.

nojautua kapeisiin teknisiin mittareihin. Sen sijaan vaikutustenarvioinnin tulisi olla osa algoritmien pakollista sääntelyä, julkisten standardien ohjaamaa, sisältää avointa auditointia ja eri sidosryhmien osallistumista prosessiin ollakseen tehokasta.[246] Yksi algoritmisen vaikutustenarvioinnin päätehtävistä onkin pyrkiä tekemään julkisesti näkyväksi tekoälyjärjestelmien perustana olevat valinnat, kuten miten sen haitat, hyödyt ja riskit on määritelty. Tämän tiedon dokumentointi mahdollistaa muille toimijoille, kuten kansalaisjärjestöille tekoälysovelluksen hyväksyttävyyden ja tarkoituksenmukaisuuden arvioinnin sekä tarvittaessa julkisen kritisoinnin. Tällä hetkellä useimmat käytössä olevat algoritmisen vaikutuksenarvioinnin menettelyt eivät kuitenkaan ole järin läpinäkyviä, selkeitä vastuiden jakautumisen kannalta ja lisäksi niistä puuttuvat selkeät mahdollisuudet kansalaisten ja sidosryhmien osallistumiseen.[247]

Algoritmiseen vaikutustenarviointiin ei ole vain yhtä mallia, vaan ne vaihtelevat laajuutensa, velvoittavuutensa ja kriteeriensä suhteen merkittävästi. Kirjallisuudesta on tunnistettavissa ainakin kolme erillistä tapaa: 1) kyselylomake, 2) tietosuojaa koskeva vaikutustenarviointi (DPIA) ja 3) julkisen viraston malli.[248] Kyselylomakemallissa vaikutustenarviointi täytetään kysymys- ja vastausmuodossa, jonka perusteella saadaan jonkinlainen riskiluokitus tai pisteytys järjestelmälle. Kysymysten laajuus ja vastausvaihtoehdot voivat vaihdella paljon eri työkalujen välillä. Esimerkki kyselylomakemallista on Kanadan keskushallinnon *Algorithmic Impact Assessment (AIA)* -työkalu, joka valtion virastojen on suoritettava ennen automaattisen päätöksentekojärjestelmän käyttöönottoa. Myös tässä hankkeessa tuotettu arviointikehikko on lähimpänä kyselymallia. Tietosuojaa koskeva vaikutustenarviointi eli DPIA vaaditaan yleisen tietosuoja-asetuksen eli GDPR:n alla silloin, kun suunniteltu henkilötietojen käsittely todennäköisesti aiheuttaa korkean riskin ihmisten oikeuksille. Vaikka DPIA-malli keskittyy ensisijaisesti riskeihin henkilötietojen käsittelyssä, koskee se myös automatisoitua päätöksentekoa, tarjoten näin yhden mallin algoritmisen vaikutustenarvioinnin toteuttamiseen[249]. Arvioinnin avoimuus ja sidosryhmien osallistuminen ei ole kuitenkaan tähän mennessä toteutunut toivotusti tietosuojaa koskevassa vaikutustenarvioinnissa. Julkisen viraston malli puolestaan painottaa julkisen sektorin toimijoiden roolia, toimintatapoja ja vastuun jakoa algoritmien vastuullisessa käytössä. Se kattaa muun muassa hankintakäytännöt, arviointien julkaisemisen, palautteen keräämisen ja järjestelmien auditoinnin. Yleisesti malli korostaa avoimuutta ja vuoropuhelua sidosryhmien, kuten kansalaisten kanssa. AI Now Instituten algoritmisen vaikutustenarvioinnin kehikkoa[250] vuodelta 2018 voi pitää esimerkkinä julkisen viraston mallista.

246 Moss ym., 2021.

247 Ada Lovelace Institute, AI Now Institute ja Open Government Partnership, 2021, 22-23.

248 Institute for the Future of Work, 2021.

249 Kaminski & Malgieri, 2021.

250 Reisman ym., 2018

Samaan aikaan kun tekoälyn sovellusalat ja tutkimustieto algoritmisen syrjinnän riskeistä ovat lisääntyneet[251], käsitys vaikutustenarvioinnin hyödyllisyydestä sääntelyjärjestelmän osana ja sitä täydentävänä hallinnon käytäntönä on vahvistunut[252]. Esimerkiksi Euroopan komission tekoälyä käsittelevä korkean tason asiantuntijaryhmä julkaisi 2020 kesällä luotettavan tekoälyn arviointiluettelon[253]. Myös Euroopan Unionin perusoikeusvirasto (FRA) on kannattanut pakollista perusoikeuksiin keskittyvää vaikutustenarviointia tekoälyjärjestelmille. Se toteutettaisiin ennen järjestelmien käyttöönottoa sekä yksityisellä että julkisella sektorilla ja voitaisiin toistaa säännöllisesti käytön aikana, arvioinnin tulosten ollessa julkisia[254]. Lisäksi Euroopan neuvoston tekoälyn komitea on edistänyt tekoälyn ihmisoikeuksiin keskittyvää arviointikehikkoa yhteistyössä Alan Turing Instituten kanssa.[255] Isossa-Britanniassa samankaltaisia työkaluja on kehittänyt julkisen sektorin käyttöön muun muassa Centre for Data Ethics and Innovation. Globaalilla tasolla UNESCO julkaisi syksyllä 2021 oman julkilausumansa ja suosituksensa tekoälyn eettisestä kehittämiseksi ja hyödyntämiseksi[256] ja valmistelee nyt tekoälyn eettisen vaikutustenarvioinnin kehikkoa ulkoisten asiantuntijoiden johdolla. Myös Euroopan komission huhtikuussa 2021 julkaisema ehdotus uudeksi tekoälyasetukseksi edellyttää tekoälysovellusten kehittäjiltä vaatimustenmukaisuuden arviointia (*conformity assessment*), jolla voidaan nähdä yhteyksiä algoritmiseen vaikutustenarviointiin. Algoritminen vaikutustenarviointi yhdistyy siis luontevasti sekä eurooppalaiseen että kansainväliseen ennakoitavissa olevaa kehitykseen, jossa eettinen tekoäly ja perusoikeuksien huomioiminen pyritään sisällyttämään osaksi digitalisaatiota koskevaa sääntelyä.

Kansallisesta valtionhallinnon perspektiivistä vaikutustenarviointi on myös tärkeä osa lainsäädäntötyötä, kun tarkastellaan eri säädösten mahdollisia vaikutuksia esimerkiksi eri väestöryhmiin osana lainvalmistelua. Lisäksi hallinnon työtä tukee arviointityö, joka on osa säädösten toimeenpanoa. Esimerkki lainsäädäntötyön osana tehtävästä arvioinnista ovat päätösten vaikutusten ennakkoarviointi ja erilaiset ihmisiin kohdistuvien vaikutusten arvioinnit, kuten sukupuolivaikutusten arviointi, jonka toteuttamiseksi esimerkiksi Terveyden ja hyvinvoinnin laitos on laatinut toimintaohjeet[257]. Yleisesti Suomessa osana lainsäädäntötyötä tehtävään vaikutustenarviointiin on olemassa oikeusministeriön

251 Ks. raportin luku 1.1.

252 Ks. esim. Reisman ym., 2018 ja Institute for the Future of Work, 2021.

253 Europpan komissio, AI HLEG, 2020.

254 Euroopan unionin perusoikeusvirasto, 2020..

255 The Alan Turing Institute, 2021.

256 Uutinen UNESCOn yleiskokouksen linjauksista ja suositukset ladattavissa https://en.unesco.org/artificial-intelligence/ethics

257 THL (2021). Sukupuolivaikutusten arviointi. https://thl.fi/fi/web/sukupuolten-tasa-arvo/tasa-arvon-edistaminen/sukupuolivaikutusten-arviointi

laatimat ohjeet vuodelta 2007[258][259]. Säädösvaikutusten arvioitiin ihmisiin kohdistuvien vaikutusten näkökulmasta on olemassa sosiaali- ja terveysministeriön ohjeet vuodelta 2016[260] ja lisäksi sisäministeriö on julkaissut ohjeen yhdenvertaisuusvaikutusten arviointiin säädösvalmistelussa vuonna 2014[261]. Tällä hetkellä oikeusministeriössä on meneillään hanke[262], jossa perus- ja ihmisoikeusvaikutusten arviointia kehitetään säädösvalmistelussa. Tietosuojaa koskeva vaikutustenarviointi on puolestaan esimerkki osana säädösten toimeenpanoa toteutettavasta vaikutustenarvioinnista. Tietosuojavaltuutettu on korostanut, että rekisterinpitäjien on aktiivisesti varmistettava, että käsiteltävät henkilötiedot ja niiden käsittelyyn käytetyt algoritmit eivät johda syrjintään, joten yhdenvertaisuusvaikutukset tulisi jo huomioida tietosuojan vaikutustenarvioinnissa. Tietosuojavaltuutetun toimisto on tuottanut ohjeen ja excel-työkalun tietosuojan vaikutustenarvioinnin tekemistä varten.[263]

Tietosuojaa koskeva vaikutustenarviointityökalu on toiminut osittain innoittajana myös nyt käsillä olevalle tekoälyn syrjiviä vaikutuksia arvioivalle kehikolle yhdessä kansainvälisten, algoritmisten vinoumien vaikutustenarviointiin suunniteltujen työkalujen kanssa. Koska keskeistä tässä hankkeessa kehitetylle kehikolle on erityisesti algoritmisen syrjinnän riskien ehkäisy sekä yhdenvertaisuuden edistämisen tekoälyjärjestelmissä, se kytkeytyy ensisijaisesti yhdenvertaisuuslakiin ja siten yhdenvertaisuusvaikutusten arviointiin. Yhdenvertaisuuden edistämisvelvollisuuden (YVL 5-7 §) perusteella viranomaisten, koulutuksen järjestäjien sekä työnantajien on lain mukaan arvioita yhdenvertaisuuden toteutumista toiminnassaan ja ryhdyttävä tarvittaessa toimenpiteisiin sen edistämiseksi. Tästä johtaa myös velvollisuus arvioida tekoälyjärjestelmien yhdenvertaisuusvaikutuksia niitä käyttöönotettaessa. Arvioinnissa voidaan hyödyntää yhdenvertaisuuden arvioinnin työkalua[264], joka sisältää eri osa-alueet ja perusteet, joita peilataan suhteessa arvioitavaan toimenpiteeseen. Arviointivelvollisten työn tueksi suunniteltu työkalu auttaa tunnistamaan syrjintää sekä toimenpiteitä, joiden avulla toiminnan yhdenvertaisuutta voidaan edistää esimerkiksi työpaikalla.

258 http://urn.fi/URN:ISBN:978-952-466-431-8

259 Esim. Keinänen ja Pajuoja (2020) ovat arvioineet järjestelmän toimivuutta ja esittäneet kehitysehdotuksia eri näkökulmista.

260 http://urn.fi/URN:ISBN:978-952-00-3771-0

261 http://urn.fi/URN:ISBN:978-952-491-963-0

262 https://oikeusministerio.fi/hanke?tunnus=OM020:00/2022

263 Tietosuojavaltuutetun toimisto (2021). Vaikutustenarvioinnin tekeminen. https://tietosuoja.fi/vaikutustenarvioinnin-tekeminen

264 Finlex. Yhdenvertaisuuden arvioinnin työkalu. http://yhdenvertaisuus.finlex.fi/

Aiemmin kehitetyt algoritmiset vaikutustenarviointimallit ovat keskittyneet erityisesti läpinäkyvyyteen ja luotettavuuteen. Syrjimättömyyden näkökulma ja erityisesti yhdenvertaisuuden edistämisen keinot ovat tähän mennessä saaneet vain rajallisesti (ks. seuraava luku). Raportin toisessa osassa kuvailtu Suomen oikeudellinen konteksti antaa mahdollisuuden kytkeä algoritmisen syrjinnän vaikutustenarviointi osaksi laajaa yhdenvertaisuuden edistämisen velvoitetta ja näin istuttaa teknologian hyödyntäminen laajempaan hallinnolliseen käyttöyhteyteen.

Käsillä oleva arviointikehikko pyrkii siis kiinnittymään käynnissä olevaan kansainväliseen keskusteluun algoritmisesta syrjinnästä ja vinoumista, toimimaan algoritmisen vaikutustenarvioinnin kansainvälisten esikuviensa sovelluksena Suomen kontekstissa, asettumaan luontevaksi osaksi hallinnon ohjaustyökaluja sekä osoittamaan niin hallinnon kuin teknologian kehittäjienkin vastuut osana vaikutustenarviointia ja yhdenvertaisuuden edistämistä.

3.2 Menetelmät arviointikehikon muodostamiseksi: vertaileva analyysi ja yhteiskehittäminen

Arviointikehikko syrjimättömälle tekoälylle koostettiin yhdistelemällä vertaileva analyysi ja yhteiskehittäminen. Hankkeen ensimmäinen ja toinen osa esittelivät kehikon kansallisen ja oikeudellisen viitekehyksen. Hankkeen kolmannen osatehtävän alussa puolestaan luotiin katsaus olemassa oleviin kansainvälisiin arviointikehikkoihin erityisesti algoritmisen vaikutustenarvioinnin näkökulmasta. Vertaileva analyysi tarkasteli eri toimijoiden suosittelemia algoritmien vaikutustenarviointikehikkoja ja -menetelmiä tekoälyn eettiseen käyttöön ja syrjimättömyyteen. Tutkijaryhmä kartoitti noin 30 kansainvälisistä esimerkkiä kehikoista. Metodologisesti kartoitus toteutettiin käyttämällä Google-hakukonetta ja seuraavia hakusanoja:

- AI bias + discrimination + fairness + ethics + diversity + inclusion framework
- Algorithmic impact assessment + auditing + accountability framework

Haut toteutettiin ensin tammikuussa 2022 ja myöhemmin toukokuussa 2022 aineiston täydentämiseksi, sillä kyseessä on alati kehittyvä keskustelu ja uusia työkaluja kehitetään aktiivisesti. Kartoitetut arviointikehikot ja työkalut on dokumentoitu raportin liitteessä 4.

Vertailevan analyysin tulokset osoittavat, että olemassa olevat kehikot erosivat painotuksiltaan tämän hankkeen arviointikehkon tavoitteista käsittelemällä erityisesti luotettavuutta ja läpinäkyvyyttä tekoälyjärjestelmien hyödyntämisessä. Otoksen vähemmistön muodostivat kehikot, jotka painottivat syrjinnän näkökulmaa. Näissä kuitenkin

näkökulmana olivat erityisesti algoritmiset vinoumat ja syrjinnän kielto, ei yhdenvertaisuuden edistäminen. Vertailtujen työkalujen julkaisumaissa painottuivat jotkin valtiot; Euroopasta erityisesti Alankomaat ja Iso-Britannia erottuivat edukseen tekoälyn eettisessä soveltamisessa julkisella sektorilla julkaistujen kehikkojen perusteella. Myös Yhdysvalloista dokumentoitiin useita työkaluja, mutta julkinen sektori ei painottunut niissä juurikaan. Nämä painotukset johtunevat osittain englanniksi toteutetusta hausta. Yhteydenpito parhaiden käytäntöjen valtioihin voi kuitenkin tarjota Suomen valtionhallinnolle mahdollisuuden vertaisoppimiseen tekoälyn syrjimättömän käytön varmistamiseksi. Vertailevan arvioinnin perusteella on joka tapauksessa selvää, että toistaiseksi syrjintään ja erityisesti yhdenvertaisuuden edistämiseen on keskitytty vain rajallisesti olemassa olevissa työkaluissa.

Vertailevan arvioinnin pohjalta tutkijaryhmä muodosti alustavan käsityksen kehikon sisältämistä kysymyksistä, formaatista ja visuaalisuudesta. Näiden pohjalta tuloksena oli ns. elinkaarimalliin nojaava formaatti, jossa tiedossa olevista arviointikehikoista poiketen huomioitiin ja osoitettiin myös mahdollisuus yhdenvertaisuuden edistämiseen elinkaaren eri vaiheissa[265].

Hankkeen sidosryhmät osallistuivat kehikon luonnosteluun ja kommentoivat sen sisältöä ja muotoa eri vaiheissa. Hankkeen ohjaus- ja asiantuntijaryhmä[266] kommentoivat ja kehittivät kehikkoa sekä kirjallisesti että suullisesti sitä varten järjestetyissä tilaisuuksissa. Kansainvälisen asiantuntijaryhmän kanssa järjestettiin kaksi tapaamista helmikuussa ja toukokuussa 2022. Hankkeen ohjausryhmä kommentoi työtä niin ikään ohjausryhmäkokouksissa keväällä 2022.

Avoimia sidosryhmätyöpajoja järjestettiin kevään 2022 aikana kaksi, joista molemmat järjestettiin verkossa. Maaliskuussa pidetyn ensimmäisen työpajan osallistujat edustivat erityisesti yksityisiä tekoälyn kehittäjiä ja alan tutkijoita yliopistoista, kun taas jälkimmäinen työpaja huhtikuussa kohdistettiin erityisesti julkisen hallinnon henkilöstölle eri tasoilta ja sektoreilta (mm. ministeriöistä ja virastoista). Molempiin työpajoihin osallistui noin 30 sähköpostitse kutsuttua alan suomalaista asiantuntijaa. Osallistuneita henkilöitä ei kuitenkaan listata erikseen raportissa, eikä esiin nostettuja luottamuksellisia näkökulmia yksilöidä.

265 Ks. luku 3.3.

266 Ks. liite 5 hankkeen kansainvälisestä asiantuntijaryhmästä ja työpajoista.

Sidosryhmätyöpajojen rakenne oli yhteneväinen ja pienryhmäkeskustelujen kysymykset samat molemmissa tilaisuuksissa. Ilmoittautuneille lähetettiin etukäteen luonnos arviointikehikosta, jotta heillä oli halutessaan mahdollisuus tutustua sen sisältöön etukäteen. Alustuksen ja yleisen keskustelun jälkeen pienryhmissä keskityttiin erityisesti kuulemaan eri näkökulmia arviointikehikon käytettävyydestä. Kysymykset olivat:

1. Miten viitekehys olisi käyttäjäystävällinen ja käytännöllinen eri kohderyhmille (virkahenkilöille, tekoälyn kehittäjille ym.)? Mitä olisi huomioitava arviointikehikon muodossa tai käyttöliittymässä? Miten se olisi sinulle hyödyllinen?
2. Miten veisit arviointikehikon käytäntöön kehittäjille ja/tai integroisit sen osaksi valtionhallinnon käytäntöjä? Onko sillä yhteyksiä esimerkiksi vaikutustenarviointiin, hankintalainsäädäntöön tai EU:n tekoälysäädökseen?
3. Mitä arviointikehikon tulisi vielä sisältää? Pitäisikö esim. tarkemmin huomioida eri tekoälysovellusten tai käyttösektorien erot? Mitä olisi vielä tuotava esiin suunnittelussa, kehityksessä ja käyttöönottossa?

Työpajoissa käydyt keskustelut dokumentoitiin ottamalla yleisestä keskusteluista yksityiskohtaiset muistiinpanot, kirjoittamalla pienryhmäkeskustelussa esiin nousseet näkökulmat ryhmäkohtaisille kalvoille sekä tallentamalla chat-kommentit. Työpajakeskustelujen dokumentaatiota on hyödynnetty hankkeen loppuraportin laatimisessa, arviointikehikon kysymysten tarkennuksessa ja rajauksessa sekä politiikkasuositusten laatimisessa.

Sidosryhmätyöpajoissa keskusteltiin laajasti siitä, millaiset arvointikehikon kysymykset ja ohjeet palvelisivat eri käyttäjäkuntia ja miten sen sisällön jäsentäminen ohjaisi käyttäjää pohtimaan yhdenvertaisuuskysymyksiä omasta näkökulmastaan. Keskusteluissa korostui erityisesti kehikon käytettävyys ja kontekstisidonnaisuus. Varsinkin yhteistyö ja roolien jako eri organisaatioiden välillä, kuten esimerkiksi tekoälyjärjestelmän julkisen tilaajan ja yksityisten kehittäjien välillä nousi toistuvasti esiin. Myös arviointikehikon ymmärrettävyys eri käyttäjäryhmille nousi esiin ja tämän perusteella lopulliseen versioon lisättiin auttavaa ohjeistusta, asiasanasto ja esimerkkejä käytöstä. Tämän ohella nousi esiin arviointikehikon yhteys hankintalainsäädäntöön. Tämä vaikutti hankkeen kohderyhmän kirkastamiseen erityisesti hallinnon toimijoiden näkökulmaa painottavaan suuntaan. Toisaalta keskusteltiin myös siitä, ettei kehikon tulisi olla byrokraattinen rasite, vaan myös tekoälyn kehittäjiä palveleva helppokäyttöinen työkalu sovellusten yhdenvertaisuuden edistämiseen.

Kahden sidosryhmätyöpajan ja asiantuntijaryhmätilaisuuden sekä ohjausryhmässä käytyjen keskustelujen jälkeen kehikko viimeisteltiin palautteen perusteella. Lisäksi joitakin kehittäjäyhteisön edustajia lähestyttiin vielä arviointikehikon käyttökohteiden ja esimerkkien tarkentamiseksi. Myös ohjausryhmällä oli mahdollisuus esittää näkemyksiä samasta versiosta yhden viikon ajan. Tämän jälkeen hankkeen tutkimusryhmä viimeisteli Demos

Helsingin johdolla kehikon. Ratkaisu julkisen sektorin painottamiseen pohjautuu paitsi yhden hankkeen rajallisiin mahdollisuuksiin huomioida käyttäjäryhmät eri sektoreilla, toimialoilla ja teknologian kehittämisessä, myös alkuperäisen tietotarpeen tavoitteisiin.

3.3 Suomalainen arviointikehikko syrjimättömille tekoälysovelluksille

Tämä luku esittelee varsinaisen arviointikehikon syrjimättömille tekoälyjärjestelmille, sen käyttökohteet ja -tilanteet sekä ohjeistuksen kehikon käyttämiseen. Kehikko tukee tekoälysovellusten syrjintä- ja yhdenvertaisuusvaikutusten arviointia erityisesti julkisen sektorin kontekstissa. Yhdenvertaisuuslain 5 § sekä perustuslain perusteella viranomaisen ei tule hyödyntää tekoälyä tai automaattista päätöksentekoa tavalla, joka loukkaisi ihmisten perusoikeuksia tai olisi ristiriidassa yhdenvertaisuuden edistämisvelvoitteen kanssa. Kehikolla on yhteneväisyyksiä myös tietosuojaa koskevaan vaikutustenarviointiin, täydentäen sen automaattisen päätöksenteon osuutta. Tekoälyn kohdalla vaikutustenarviointi edistää myös järjestelmän läpinäkyvyyttä, luotettavuutta, sidosryhmien osallistumista sekä selkeää vastuun jakautumista järjestelmän käytöstä. Lisäksi arviointikehikko tukee yleistä riskienhallintaa tekoälyn kontekstissa korostamalla vastuullisia käytäntöjä ja mekanismeja, kuten moninaisuuden lisäämistä kehityksessä.

Arviointikehikko antaa myös teknisempiä ohjeita ja avaa parhaita käytäntöjä, joilla voidaan kitkeä syrjiviä algoritmisia vinoumia läpi tekoälyn elinkaaren. Keskeisenä erona aiempiin vastaaviin julkaisuihin tekoälyn vastuullisesta kehittämisestä, arviointikehikko keskittyy ensisijaisesti algoritmiseen syrjintään sekä keinoihin edistää yhdenvertaisuutta tekoälyjärjestelmien käytössä.

3.3.1 Kohderyhmä ja käyttäjät

Arviontikehikon pääasiallisena kohderyhmänä ovat **julkiset organisaatiot ja viranhaltijat,** jotka voivat käyttää kehikkoa tunnistaessaan ja arvioidessaan tekoälyjärjestelmien mahdollisia syrjiviä vaikutuksia läpi niiden elinkaaren suunnittelusta ja kehityksestä käyttöönottoon. Lisäksi se auttaa myös **tekoälyn ja palveluiden kehittäjiä** omien tekoälyjärjestelmiensä ja prosessien syrjivien vaikutusten arvioinnissa, välttämisessä sekä yhdenvertaisuuden edistämisessä, varsinkin kun kyse on julkisten palvelujen tuottamisesta. Täten se toimii joustavana itsearviointityökaluna sekä tekoälyjärjestelmiä tilaaville ja kehittäville julkisille toimijoille, että niiden toimittajille. Arviointikehikon kohderyhmät voidaan laajemmin jaotella seuraavasti:

1. Julkisen sektorin toimijat, jotka työssään vastaavat esimerkiksi palveluntarjoajien valinnasta, päättävät tekoälyn hyödyntämisestä omalla hallinnonalallaan tai osallistuvat palvelujen digitalisoinnin kehittämiseen
2. Tekoälyn kehittäjät, datatieteilijät ja ohjelmoijat
3. Tutkijat, jotka arvioivat tekoälyn yhteiskunnallisia vaikutuksia ja näihin liittyviä työkaluja
4. Kansalaiset, joiden arkielämään tekoälysovellukset vaikuttavat

Julkisella sektorilla arviointikehikko on suunnattu valtio-, alue- ja kuntatason organisaatioiden käytettäväksi tilanteissa, joissa suunnitellaan, hankitaan tai otetaan käyttöön tekoälysovelluksia tai automatisoidaan palvelun tai prosessin osaa. Kehikko luo ymmärrystä algoritmipohjaisten sovellusten mahdollisista syrjivistä vaikutuksista sekä keinoista vähentää niihin liittyviä riskejä. Sitä voidaan soveltaa monenlaisiin käyttötilanteisiin. Kehikon käyttäminen on erityisen tärkeää tilanteissa, joissa:

- Tekoälyjärjestelmä vaikuttaa välittömästi tai välillisesti kansalaisten perusoikeuksiin (esim. terveys, asuminen, luotonanto, pääsy koulutukseen, poliittinen osallisuus)
- Sovelluksen kohteena on haavoittuvia ryhmiä ja vähemmistöjä tai sen opetusdata sisältää ihmisten henkilökohtaisia ominaisuuksia ja käyttäytymistä (esim. koulutustieto)
- Tekoälysovellukseen liittyy puoliautomaattista, automaattista tai täysin automatisoitua päätöksentekoa (vrt. tiukan valvonnan tekoälysovellukset)

Arviointikehikko toimii parhaiten tilaajan ja kehittäjän välisenä vuoropuheluna, mutta sitä on mahdollista käyttää myös itsenäisesti sovelluksen hankinnan ja kehittämisen eri vaiheissa. Sekä tekoälyjärjestelmien tilaajien että kehittäjien on huomioitava eri operatiiviset tasot omien organisaatioidensa sisällä: yhteistyö muun muassa ylimmän johdon, asiantuntijoiden, lakiosaston ja viestinnän välillä on tärkeää läpi tekoälyn elinkaaren (kehittäjäpuolella myös mm. tuotekehitys, myynti, mainonta). Siksi kehikko pyrkii osaltaan huomioimaan myös organisaatiotason laajemmat prosessit ja niiden resurssoinnin. Näin se voi myös edistää jaettua ymmärrystä kunkin osaston rooleista ja vastuualueista tekoälyjärjestelmien syrjintäriskien hallinnassa.

3.3.2 Arviointikehikon elinkaarimalli

Arviointikehikko pohjaa tekoälyn elinkaarimalliin, joka korostaa, että tekoälyjärjestelmien syrjivät vaikutukset voivat syntyä ja ilmentyä useassa eri vaiheessa: mm. kyseenalaisten suunnitteluperusteiden takia, datan epäedustavuuden tai käyttöönotossa tarkoituksenmukaisemattomassa ympäristössä. Elinkaarimalli mahdollistaa kokonaisvaltaisen syrjinnän

riskien hallinnan, sen sijaan, että järjestelmää arvioitaisiin vain joillain tietyillä, muuttumattomilla perusteilla, jotka vanhenevat nopeasti teknologian kehittyessä. Näin arviointikehikko toimii vaikutustenarviointiprosessina syrjintäriskeihin varatumisessa sekä yhdenvertaisuuden edistämisessä läpi tekoälyjärjestelmän elinkaaren.

Kehikko ja sen sisältämät kysymykset on jaoteltu elinkaarimallin mukaisesti kolmeen vaiheeseen:

1. **Suunnittelu**, eli tekoälyjärjestelmän tavoitteiden, käytön motiivien, tarpeellisuuden ja yhdenvertaisuusvaikutusten alustava arviointi ja määrittely. Suunnittelun merkitys korostuu julkisen toimijan tilatessa tai ottaessa käyttöön tekoälysovelluksen, jonka kehitysprosessiin sillä on vain rajallinen pääsy. Yhdenvertaisuuden ja sen edistämisen näkökulmasta suunnittelussa on huomioitava erityisesti järjestelmän tavoitteet, haavoittuvien sidosryhmien osallistuminen ja sovelluksen vaikutukset suhteessa vallitsevaan yhteiskunnalliseen eriarvoisuuteen.
2. **Kehitys**, joka kattaa kolme osa-aluetta: data ja sen esivalmistelu, algoritmisen mallin opetus sekä mallin validointi. Olemassa olevaa eriarvoisuutta heijastava opetusdata on yksi algoritmisen syrjinnän tärkeimmistä tekijöistä. Samaten algoritmiset vinoumat voivat johtaa syrjiviin lopputuloksiin mallin opetusvaiheessa. Lisäksi syrjintää voi ilmetä mallin testaus- ja validointivaiheessa, tarkasteltaessa toimiiko se tarkoitetulla tavalla. Kehitysvaiheen kohdissa voidaan soveltaa teknisiä esi-, malli- ja jälkikäsittelymenetelmiä sekä reiluusmetriikoita algoritmisten vinoumien oikomiseeen.
3. **Käyttöönotto**, jossa tärkeää on muun muassa järjestelmän ihmisohjauksen, läpinäkyvyyden ja seuranta-mekanismien varmistaminen. Tekoälysovelluksen luotettavuus ja päätösten tarkkuus saattaa horjua, johtaen syrjintään, jos se otetaan käyttöön eri ympäristössä tai kohderyhmässä kuin mihin se on alunperin suunniteltu. Huomionarvoista on, että tekoälyn elinkaari ei lopu käyttöönottoon, vaan sen toiminnan seuranta- ja ylläpito informoi aiempia kohtia iteratiivisesti järjestelmää jatkokehitettäessä.

Yksityisen kehittäjän tai palveluntarjoajan perspektiivistä elinkaarimallin voi samoin ajatella koostuvan a) tekoälyn tuote- ja palvelukonseptin suunnittelusta, sidosryhmien tunnistamisesta, ongelman- ja tavoitteiden määrittelystä; b) teknisestä kehityksestä, arvioinnista ja testaamisesta; sekä (c) järjestelmän käyttöönotosta eri osapuolien toimesta, sen markkinoinnista, vaikutustenseurannasta, kunnossapidosta ja lopulta tuotteen poistamisesta käytöstä.

3.3.3 Arviointikehikon käyttäminen

Jokainen tekoälyjärjestelmän elinkaaren vaihe on jaettu arviointikehikossa kysymyksiin, joiden tarkoituksena on kuvata tekoälyjärjestelmän eri vaiheisiin liittyviä syrjintäriskejä sekä mahdollisuuksia niiden minimoimiseen ja yhdenvertaisuuden edistämiseen. Kysymyksiin vastataan arvioimalla onko syrjintäriskit tai yhdenvertaisuuden edistäminen huomioitu sovelluksessa hyvin, osittain tai ei ollenkaan. Tämän perusteella järjestelmälle kertyy riskipisteitä, jotka ohjaavat sen käytössä. Osa aiheista ja kysymyksistä on merkitty yhdenvertaisuuden kannalta välttämättömiksi kohdiksi, jotka on huomioitava edes osittain kehityksessä. Koska tekoälyä voidaan soveltaa laajasti eri konteksteihin, arviointikehikko voidaan räätälöidä vastaamaan kehitettävää tai hankittavaa järjestelmää. Tämän vuoksi käyttäjä voi halutessaan lisätä kysymyksille painotuksia, määritellä eri asiat välttämättömiksi tai tarvittaessa jättää vastaamatta sovellukselle epärelevantteihin kysymyksiin. Jotta järjestelmän aiheuttamien riskien luonnetta voidaan ymmärtää, kehikon kaikki kysymykset suositellaan kuitenkin käymään läpi joka tapauksessa. Lopuksi kehikon läpikäymisellä voidaan arvioida onko sovelluksen kanssa suhteellisen turvallista edetä, pitäisikö käyttöä vielä harkita vai liittyykö sovellukseen niin perustavanlaatuisia syrjintäriskejä, ettei sen käyttöä voida suositella ilman muutoksia.

On oletettavaa, että kysymyksiin vastaaminen vaatii toisinaan lisäselvitystä sekä tilaajan ja kehittäjän välistä vuoropuhelua. Kehikko on rakennettu siten, että tilaajan ja kehittäjän näkökulmat painottuvat eri vaiheissa riippuen siitä, kenen asiantuntemus ja vastuu mahdollistaa syrjimättömyyttä tukevien valintojen tekemisen. Parhaiten arviointikehikon käyttäminen onnistuu, jos yhteistyö tilaajan ja kehittäjän välillä käynnistyy jo suunnitteluvaiheessa, jotta kehittäjällä on riittävä näkyvyys aiotun sovellusalan erityispiirteisiin. Sitä voidaan soveltaa kuitenkin myös valmiiden sovellusten arviointiin.

Arviointikehikkoa voidaan käyttää useaan eri käyttötarkoitukseen ja kontekstiin:

- **Hankintakäytäntöjen kehittämiseen sekä hankinnan ohjaamiseen**, niin että hankittavan sovelluksen tai järjestelmän syrjintäriskit on minimoitu.
- **Roolien ja vastuiden selkeyttämiseen:** Arviointikehikkoa voidaan käyttää työkaluna eri toimijoiden roolien ja vastuiden selkeyttämiseen syrjintäriskien minimoimisessa sekä organisaatioiden sisäisesti, että niiden välillä. Projektipäälliköllä, datatietelijällä, algoritmien kehittäjällä, toimialueen asiantuntijalla, organisaation lakiosastolla, henkilöstöhallinnolla, johdolla, myynnillä ja viestinnällä on erilaiset mutta tärkeät roolit syrjintäriskit minimoimisessa, ja nämä vaihtelevat tilanteittain.
- **Koulutukseen ja tietoisuuden lisäämiseen:** Arviointikehikon käyttäminen ja siitä viestiminen lisää tietoisuutta ja ymmärrystä tekoälyn syrjintäriskeistä ja sitä voidaan hyödyntää julkisen sektorin työntekijöiden koulutuksessa ja

osaamisen kehittämisessä sekä yhteisten käytäntöjen luomisessa ja hyvän hallinnon periaatteiden vahvistamisessa.

- **Luottamuksen vahvistamiseen**: Kun vastuut ovat sovelluksen elinkaaren ajalta selkeitä, voidaan aktuaalisia syrjintäriskejä minimoida sekä viestiä sitoutumisesta eettisten tekoälyjärjestelmien kehittämiseen. Tämä mahdollistaa luottamuksen eri toimijoiden välillä sekä vahvistaa kansalaisten ja käyttäjien luottamusta julkisiin palveluihin ja julkiseen hallintoon.
- **Virheisiin vastaamiseen ja niiden korjaamiseen**: Syrjintäriskien realisoituessa arviointikehikko auttaa sekä julkista hallintoa että kehittäjiä tunnistamaan syrjinnän lähteen järjestelmässä, mahdollistaen virheen korjaamisen.
- **Yhdenvertaisuuden edistämisen toimenpiteiden tunnistamiseen** tekoälysovellusten kehittämisessä ja käyttöönottamisessa

Näin Arviointikehikkoa käytetään:

1. **Tutustu kehikkoon ja sovella se käyttökontekstiin sopivaksi**

Arviointikehikon on tarkoitus olla proaktiivista riskienhallintaa ja yhdenvertaisuuden huomiointia edesauttava työkalu. Se on toisin sanoen esimerkkinä toimiva malli tekoälysovellusten yhdenvertaisuusvaikutusten arviointia varten. Kehikon avulla laskettava pistemäärä on tarkoitettu tulkittavaksi suuntaa antavana viitearvona, ei lopullisena kvantitatiivisena totuutena tarkastellun järjestelmän syrjintäriskeistä ja odotettavista yhdenvertaisuusvaikutuksista. Kannustamme käyttämään kehikkoa kuhunkin käyttökontekstiin soveltuvin osin ja painottamme monialaisen yhteistyön tärkeyttä kehikon kysymysten pohdinnassa.

Suosittelemme kehikon kysymysten läpikäymistä, niiden merkityksen tunnistamista kyseisessä käyttökontekstissa sekä tarvittaessa painotusta jo ennen käyttöönottoa. Osaa kysymyksistä on tärkeä arvioida systemaattisesti, kun taas osa voi ennemminkin nostaa esille asioita ja näkökulmia, joihin voidaan kiinnittää huomiota riippuen tapauksesta. Joukossa voi olla myös kysymyksiä, jotka eivät ole merkityksellisiä kaikissa konteksteissa, jos ne on esimerkiksi jo huomioitu osana muita tietojärjestelmiin liittyviä velvoitteita tai mikäli sovelluksella ei ole välittömiä oikeudellisia vaikutuksia luonnollisiin henkilöihin. Myös esimerkiksi se, käytetäänkö sovelluksessa henkilötietoja vai ei, on keskeistä arvioinnille.

Kehikon soveltamiseen vaikuttaa myös se, käytetäänkö arviointikehikkoa valmiin vaiko vielä kehitettävän sovellusten arviointiin. Vaikka kyseessä olisi valmis tekoälytuote- tai palvelu, kannustamme silti koko elinkaareen liittyvien kysymysten läpikäymiseen, jolloin painopiste voi olla kysymyksissä, jotka liittyvät käytön suunnitteluun sekä käytön ja sen vaikutusten monitorointiin.

2. **Suunnittele kehikon käyttöprosessi**

Kehikon käyttämisessä korostuu keskustelu ja yhteistyö keskeisten sidosryhmien välillä (tilaaja, kehittäjä jne.) ja kysymyksistä oppiminen. Kehitys- ja suunnitteluprosessin (mukaan lukien kehikon käytön) ja tulosten dokumentointi ja hyödyntäminen ovat myös keskiössä. Kehikon käyttöä suunniteltaessa on syytä kiinnittää huomiota: a) toimijoihin, joiden yhteistyönä kehikkoa käytetään ja luoda prosessi, joka sopeutuu organisaation kulttuuriin ja prosesseihin sekä mahdollistaa yhteistyön ja keskustelun sen aikana; b) miten riskeihin varaudutaan, miten esiintulevia mahdollisia ongelmia tuodaan esiin, käsitellään ja lähdetään korjaamaan.

3. **Varmista riittävä ymmärrys ja osaaminen**

Kontekstista riippumatta arviointikehikko sisältää kysymyksiä, jotka vaativat teknistä asiantuntemusta, jolloin käyttäjien on varmistettava, että saatavilla on riittävä ymmärrys ja osaaminen esimerkiksi tilastollisten analyysien ja muiden vaikutuksenarviointimenetelmien hyödyntämiseen mahdollisesti syrjivien vinoumien tunnistamisessa. Lisäksi arviointikehikon kysymyksiin vastaaminen edellyttää riittävää ymmärrystä järjestelmän käyttökontekstista sekä syrjintään ja yhdenvertaisuuteen liittyvistä kysymyksistä ja velvollisuuksista. Näihin molempiin tämä raportti antaa myös eväitä.

4. **Vastaa kysymyksiin relevanttien toimijoiden kanssa keskustellen**

Kun kysymyksiin on tutustuttu ja niitä on tarvittaessa priorisoitu ja painotettu, arviointikehikon kysymyksiin vastataan huolellisesti. Yleensä tämä tarkoittaa julkisen tilaajan ja kehittäjän välistä vuoropuhelua sekä tiimien sisäistä vastuun jakoa. Ensimmäisen vaiheen kysymyksiin voidaan vastata heti, kun tekoälyjärjestelmän sovellusala on selvillä, ja sen kriteerejä voidaan soveltaa myös teknisen kehittäjän valintaan sekä täydentää yhdessä kehittäjän kanssa, kun toteuttaja on valittu.

5. **Dokumentoi tulokset ja tunnista tarvittavat, korjaavat toimenpiteet**
Arviointikehikon tulosten tulkitseminen riippuu käyttökontekstista: jos julkinen organisaatio on käyttänyt sitä eri sovelluskehittäjien arviointiin, tulos voi auttaa valitsemaan tilattavan sovelluksen tai kehittäjän, joka on parhaiten ottanut huomioon yhdenvertaisuusnäkökulman. Jos taas kyseessä on jo käytössä oleva sovellus, voidaan tulosten avulla palata riskitekijöitä koskeviin kysymyksiin ja suunnitella korjaavat toimenpiteet. Toimenpiteiden valitsemisessa on tärkeää myös tunnistaa mihin ja miten kehikon tuloksia pitäisi viestiä sekä millaista yhteistyötä riskien minimointi tai yhdenvertaisuuden edistäminen edellyttää kussakin yhteiskunnallisessa käyttökontekstissa.

Useat eri organisaatiot voivat käyttää arviointikehikkoa omiin tarpeisiinsa. Alla kolme esimerkkiä arviointikehikon käyttämisesestä julkisen sektorin kontekstissa:

Käyttöesimerkit:

1. **Kaupunki kehittää palvelujärjestelmää ja hankkii tekoälyyn pohjautuvan sovelluksen tukemaan erityisesti vammaisten henkilöiden palveluohjausta.**
Kaupunki käynnistää hankinnan ja edellyttää arviointikehikon käyttämistä osana hankintakriteerejä. Tilaajaorganisaatio, eli kaupunki käy läpi palveluntarjoajien kanssa kehikon relevantit kysymykset joissa suunnittelua arvioidaan retrospektiivisesti. Arviointikehikon tulos vaikuttaa lopullisen palvelun valintaan. Valittavan palveluntarjoajan kanssa sovitaan mahdollisista korjaavista toimenpiteistä ennen sovelluksen hankkimista ja käyttöönottamista: esimerkiksi jos prosessin aikana on todettu että vammaisten henkilöiden oikeutta kohtuullisiin mukautuksiin ei ole riittävästi huomioitu, palveluntarjoajaa vaaditaan huomioimaan tämä sovelluksen toimittamisessa.

Lopputuloksena tilaajaorganisaatio on varmistanut, että hankittava sovellus ei sisällä merkittäviä riskejä yhdenvertaisuuden näkökulmasta ja se mahdollistaa yhdenvertaisuuden edistämisen vammaisten palveluissa. Jos sovellukseen liittyy riskejä, ne on tunnistettu ja minimoitu muu soveltuva lainsäädäntö (esim. tietosuojaa koskevat säädökset) huomioiden jo ennen sovelluksen hankkimista ja käyttöönottamista. Tilaajaorganisaatio viestii tästä päätöksen yhteydessä sekä opastaa sovelluksen käyttäjiä myöhemmin käyttöönotossa.

2. **Valtion virasto tai laitos, esimerkiksi Kela, kehittää koneoppimiseen pohjautuvan päätöksentekojärjestelmän sosiaalietuushakemusten yhdenvertaisempaan käsittelyyn oman ICT-tiiminsä avulla.**

Arviointikehikko otetaan heti osaksi sovelluksen suunnittelua ja kehittämistä, alkaen järjestelmän hyväksyttävän tavoitteen asettelusta ja keinojen oikeasuhtaisuuden arvioinnista. Kaikki kysymykset käydään läpi soveltaen kukin kysymys tekoälysovelluksen käyttökontekstiin. Järjestelmän yhdenvertaisuusvaikutuksia pyritään parantamaan opettamalla malli siten, että sen tuottamien ennusteiden virhetasot ovat yhtäläiset läpi vaikutuksenalaan kuuluvien väestöryhmien. Järjestelmän käyttöön ja valvontaan nimetään vastuutiimi, joka koostuu kehittäjistä, henkilöstöhallinnosta sekä sovelluksen tulevista käyttäjistä. Kehityksen aikana huomataan, että malli on vinoutunut siten, että se tuottaa keskimäärin useammin virheitä Lähi-idästä Suomeen muuttaneiden ihmisten kohdalla. Tiimi päättelee, että erot virhetasoissa johtuvat siitä, että kyseiseen ryhmään kuuluvien henkilöiden työhistoriasta on saatavilla vähemmän tietoa. Virhetasojen eroja pienennetään asettamalla mallille optimointirajoituksia opetusvaiheessa, minkä kehittäjät huomaavat heikentävän mallin yleistä osumatarkkuutta muutamalla prosenttiyksiköllä. Järjestelmän käyttäjille ilmoitetaan tarvittavat tiedot, jotta he osaavat arvioida järjestelmän tuottamia ennusteita asianmukaisesti.

Käyttämällä kehikkoa virasto kuitenkin varmistaa, että sovellus ei syrji asiakkaita heidän alkuperänsä perusteella. Samalla virasto kasvattaa työntekijöiden kyvykkyyksiä ymmärtää ja arvioida syrjintäperusteisiä riskejä tekoälyn käytössä sekä voi nostaa teemaa esille yleisessä keskustelussa hallinnonalan sisällä. Mahdollisiin syrjintäriskeihin osataan varautua ennalta, mikä helpottaa vastaavien sovellusten kehittämistä tulevaisuudessa.

Esimerkki 3: Ministeriö, esimerkiksi valtiovarainministeriö integroi arviointikehikon oman hallinnonalansa ohjaukseen tekoälysovellusten käytössä ja velvoittaa virastoja toteuttamaan arvioinnin ennen käyttöönottoa.

Arviointikehikon käyttö sisällytetään tulosohjauksen kehittämiseen ja täten ministeriön ja viraston välille laadittavaan tulossopimukseen. Näin ministeriö voi edellyttää omalla hallinnonalallaan kehikon esitäyttämistä tekoälypohjaisten sovellusten käyttöönottoa suunniteltaessa. Lisäksi ministeriö hyödyntää sitä virkahenkilöiden koulutuksessa yhteistyössä koulutusta tarjoavien tahojen kanssa sekä nimeää vastuuhenkilön seuraamaan ja ohjaamaan yhdenvertaisuuden edistämistä hallinnonalan tekoälysovelluksissa. Asiasta luodaan konkreettinen ohjeistus sekä virastoille, mutta myös muiden ministeriöiden hyödynnettäväksi.

Tällöin kehikkoa käytetään kokonaisuutena lisäämään hallinnonalan valmiuksia ja kyvykkyyksiä tunnistaa ja minimoida syrjintäperusteita riskejä, joita tekoälysovelluksiin liittyy. Arviointikehikon hyödyntämisen tuloksena huomataan, että palveluiden yhdenvertaisuuden edistäminen on yhdistettävissä niiden kustannustehokkuuteen.

Arviointikehikko on sovelletun teknologian, datan ja algoritmien suhteen agnostinen. Tämä tarkoittaa, että sitä ei ole suunniteltu vain jotakin tiettyä tekoälyteknologiaa (esim. suosittelujärjestelmä tai konenäkö) tai mallia varten, vaan se pyrkii olemaan sovellettavissa laajasti eri tekoälyjärjestelmien suunnittelu-, kehitys- ja käyttöönottoprosesseihin. Kehikkoa ei ole myöskään tarkoitettua vain jollekin tietylle alalle tai sektorille, vaan se tarjoaa peruslähtökohdat syrjinnän arvioinnille sovellusalasta riippumatta. Todettakoon kuitenkin, että arviointikehikon kehitystä ohjanneena tyyppiesimerkkinä voidaan pitää koneoppimista soveltavaa automatisoitua algoritmista päätöksentekoa ihmisten perusoikeuksiin liittyvissä asioissa, kuten sosiaalietuuksien jaossa tai sisäänottoprosesseissa. Huomionarvoista on myös, että vaikka arviointikehikko pyrkii kuvaamaan riskejä syrjivien vinoumien synnylle eri elinkaaren vaiheissa, se ei suoranaisesti kerro riskien vakavuudesta tai todennäköisyydestä, jotka vaihtelevat käyttökontekstin mukaan.

Arviointikehikon pisteytys:

Arviointikehikon aihealueisiin ja kysymyksiin vastataan arvioimalla, miten kyseinen asia on huomioitu sovelluksen elinkaaren vaiheessa. Vastauksista kertyy riskipisteitä. Ne pisteytetään seuraavanlaisesti:

Hyvin: 0 pistettä

Kysymys on huomioitu vahvasti osana järjestelmän kehitystyötä (esimerkiksi toteuttamalla korjaustoimenpiteitä, ottamalla käyttöön soveltuvia prosesseja tai määrämäällä henkilöitä vastuuseen kyseisestä aihealueesta).

Osittain: 0,5 pistettä

Kysymys on otettu huomioon jossain määrin tai kohtuullisesti (esimerkiksi suunnittelemalla toimenpiteitä, perehtymällä aiheeseen tai keskustelemalla siitä organisaatiossa tai kehitystiimissä).

Ei huomioitu: 1 piste

Kysymykseen ei ole kiinnitetty juuri lainkaan huomiota.

Jokaisen aihealueen lopussa on kolme yhdenvertaisuuden edistämiseen liittyvää kysymystä. Ne pisteytetään päinvastaisesti, eli niiden huomioiminen vähentää riskipisteitä. Toisin sanoen esimerkiksi edistämisvelvoitteen hyvä huomiointi osana suunnittelua vähentää yhden (-1) riskipisteen.

Vastauskentässä oleva ▲ -symboli tarkoittaa yhdenvertaisuuden kannalta välttämätöntä tai erittäin tärkeää kohtaa. Nämä kohdat on huomioitava vähintään kohtuullisesti ennen etenemistä, eli niitä ei voi jättää huomiotta.

Kaikki arviointikehikon kysymykset eivät välttämättä sovi jokaisen tekoälyjärjestelmän arviointiin. Näissä tapauksissa kysymykseen voi jättää vastaamatta vaihtoehdolla N/A.

Jokaisesta sovelluksen elinkaaren vaiheesta lasketaan vastausten perusteella **osapisteet**. Aihealueiden osapisteet vaihtelevat 12 ja 17 välillä riippuen kysymysten määrästä, ollen keskimäärin x/15 (alla esimerkki). Pisteet auttavat arvioimaan kyseisen vaiheen syrjintäriskejä:

0-3,5 pistettä/osavaihe: Suhteellisen turvallista edetä seuraavaan vaiheeseen.

4-8 pistettä/osavaihe: Harkitse seuraavaan vaiheeseen etenemistä ellei korvaavia toimenpiteitä tunnisteta ja toteuteta.

8,5-15 pistettä/osavaihe: Älä etene seuraavaan vaiheeseen ennen muutoksia.

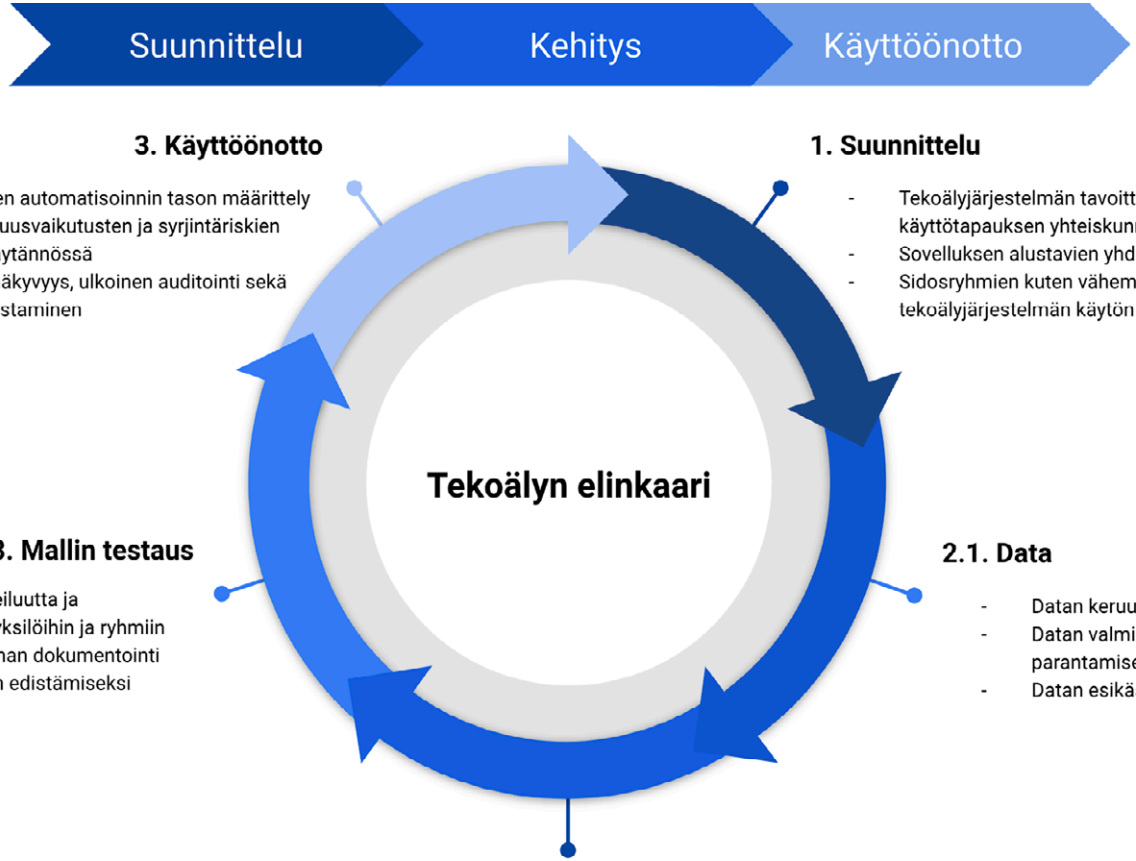
Kun kaikki vaiheet on käyty läpi, kehikon avulla lasketaan kokonaispisteet (x/73). **Kokonaispisteet** eivät saa olla punaisella (39-73 pistettä), jotta sovellusta voidaan ottaa käyttöön ennen korjaavia toimenpiteitä. Pisteytyksessä oletetaan, että kaikki kysymyksiin vastataan.

0-18,5: Sovelluksen käytön kanssa on suhteellisen turvallista edetä.

19-38,5: Harkitse käyttöä ellei korjaavia toimenpiteitä toteuteta. Jos korjaavia toimenpiteitä ei tehdä, mutta käyttöönottoa jatketaan, on syytä käydä huolellisesti läpi, miten sovelluksen syrjintäriskeihin varaudutaan.

39-72: Sovellus sisältää merkittäviä syrjintäriskejä eikä sen käyttöön ottamista voida suositella ilman korjaavia toimenpiteitä. Tuloksen syy on analysoitava ja keskusteltava osapuolten kanssa, millaisilla toimenpiteillä syrjintäriskit minimoidaan.

Seuraavilla vaakamuotoisilla sivuilla esitetään itse arviointikehikko kysymyksineen. Se on saatavilla tämän raportin yhteydessä myös helppokäyttöisessä Excel-muodossa.



1. Suunnittelu

- Tekoälyjärjestelmän tavoitteiden määrittely ottaen huomioon käyttötapauksen yhteiskunnallisen ja oikeudellisen kontekstin
- Sovelluksen alustavien yhdenvertaisuusvaikutusten arvioiminen
- Sidosryhmien kuten vähemmistöjen osallistuminen tekoälyjärjestelmän käytön määrittelyyn

2.1. Data

- Datan keruu, alkuperän ja laadun varmistaminen
- Datan valmistelu ja validointi edustavuuden parantamiseksi
- Datan esikäsittely vinoumien vähentämiseksi

2.2. Algoritmisen mallin opetus

- Mallin reiluuden arviointi reiluumetriikoin ja mallin ominaisuuksien valinta
- Moninaisuus kehityksessä ja yhteiskehittäminen sidosryhmien kanssa
- Mallin käsittely, kuten regularisointi ja optimointirajoitukset vinoumien oikomiseksi

2.3. Mallin testaus

- Testataan ja arvioidaan mallin reiluutta ja yhdenvertaisuutta suhteessa eri yksilöihin ja ryhmiin
- Mallin ominaisuuksien ja toiminnan dokumentointi
- Jälkikäsittely yhdenvertaisuuden edistämiseksi

3. Käyttöönotto

- Käytön skaalan ja päätösten automatisoinnin tason määrittely
- Järjestelmän yhdenvertaisuusvaikutusten ja syrjintäriskien seuranta ja monitorointi käytännössä
- Järjestelmän ylläpito, läpinäkyvyys, ulkoinen auditointi sekä oikeussuojakeinojen varmistaminen

1. SUUNNITTELU

Suunnittelu on yksi tärkeimmistä vaiheista tekoälyjärjestelmien yhdenvertaisuuden ja syrjimättömyyden näkökulmasta, erityisesti julkisen sektorin hankinnoissa, joissa ei ole suoraa pääsyä tekoälyn kehitysprosessiin. Tässä vaiheessa määritellään järjestelmän perusteet, tavoitteet ja käytön motiivit, jotta sen käyttö olisi oikeasuhtaista yhteiskunnalliseen kontekstiin ja väestöön.

Vastuu: Ensisijainen vastuu suunnittelusta on tekoälyjärjestelmän tilaavalla ja käyttöönottavalla organisaatiolla, edellyttäen kuitenkin myös yhteistyötä kehittäjien kanssa.

KONSEPTI JA ESISUUNNITTELU		
Aihe	**Kysymykset**	**Huomioitu**
Perusteet järjestelmän kehittämiselle	Ovatko tekoälyjärjestelmän kehittämisen tavoitteet ja perusteet arvioitu yleisen edun mukaisiksi sekä yleisesti hyväksyttäviksi? Mihin yhteiskunnalliseen ongelmaan se pyrkii vastaamaan?	▲
Järjestelmän käytön tavoitteet	Onko tekoälyjärjestelmän käytön tavoitteet ja päämäärät määritelty selkeästi, sisältäen ongelman muotoilun, toiminnan, käyttökohteen, käyttäjät ja henkilöt, joihin sovellus vaikuttaa?	
Sovelluksen tarpeellisuus	Onko suunniteltu järjestelmä tai tekoälyn käyttötapa välttämätön ja oikeasuhteinen asetetun tavoitteen saavuttamiseksi?	
Yhteiskunnallinen käyttökonteksti	Onko suunnittelussa huomioitu käyttötapauksen yhteiskunnalliseen kontekstiin liittyvä historiallinen ja sosiaalinen eriarvoisuus, jotta järjestelmän toiminta ei uusinna näitä epätasa-arvoja?	
Toimialakohtaiset riskit	Käytetäänkö tekoälyjärjestelmää julkisella tai puolijulkisella alalla sellaisten päätösten tekemiseen tai toiminnan tueksi, jotka vaikuttavat välittömästi tai välillisesti kansalaisten perusoikeuksiin (esim. terveys, turvallisuus, pääsy koulutukseen ja työhön)?	▲
YHDENVERTAISUUSVAIKUTUKSET		
Yhdenvertaisuusarvioinnin resurssointi	Onko organisaatiossa osoitettu henkilö, joka on vastuussa järjestelmän yhdenvertaisuusvaikutusten arvioinnista, tiedostaen yhdenvertaisuus- ja tasa-arvolainsäädännössä asetetut vaatimukset ja mahdollisuudet? Onko hän myös julkisesti tavoitettavissa?	
Erilaisen kohtelun huomiointi	Voiko järjestelmän käyttö johtaa erilaiseen kohteluun kiellettyjen syrjintäperusteiden perusteella (esim. ikä, kansalaisuus, kieli, uskonto, vammaisuus tai muu henkilöön liittyvä syy, YVL 8 §)? Miten tähän on varauduttu jo osana suunnittelua?	▲
Ryhmäkohtaisten vaikutusten arvointi	Onko arvioitu, mihin ryhmään tai ryhmiin tekoälyjärjestelmän käyttö saattaa vaikuttaa kohtuuttomasti joko suoraan tai välillisesti? Ovatko nämä henkilöt erityisen haavoittuvassa asemassa intersektionaalisesti? Liittyykö haitta esimerkiksi palvelun tarkkuuteen, saatavuuteen tai sen konkreettisiin vaikutuksiin?	
SIDOSRYHMIEN OSALLISTUMINEN		
Osallistumisen suunnittelu	Onko laadittu suunnitelma ja nimetty vastuuhenkilö sidosryhmien ja kansalaisten aktiiviseksi osallistumiseksi järjestelmän suunnittelu- ja kehitysprosessiin?	
Sidosryhmien kuuleminen	Aiotaanko asianomaisia yhteisöjä ja ryhmiä kuulla tekoälyn käyttötapauksesta, sen hyödyistä ja tavoitteiden asianmukaisesta määrittelystä? Miten varmistetaan, että heidän huolensa ja kokemuksensa todella huomioidaan järjestelmän suunnittelussa?	

Haavoittuvimpien osallistuminen	Sisältävätkö kuultavat ja suunnitteluun osallistuvat sidosryhmät myös heikoimmassa asemassa olevat ihmiset, esimerkiksi paikallis- ja kansalaisjärjestöjen kautta?	▲
SAAVUTETTAVUUS		
Vähemmistöjen tarpeet	Onko saavutettavuuden ja esteettömyyden periaatteet otettu lukuun eri vähemmistöjen tarpeiden huomioimiseksi järjestelmää suunniteltaessa?	
Kohtuulliset mukautukset	Onko järjestelmän suunnittelussa varmistettu, että vammaiset henkilöt saavat asianmukaisia kohtuullisia mukautuksia?	▲
Yhdenvertaiset mahdollisuudet	Hyödyttääkö järjestelmä ihmisiä tasapuolisesti ja turvaako se kaikkien henkilöiden yhtäläiset mahdollisuudet käyttää ja saada palveluita?	
VASTUULLISUUS		
Hankintaperiaatteet ja niistä viestiminen	Ovatko hankinta- tai kehityskäytännöt, kuten yhdenvertaisuusvaatimukset viestitty ja välitetty selkeästi eteenpäin tekoälyjärjestelmän teknisille kehittäjille? Miten jatkuva kommunikaatio läpi kehitysvaiheiden varmistetaan?	
Mahdollinen väärinkäyttö	Onko järjestelmän mahdolliseen väärinkäyttöön liittyvät riskit dokumentoitu ja onko niihin varauduttu jo osana suunnittelua? Kattavatko nämä sekä käyttäjien tahattoman väärinkäytön että tahallisen väärinkäytön ilkivaltaisten toimijoiden toimesta?	
Roolit ja vastuun jakautuminen	Onko vastuun jakautuminen, roolit ja läpinäkyvyys kehityksen osalta määritelty sekä siihen osallistuvien organisaatioiden välillä että heidän tiimiensä sisällä jo tässä vaiheessa? Osallistuuko kehitykseen työntekijöitä eri osastoilta, kattaen esimerkiksi johdon, toimialueen asiantuntijat, lakipuolen ja tekniikan?	
YHDENVERTAISUUDEN EDISTÄMINEN		
Edistämisvelvoitteen huomiointi	Onko suunnittelussa huomioitu viranomaisten ja muiden toimijoiden laaja edistämisvelvoite, eli onko järjestelmä nimenomaisesti suunniteltu edistämään yhdenvertaisuutta, esimerkiksi huomioimalla paremmin toimintarajoitteisten ihmisten tarpeet?	
Syrjinnän havaitseminen	Hyödynnetäänkö tekoälyä nykyisissä järjestelmissä tai päätöksissä ilmenevän syrjinnän havaitsemiseksi, ennaltaehkäisemiseksi tai siihen puuttumiseksi?	
Positiiviseen erityiskohtelu	Hyödynnetäänkö tekoälyä positiiviseen erityiskohteluun keinona tosiasiallisen yhdenvertaisuuden edistämiseksi, kuten esim. koulujen kielikiintiöiden toimeenpanemiseksi?	
TULOS (0–4, 4,5–9, 9,5–17)		**X / 17**

2.1. KEHITYS - DATA

Aiempaa syrjintää ja eriarvoisuutta heijastava opetusdata on yksi algoritmisen syrjinnän tärkeimmistä syistä. Data ja sen esikäsittelyvaihe kattaa 1) datan keruun, joissa otetaan huomioon datan alkuperä, määrä ja laatu, 2) datan valmistelun, jolla varmistetaan sen edustavuus ja haitallisen sisällön poistaminen, ja lopuksi 3) datan validoinnin ja tasapainottamisen.

Vastuu: Kokonaisuudessaan kehitysvaiheesta päävastuussa on tekoälyjärjestelmän tekninen kehittäjä, vaatien myös jatkuvaa yhteydenpitoa ja yhteistyötä tilaaja-organisaation kanssa.

DATAN KERUU		
Aihe	**Kysymykset**	**Huomioitu**
Datan alkuperä	Käytetäänkö harjoitusdatana tilaaja- tai kehittäjäorganisaation omaa, valmiiksi validioitua dataa? Jos käytetään ulkopuolisia datalähteitä, onko niiden tarkoituksesta, otannasta, puutteista, omistuksesta, käytön ja säilytyksen rajoitteista sekä ajantasaisuudesta varmuutta?	
Datan määrä	Hyödyntääkö järjestelmä suuria datamassoja useista eri lähteistä? Miten niiden alkuperä, laatu ja yhteensopivuus on varmistettu?	
Datan laatu	Miten on huomioitu, että datan laatu ei välttämättä ole yhtenäinen eri ryhmien tai lähteiden välillä? Onko laatua arvioitu esimerkiksi kohdemuuttujien jakautumisen, väestöryhmien ominaisuuksien, aggregoinnin vaikutusten tai puuttuvien tietokenttien suhteen?	
Henkilötietojen käyttö	Käytetäänkö harjoitus- tai syötedatassa erityisiä (arkaluontoisia) henkilötietoja tai muita henkilökohtaisia ominaisuuksia, kuten ikää, kieltä tai uskontoa? Jos käytetään, onko se välttämätöntä järjestelmän toiminnalle ja yhdenvertaisuudelle? Ovatko tietosuojaan liittyvät asiat, kuten yksityisyydensuoja ja tiedonkäsittelyn reiluus huomioitu?	▲
Datasubjektien informointi	Onko datan keruusta ja käytöstä, sisältäen sen perustelut, tarkoituksen ja kategorisoinnit viestitty selkeästi ja saavutettavasti henkilöille, joita se koskee? Onko henkilöiltä saatu tietoon perustuva suostumus datan käyttöön?	
Datan säilytys ja dokumentaatio	Onko data säilytetty ja dokumentoitu turvallisesti, kattaen aiemmat kohdat, jotta harjoitusdata voidaan tarvittaessa luoda uudelleen, todentaa ja päivittää? Onko nimitetty henkilö, jonka vastuulla datan hallinta, päivittäminen ja yhteensopivuus on?	
DATAN VALMISTELU		
Laadun varmistaminen	Onko varmistettu, että käytetyt dataluokat ja -nimikkeet soveltuvat järjestelmän käyttötarkoitukseen?	▲
Vastuullinen datan annotointi	Onko datapisteiden nimeämisestä vastuussa oleva henkilö tai tiimi pätevä valmistelemaan ja annotoimaan järjestelmässä hyödynnettävää, erityisesti arkaluontoista dataa (esim. väestö- tai käyttäymisdataa)?	
Ennakkotapaukset	Onko sovellettavaa dataa tai datajoukkoa käytetty samankaltaisissa sovelluksissa aiemmin? Onko kyseisen tai samankaltaisen datan käytöstä esimerkiksi onnistuneita kokeiluja tai ennakkotapauksia?	
Syy-seuraus-suhteet datan taustalla	Onko käytettävän datan valinta perusteltua käyttökontekstin kausaalisen rakenteen valossa? Onko hyödynnetty informaatio merkityksellistä syy-seuraus-suhteiden näkökulmasta kyseisessä päätöksenteon kontekstissa tai puuttuuko siitä päätöksenteon kannalta merkittäviä tekijöitä?	

Historiallinen eriarvoisuus	Millaisia tavoitteita, käsityksiä, uskomuksia ja mahdollisia historiallisia, yhteiskunnallisia tai rakenteellisia eriarvoisuuden muotoja käytettävä data heijastaa?	▲
DATAN VALIDOINTI		
Edustavuus	Edustaako harjoitusdata kattavasti eri väestöryhmiä joihin järjestelmä vaikuttaa, erityisesti vähemmistöjä? Onko datan edustavuutta verrattu soveltuvaan referenssipopulaatioon (esim. väestötilastoihin)?	▲
Otanta- ja nimikevinoumat	Onko harjoitus- ja syötedatan mahdolliset otantavirheet ja vinoumat nimikkeiden jakaumissa ryhmien välillä huomioitu? Onko datapisteet nimetty luotettavasti, johdonmukaisesti ja syrjimättömällä tavalla?	
Datan tasapainottaminen	Onko harjoitusdataa tarvittaessa tasapainotettu esimerkiksi yliotannalla edustavuuden ja ennustetarkkuuden parantamiseksi? Onko erityisesti harkittu yliotantaa vähemmistöjen kohdalla?	
Korvikemuuttujien huomiointi	Onko datasta identifioitu ja siivottu selkeät sijais- tai korvikemuuttujat henkilökohtaisille ominaisuuksille, jotka voivat johtaa syrjiviin lopputuloksiin?	▲
Loukkaava sisältö	Onko datasta identifioitu ja siivottu mahdollinen syrjivä, loukkaava ja psykologista haittaa aiheuttava sisältö, kuten halventavat kielelliset ilmaisut tai syrjiviä stereotypioita vahvistava kuvamateriaali?	
YHDENVERTAISUUDEN EDISTÄMINEN		
Inklusiivinen datan keruu	Onko järjestelmää varten kerätty dataa inklusiivisesti aliedustetuista ryhmistä heidän kanssaan yhdenvertaisuuden edistämiseksi, yksityisyydensuoja ja GDPR:ssä asetetut datasubjektin oikeudet huomioiden?	
Ei-binääriset datanimikkeet	Onko huomioitu binääristen nimikkeiden ja luokittelujen mahdollisesti aiheuttama haitta vähemmistöille esimerkiksi soveltamalla inklusiivisempia, ei-binäärisiä muuttujia tai nimikkeitä datassa?	
Datan esikäsittely (pre-processing)	Onko kehitysvaiheessa suoritettu muuta datan esikäsittelyä esimerkiksi otos-, mittaus- ja nimikevinoumien korjaamiseksi ja yhdenvertaisuusvaikutusten parantamiseksi?	
TULOS (0–4, 4,5–8,5, 9–16)		**X / 16**

2.2. KEHITYS - ALGORITMISEN MALLIN OPETUS

Syrjivät vinoumat voivat ilmetä myös mallin ja algoritmien opetusvaiheessa, johtuen malliin ja metriikoihin liittyvistä valinnoista. Tärkeää on myös organisaatioiden ja kehitystiimien moninaisuus sekä koulutus yhdenvertaisuudesta.

MALLIN REILUUDEN ARVIOIMINEN		
Aihe	**Kysymykset**	**Huomioitu**
Perustasojen määrittely	Onko mallintamisen yhteydessä tunnistettu perustasot mallinnettavalle ilmiölle tai ominaisudelle (vrt. kohdemuuttujan arvoille) demografisten ryhmien sisällä?	
Reiluusmetriikoiden käyttö	Onko kehitystyössä tutustuttu algoritmien yhdenvertaisuusvaikutusten arviointiin? Onko hyödynnetty esimerkiksi reiluumetriikoita (engl. fairness metrics), kuten ehdollista tilastollista pariteettia tai yhtäläisiä virhetasoja tekoälyjärjestelmän vinoumien ja mahdollisen syrjinnän riskien tunnistamiseksi?	▲
Reiluusmetriikoiden valinta	Onko soveltuvien yhdenvertaisuus- ja reiluusmetriikoiden määrittely ja valinta perusteltua, ottaen huomioon kyseessä olevan yhteiskunnallisen käyttötapauksen sekä soveltuvan lainsäädännön? Onko käytettyjen reiluusmetriikoiden metodologiset oletukset ja rajoitteet tunnistettu?	
Vertailuluokkien valinta	Ovatko mallin reiluuden arvioimisessa vertailuluokkia määrittävät ominaisuudet (esim. alkuperä ja ikä) merkityksellisiä ja perusteltuja välittömän ja välillisen syrjinnän riskien tunnistamisen kannalta? Onko analyysejä suoritettu esimerkiksi kiellettyjen syrjintäperusteiden tai niiden kanssa korreloivien korvikemuuttujien suhteen?	
Moniperusteisen syrjinnän arviointi	Suoritetaanko arviointia intersektionaalisesti (vrt. alaryhmien välillä) moniperusteisen syrjinnän varalta?	
MALLIN / OMINAISUUKSIEN VALINTA		
Mallin toiminta	Onko mallin toimintaa koskien selkeästi dokumentoitu, mitä mallin on tarkoitus ennustaa, luokitella tai suositella (sis. kohdemuuttujien valinta, luokittelutehtävä, suoritustavoitteet) ja mitkä sen tarkoitetut vaikutukset ovat? Millä sen suorituskykyä, reiluutta ja muita asiaankuuluvia tekijöitä mitataan ja arvioidaan?	▲
Mallin rajoitukset	Onko valitun tekoälymallin (esim. lineaarinen regressio, neuroverkko) soveltuvuutta tehtävään arvioitu ja sen mahdolliset rajoitukset tunnistettu ja dokumentoitu? Pyritäänkö käyttämään mahdollisimman yksinkertaista, aiemmin validoitua ja läpinäkyvää mallia tehtävästä suoriutumiseen?	
Mallin alttiudet	Onko tunnistettu ja dokumentoitu mitkä mallin piirteet saattavat altistaa sen tuottamaan syrjiviä tuloksia (esim. läpinäkymättömyys tai syrjivien korvikemuuttujien käyttö)? Miten alttiuksiin on varauduttu tapauksissa, joissa kyseessä on jatkuvasti käyttöympäristöstään datan perusteella oppiva algoritmi?	
Muuttujien valinta	Onko varmistettu, että mallin muuttujina ei käytetä kiellettyjä syrjintäperusteita ilman oikeudellisesti perusteltua ja oikeutettavaa syytä?	▲
Reiluuden kohdearvot	Onko mallin kehittämisen yhteydessä määritelty ja dokumentoitu hyväksyttäviksi arvioidut tavoite- tai kohdearvot mallin reiluudelle (esim. hyväksyttävät erot väärissä positiivisissa ja väärissä negatiivisissa tuloksissa ryhmien välillä)? Onko algoritmin tavoitefunktioon nimenomaisesti sisällytetty jokin reiluustavoite?	

MONINAISUUS KEHITYKSESSÄ		
Kehittäjien monitieteisyys- ja arvoisuus	Onko järjestelmän kehitykseen osallistuva tiimi inklusiivinen ja moninainen väestöllisesti, arvojen, koulutuksen ja hankittujen tietojen ja taitojen suhteen? Sisältääkö tiimi henkilöitä, jotka ovat saaneet koulutusta yhdenvertaisuudesta ja eettisestä teknologiasuunnittelusta?	
Yhteistyö tilaajan kanssa	Tehdäänkö järjestelmän kehityksessä jatkuvaa yhteistyötä tilaaja-organisaation tai muiden ei-teknisten asiantuntijoiden kanssa, erityisesti yhdenvertaisuusperiaatteen paremmaksi huomioimiseksi?	
Yhteiskehittäminen sidosryhmien kanssa	Osallistetaanko kehitysprosessiin loppukäyttäjiä ja järjestelmän vaikutuksenalaan kuuluvia henkilöitä esimerkiksi sidosryhmätyöskentelyn tai yhteiskehittämisen muodossa? Miten erityisesti haavoittuvien henkilöiden tai aliedustettujen ryhmiin kuuluvien intressit otetaan huomioon?	
YHDENVERTAISUUDEN EDISTÄMINEN		
Osaaminen ja koulutus vinoumista ja oikomismenetelmistä	Onko varattuja resursseja, koulutusta tai muuten varmistettu, että kehittäjät pystyvät huomioimaan yhdenvertaisuusnäkökulmat työssään? Onko esimerkiksi varmistettu, että syrjivien vinoumien tunnistamiseen ja tekniseen oikomiseen käytettäviä työkaluja osataan käyttää pätevästi?	
Vinoumia muuttavien reiluusmetriikoiden käyttö	Onko mallin opettamisessa harkittu vinoumia muuttavien reiluusmetriikoiden käyttöä (engl. bias transforming metrics), joilla pyritään edistämään ryhmien tosiasiallista yhtäläistä pääsyä hyödykkeisiin ja palveluihin, sikäli kun niiden käyttö on arvioitu oikeutetuksi?	
Mallin käsittely (in-processing)	Onko hyödynnetty mallin opettamista ja prosessointia koskevia oikomismenetelmiä, kuten regularisointia, kilpailevia oikomismalleja tai optimointirajoituksia yhdenvertaisuuden edistämiseksi mallin opetusvaiheessa?	
TULOS (0–3, 3,5–7, 7,5–13)		**X / 13**

2.3. KEHITYS - MALLIN TESTAAMINEN
Opetettua mallia on testattava ja validoitava sen varmistamiseksi, että se toimii tarkoitetulla tavalla, eivätkä sen tulosteet heijasta syrjiviä vinoumia. Validoinnin ja testauksen perusteella mallia voidaan joutua tarkistamaan ja korjaamaan syrjinnän ehkäisemiseksi ja yhdenvertaisuusvaikutusten parantamiseksi. Huomiota on kiinnitettävä myös lopullisen mallin toiminnan dokumentaatioon.

Aihe	Kysymykset	Huomioitu
MALLIN TESTAUSPROSESSI		
Mallin testi- ja arviointidata	Millaisella testidatalla mallin toimintaa (esim. tulosteiden laatu, osumatarkkuus, virhetasot) arvioidaan? Eroaako testidata tarpeeksi harjoitusdatasta mallin ylisovittamisen välttämiseksi ja mahdollisten vinoumien tunnistamiseksi?	
Mallin reiluuden testaaminen	Onko mallia testattu syrjivien vinoumien varalta? Onko mallin toimintaa esimerkiksi arvioitu ja vertailtu tässä vaiheessa soveltuvien reiluusmetriikoiden avulla?	▲
Välittömän syrjinnän arviointi	Onko mallia ja sen toimintaa arvioitu yksilöiden tasolla välittömän syrjinnän riskien varalta?	
Välillisen syrjinnän arviointi ryhmätasolla	Onko mallin toimintaa ja reiluutta testattu suhteessa eri väestöryhmiin välillisen syrjinnän riskien varalta, esimerkiksi intersektionaalisten vertailuluokkien välillä? Jakautuvatko mallin virheet tasaisesti läpi väestöryhmien ja voivatko ne olla vahingollisempia tietyille ryhmille?	
MALLIN VALIDOINTI		
Oikeutettavat erot	Jos mallin toiminnassa tai tuloksissa on eroja eri yksilöiden tai väestöryhmien välillä (varsinkin syrjintäperusteiden suhteen) voidaanko eroja pitää tarpeellisina, oikeasuhteisina ja perusteltuina järjestelmän toiminnalle? Onko eroja liittyen tulosteiden tai ennustetarkkuuden jakaumiin arvioitu suhteessa käyttökohteen oikeuskäytäntöön?	▲
Vinoumien ehkäiseminen	Onko tunnistettu, vertailtu ja sovellettu erilaisia strategioita syrjivien vinoumien välttämiseksi tai minimoimiseksi?	
Tavoitearvojen yhteensopivuus ja mahdolliset konfliktit	Onko tunnistettu mahdollisia konflikteja (engl. trade-off) esimerkiksi mallin yleisen osumatarkkuuden ja reiluuden tai erilaisten reiluuden kohdearvojen välillä? Onko dokumentoitu läpinäkyvät periaatteet mahdollisten ristiriitojen ratkaisemiseksi järjestelmän reiluuden ja yleisen tarkkuuden välillä?	
Pitkän aikavälin vaikutukset	Onko mallin ja järjestelmän käytön pidemmän aikavälin vaikutuksia arvioitu? Onko suunniteltu tai otettu käyttöön prosesseja pitkäaikaisvaikutusten monitoroimista, kuten ajettu malliskenaarioita tai simulaatioita? Miten mahdolliseen ajautumiseen (ns. model tai data drift) on vastattu?	
MALLIN DOKUMENTOINTI		
Vaihtoehtoiset menettelytavat	Onko vaihtoehtoisten menettelytapojen, kuten erilaisten mallinnustapojen tai täydentävien päätöksentekokäytäntöjen tarpeellisuutta ja käyttöä arvioitu yhdenvertaisuuden näkökulmasta?	▲
Muutokset mallin toiminnassa	Ovatko mallin käyttäytyminen (esim. yleinen ja suhteellinen ennustetarkkuus) tai arvioidut ryhmäkohtaiset vaikutukset muuttuneet testauksen ja mahdollisten korjaustoimenpiteiden myötä?	
Mallin toiminnan dokumentointi	Onko testaus- ja arviointiprosessin perusteella dokumentoitu kattavasti muutokset tekoälymallin toiminnassa (ml. odotetut yhdenvertaisuusvaikutukset), opetuksessa, testauksessa ja käyttötarkoituksessa?	

Tiedonvälitys ja koulutus tilaajalle	Onko järjestelmän tilaaja-organisaatiolle ja käyttäjille toimitettu dokumentaatio, lisenssit ja annettu riittävästi koulusta, jotta he voivat myös itse arvioida järjestelmän toimintaa käytännössä sekä välttää sen väärinkäytön?	
YHDENVERTAISUUDEN EDISTÄMINEN		
Aliedustettujen aseman parantaminen	Onko arvioitu mahdollisuutta ratkaista konfliktit mallin tavoitearvojen (esim. tarkkuuden ja reiluuden välillä) niiden henkilöiden hyväksi, jotka kuuluvat aliedustettuun tai marginalisoituun ryhmään? Onko samaten harkittu eri mallien tai päätöksentekokriteerien käyttämistä eri väestöryhmille yhdenvertaisuuden edsitämiseksi?	
Mallin yhdevertaisuuden dokumentaatio julkisesti	Onko mallin kehitys- ja arviointitoimenpiteet eriytisesti yhdenvertaisuden ja reiluuden osalta dokumentoitu asianmukaisesti esimerkiksi käyttämällä mallikortteja (engl. model cards)? Ovatko nämä tiedot saatavilla myös ulkoista auditointia varten?	
Jälkikäsittely (post-processing)	Onko jälkikäsittelymenetelmiä (esim. päätösten kynnysarvojen muuttamista tai ennusteluokkien edustavuuden tasaamista) hyödynnetty mallien pisteytyksen muuttamiseksi reilummaksi ja yhdenvertaisemmaksi, kuitenkin huomioiden kyseisten keinojen oikeudelliset perusteet?	
TULOS (0–3, 3,5–6,5, 7–12)		**X / 12**

3. KÄYTTÖÖNOTTO
Käyttöönotto on keskeinen osa tekoälyn elinkaarta ja potentiaalinen syrjinnän ilmenemiskohta. Jos tekoälyjärjestelmä esimerkiksi otetaan käyttöön eri ympäristössä ja väestössä kuin mihin se on suunniteltu, sen tarkkuus ja käyttäytyminen voi horjua, johtaen mahdollisesti syrjiviin vaikutuksiin. Huomiota on kiinnitettävä myös järjestelmän yhdenvertaisuusvaikutusten seurantaan, läpinäkyvyyteen, ylläpitoon ja säännölliseen auditointiin.
Vastuu: Käyttöönotosta ovat vastuussa yhdessä tekoälyjärjestelmän tilannut organisaatio sekä sen tekninen kehittäjä.

IHMISOHJAUS		
Aihe	**Kysymykset**	**Huomioitu**
Päätösten automatisoinnin taso	Ovatko järjestelmän tekemät päätökset aina ihmisen valvonnan alaisia, eivätkä täysin automatisoituja? Onko määritelty tarkasti käyttäjän vastuut ja velvollisuudet, kuten missä määrin tekoälyn suosituksen tai ennustuksen tulee ohjata päätöksiä?	▲
Luottamus järjestelmään	Miten todennäköistä on, että loppukäyttäjät ja kansalaiset luottavat tekoälyjärjestelmään ja delegoivat sille päätöksiä, ja kuinka tämä huomioidaan käyttöönotossa? Osaavatko käyttäjät esimerkiksi tulkita järjestelmän tulosteita?	
Käyttöliittymän vaikutukset ihmisiin	Voivatko järjestelmän käyttöliittymän ominaisuudet (esim. miten tulokset esitetään käyttäjille) vaikuttaa odottamattomasti ja negatiivisesti sen käyttöön, esimerkiksi kannustaen huolimattomaan käyttöön tai kärjistäen käyttäjien ennakkoluuloja ihmisistä? Miten tähän on varauduttu?	
SEURANTA		
Käytön skaala	Aloitetaanko järjestelmän käyttöönotto pienemmässä skaalassa tai koeasetelmassa, ennen kuin siirrytään sen täysimääräiseen käyttöön yhteiskunnassa, mahdollisiin syrjiviin tuloksiin varautumiseksi, varsinkin jos kyseessä on automatisoitu päätöksentekojärjestelmä kriittisellä yhteiskunnan alueella?	
Kohdeväestö- ja käyttö	Onko järjestelmä otettu käyttöön samassa kohderyhmässä ja käyttökohteessa, jota varten se myös suunniteltiin ja opetettiin? Onko yhteiskunnallisessa kontekstissa tapahtunut muutoksia, jotka vaikuttavat järjestelmän tarpeeseen, soveltuvuuteen tai yhdenvertaisuusvaikutuksiin?	
Muutokset käyttödatassa	Seurataanko käytössä aktiivisesti, että vastaako tekoälyjärjestelmän käytännön toimintaympäristössä kohtaama data riittävästi opetus- ja testidataa? Ilmoittaako järjestelmä, jos datassa tai mallissa ilmaantuu merkittäviä muutoksia (esim. malli oppii uusia kategorioita, luokkia tai havaitut todennäköisyysjakaumat muuttuvat)?	
Yhdenvertaisuusvaikutusten seuranta	Onko asetettu aikavälit ja mittarit, joilla seurataan ja arvioidaan järjestelmän toimintaa yhdenvertaisuuden ja syrjinnän osalta läpi järjestelmän elinkaaren? Onko järjestelmän käyttäjille annettu koulutusta näiden mittarien seurantaan?	▲
Seurantamittarien määrittely	Määritelläänkö käyttöönoton ja monitoroinnin mittareissa (esim. ennustetarkkuuden ja virhetyyppien osalta) standardit, joiden ylittäminen ja laiminlyönti antaa hälytyksen ja käynnistää järjestelmän tarkistuksen? Tuottaako järjestelmä tällöin virhelokin?	
Syrjivien vaikutusten tarkkailu	Onko järjestelmän käyttöönoton aikana havaittu syrjiviä vaikutuksia yksilöiden tai eri väestöryhmien välillä? Miten niiden epäillään syntyneen? Ovatko paikalliset sidosryhmät ja kansalaisjärjestöt osallistuneet syrjinnän monitorointiin ja raportointiin, esimerkiksi kyselyjen tai palautejärjestelmän kautta?	▲

YLLÄPITO		
Järjestelmän päivitys ja palauteketju	Syötetäänkö syntynyt data ja havainnot takaisin suunnittelun ja kehittämisen aikaisempiin vaiheisiin yhdenvertaisuusvaikutusten parantamiseksi? Onko takaisinkytkentään liittyvät riskit huomioitu operatiivisella ja strategisella tasolla?	
Huolto- ja kunnossapito	Onko tekoälyjärjestelmän huolto- ja hallintatoimenpiteistä sovittu sen teknisen kehittäjän kanssa läpi järjestelmän elinkaaren? Miten järjestelmän tarkkuutta, yhdenvertaisuutta ja käyttöön tarvittavaa tietotaitoa ylläpidetään ajan mittaan, erityisesti yhteiskunnallisen kontekstin ja väestörakenteen muuttuessa?	
Järjestelmän poisto käytöstä	Onko määritelty kuinka pitkään järjestelmää käytetään ja tuetaan sekä kuinka se tarpeen mukaan korvataan tai poistetaan käytöstä? Onko huomioitu, miten järjestelmän käytön lopettaminen vaikuttaa aiemmin vaikutuksenalaan kuuluneiden ihmisten perusoikeuksiin? Mitä tapahtuu esimerkiksi kerätylle datalle, kun sovelluksen käyttöä ei enää jatketa?	
LÄPINÄKYVYYS		
Vastuu järjestelmän virheistä	Ovatko tahot ja organisaatiot, jotka ovat vastuussa järjestelmän käytöstä sekä sen mahdollisiin virheisiin vastaamisesta määritelty selkeästi? Ilmoitetaanko virheistä läpinäkyvästi ulospäin asianmukaisille tahoille?	
Selitettävyys asianomaisille	Ilmoitetaanko henkilöille, joihin tekoälyjärjestelmä vaikuttaa sen käytöstä, päätöksentekoprosessista ja niiden perusteista läpinäkyvästi sekä saavutettavasti? Ovatko järjestelmän toiminta sekä päätökset selitettävissä ja ymmärrettäviä asianomaisille?	▲
Päätöksenteko prosessin kirjaaminen	Kirjataanko järjestelmän toiminta ja tekemät päätökset kattavasti läpi prosessin, sisältäen myös järjestelmään tehdyt muutokset käyttöönoton aikana ja sen jälkeen?	
YHDENVERTAISUUDEN EDISTÄMINEN		
Oikeussuojakeinojen varmistaminen	Onko järjestelmän käyttöönotossa huomioitu ja pyritty varmistamaan vaikutusten alaisten henkilöiden oikeus tehokkaisiin oikeussuojakeinoihin, kuten pääsy tuomioistuimeen ja oikeusapuun palvelun kohdalla?	
Avoin auditointi	Ovatko järjestelmä ja sen tekemät päätökset avoimesti saatavilla auditointia varten (mm. viranomaisille, tutkijoille, kansalaisyhteiskunnalle), sikäli kuin se on voimassa olevan lainsäädännön puitteissa mahdollista?	
Päätöksien korjaaminen	Onko luotu palautekanavat, jotta asianomaiset ja vaikutuksenalaiset henkilöt voivat tarvittaessa hakea muutosta, korvausta tai riitauttaa järjestelmän tekemät päätökset? Miten mahdollisista virheellisistä päätöksistä aiheutuvat vahingot korvataan henkilöille?	
TULOS (0–3,5, 4–8, 8,5–15)		**X / 15**
KOKONAISTULOS		**X / 73**

0–18	**19–38**	**39–73**
Suhteellisen turvallista käyttää	Harkitse käyttöä	Älä käytä ennen muutoksia

Arviointikehikon tuloksen tulkitsemisessa on syytä huomioida seuraavat asiat:

1. Kehikon tarkoituksena ei ole antaa yksiselitteistä arviota sovelluksen käyttämisestä vaan toimia työkaluna, joka mahdollistaa syrjintäriskien tunnistamisen ja niihin vastaamisen. Kannustamme sen tilannekohtaiseen soveltamiseen.
2. Tulosten tulkinta sekä siitä johtopäätösten tekeminen vaatii keskustelua ja yhteistyötä sovelluksen sidosryhmien kanssa. Tulkinnassa on tarpeen hyödyntää erilaista osaamista liittyen esimerkiksi data-analytiikkaan, syrjintään ja kulloiseen käyttökontekstiin.
3. Arviointikehikko ei itsessään pysty huomioimaan kaikkea algoritmiseen syrjintään liittyviä ja yleisiä oikeudellis-eettisiä näkökulmia ja mahdollisia ongelmia. Sen käyttämisen lisäksi tarvitaan laaja-alaista asiantuntijuutta sekä yhteiskunnallista keskustelua tekoälyn hyödyntämisen yhteiskunnallisista vaikutuksista.

Alla määritellään arviointikehikon kannalta keskeisiä käsitteitä ja termejä sen käytön helpottamiseksi.

Kehikon keskeiset käsitteet ja termit:

Tekoälyjärjestelmä: Tekoälyjärjestelmä tai -sovellus viittaa aiempaa kehittyneiseen ja autonomisiin tietokonepohjaisiin päätöksenteko- tai suosittelujärjestelmiin. Arviointikehikon tyyppiesimerkkinä ovat koneoppivat järjestelmät osana julkisen sektorin automatisoitua päätöksentekoa.

Tekoälyn elinkaari: Tekoälysovelluksen elin- ja kehityskaari, joka kattaa suunnittelun, kehityksen ja käyttöönoton itetariitivisena prosessina. Elinkaari-ajattelu korostaa sitä, että syrjivät vinoumat voivat syntyä ja ilmentyä läpi tekoälyn kehityskaaren.

Vaikutustenarviointi ja riskienhallinta: Vaikutustenarvioinnin tarkoituksena on tunnistaa, arvioida ja hallita esimerkiksi algoritmeihin tai henkilötietojen käsittelyyn sisältyviä riskejä. Arviointikehikko on tarkoitettu jatkuvaksi ja ennakoivaksi riskien hallinnan prosessiksi, korostaen organisaatioiden rakenteita ja kyvykkyyksiä.

Yhteiskunnallinen konteksti: Tekoälyjärjestelmän yhteiskunnallisella käyttökontekstilla ja siihen liittyvällä eriarvoisuudella, käyttötarkoituksella, käyttäjillä ja käytön skaalalla on suuri vaikutus siihen, miten ja mitkä syrjivät vinoumat voivat ilmetä. Teknisten ratkaisujen ohella on käyttökontekstin huomioiminen edellyttää osallistavaa suunnittelua ja vaikutusten alaisten ryhmien kuulemista tekoälyn käytöstä.

Yhdenvertaisuus: Yhdenvertaisuus tarkoittaa sitä, että kaikki ihmiset ovat samanarvoisia riippumatta esimerkiksi iästä, alkuperästä, kansalaisuudesta, kielestä, uskonnosta tai muusta henkilöön liittyvästä syystä (kielletyt syrjintäperusteet YVL 8 §). Suomen yhdenvertaisuuslain tarkoituksena on edistää yhdenvertaisuutta, ehkäistä syrjintää sekä tehostaa syrjinnän kohteeksi joutuneen oikeusturvaa.

Syrjintä: Syrjintä tarkoittaa sitä, että ihmistä kohdellaan huonommin kuin toista henkilöä jonkin henkilökohtaisen ominaisuuden perusteella. Syrjintä siis rikkoo ihmisten oikeutta yhdenvertaiseen kohteluun. Erilainen kohtelu ei kuitenkaan ole syrjintää, jos se perustuu lakiin, sillä on muuten hyväksyttävä tavoite ja se on oikeasuhtaisia. Tekoälyn käytöstä johtuva algoritminen syrjintä voi olla vaikeasti huomattavissa ja todistettavissa algoritmien läpinäkymättömyyden vuoksi.

Välitön ja välillinen syrjintä: Syrjintä voidaan jakaa välittömään eli suoraan syrjintään, jolloin henkilöä kohdellaan huonommin pelkästään jonkin hänen henkilökohtaisen ominaisuutensa takia, kuin jotakuta toista samanlaisessa tilanteessa. Välillinen syrjintä tarkoittaa puolestaan sitä, että näennäisesti neutraali sääntö, peruste tai käytäntö saattaa jonkun henkilön muita huonompaan asemaan epäsuorasti henkilökohtaisen ominaisuuden perusteella. Välillistä syrjintää voi pitää suurempana uhkana tekoälyjärjestelmissä.

Moniperusteinen syrjintä: Moniperusteisella syrjinnällä tarkoitetaan syrjinnän kohteeksi joutumista useiden eri syrjintäperusteen perusteella. Syrjintä voi tällöin olla moninkertaista, kumulatiivista ja risteävää. Esimerkiksi johonkin vähemmistöön kuuluva henkilö voi tulla syrjityksi sekä ikänsä että alkuperänsä vuoksi. Käsite on lähellä intersektionaalisuutta, joka korostaa eri sosiaaliryhmien ja identiteettien risteävyyttä, ja näiden yhteisvaikutuksena vahvistuvaa marginalisaatiota.

Positiivinen erityiskohtelu: Positiivinen erityiskohtelu tarkoittaa toimenpiteitä tosiasiallisen yhdenvertaisuuden turvaamiseksi, jotka ovat tarpeellisia parantamaan tietyn syrjinnälle alttiin ryhmän henkilöiden asemaa ja olosuhteita. Oikeasuhtainen erilainen kohtelu yhdenvertaisuuden edistämiseksi tai syrjinnästä johtuvien haittojen ehkäisemiseksi ei ole yhdenvertaisuuslain mukaan syrjintää. Positiivista erityiskohtelua voivat olla esimerkiksi vähemmistökiintiöt oppilaitoksiin.

Kohtuulliset mukautukset: Yhdenvertaisuuslain mukaan viranomaisen, koulutuksen järjestäjän, työnantajan sekä tavaroiden ja palvelujen tarjoajan on tehtävä kohtuullisia mukautuksia vammaiselle henkilölle. Mukautuksilla turvataan vammaisen ihmisen yhdenvertaisuutta ja niiden epääminen on syrjintää. Kohtuullisia mukautuksia voivat olla esimerkiksi vammaisille henkilöille tarpeellisten apuvälineiden ja työskentelytapojen mahdollistaminen työpaikalla tai koulussa.

Saavutettavuus: Saavutettavuus on esteettömyyttä digitaalisessa ympäristössä, jonka tarkoituksena on varmistaa, että kuka tahansa voi ymmärtää ja käyttää palveluja, kuten verkkosivuja ilman syrjintää. Se on siis yhdenvertaisuuden edistämistä, kattaen palvelujen teknisen toteutuksen, helppokäyttöisyyden ja niiden ymmärrettävyyden.

Vähemmistöt: Vähemmistöillä tarkoitetaan lähtökohtaisesti kansallisia, etnisiä, kulttuurisia, kielellisiä tai uskonnollisia vähemmistöryhmiä. Vähemmistöt ovat usein sosiaalisesti heikossa asemassa ja heihin kohdistuu negatiivisia stereotypioita ja ennakkoluuloja, jotka altistavat heitä syrjinnälle. Suomessa tällaisia väestöryhmiä ovat esimerkiksi: maahanmuuttajat, vammaiset, seksuaalivähemmistöt, lapset, vanhukset, uskonnolliset ryhmät, romanit ja saamelaiset.

Harjoitus- ja testidata: Harjoitus- tai opetusdata tarkoittaa dataa, jolla koneoppimismalli opetetaan tekemään ennusteita tai päätöksiä esimerkkien avulla. Se voi olla esimerkiksi tekstiä, numeroita, kuvaa tai ääntä. Jos data on puutteellista, epäedustavaa tai muuten virheellistä, sen käyttö voi johtaa mallin vinoumiin, jotka esiintyvät muun muassa tulosteiden epätasaisina jakaumina väestöryhmien välillä. Mallin kouluttamisen jälkeen sen tekemiä ennusteita testataan erillisellä testidatalla.

Tekoälymalli: Tekoäly- tai koneoppimismalli tarkoittaa datasta oppivia algoritmeja ja niiden toimintaa. Malli tunnistaa datasta toistuvia kuvioita ja tekee niiden perusteella ennusteita. Paljon käytettyjä mallityyppejä ovat mm. lineaarinen ja logistinen regressio, päätöspuut, klusterointi ja syväoppivat neuroverkot.

Vinoumat: Vinoumat ovat koneoppimismalliin rakentuneita tai opetusprosessissa tuotettuja korrelaatioita ja rakenteita, jotka näyttäytyvät esimerkiksi järjestelmän systemaattisena taipumuksena tuottaa epäsuotuisia tuloksia tai vääriä ennusteita johonkin väestöryhmään kuuluvien yksilöiden kohdalla. Jos tämä erilainen kohtelu perustuu kiellettyyn syrjintäänperusteeseen ilman oikeudellista syytä, vinoumaa voidaan pitää syrjivänä rakenteena järjestelmässä. Vinoumat voivat johtua mm. datasta, mallin koulutuksesta, tai sen käyttöönotosta. Esimerkiksi datan edustavuuden kohdalla otanta- ja nimikevinoumat ovat tärkeitä.

Reiluusmetriikka: Reiluusmetriikat ovat tilastollisia työkaluja tekoälymallin reiluuden ja mahdollisten vinoumien diagnosointiin ja arviointiin. Niillä voidaan analysoida ja vertailla millaisiin lopputuloksiin ja ennustuksiin malli päätyy eri ryhmiin kuuluvien henkilöiden kohdalla. Niitä voidaan käyttää myös tavoitearvoina mallille esimerkiksi rakentamalla ne optimointirajoitteeksi mallin opetusprosessille. Suosittuja reiluusmetriikoita ovat mm. tilastollinen pariteetti, yhtäläiset virhetasot (väärät positiiviset ja väärät negatiiviset) ja tasoitetut kertoimet ryhmien välillä.

Oikomismenetelmä: Oikomismenetelmät viittaavat teknisiin tapoihin poistaa tai lieventää tekoälyjärjestelmässä tunnistettuja vinoumia joko esi-, malli- tai jälkikäsittelyvaiheissa. Tämä voi tarkoittaa esimerkiksi datan siivoamista, datapisteiden uudelleen nimeämistä tai painottamista, optimointirajoitusten käyttöönottoa tai regularisoinnin suorittamista, päätöksenteossa käytettävien kynnysarvojen muuttamista tai esimerkiksi eri ryhmien edustuksen tasaamista ennusteluokissa.

Läpinäkyvyys: Tekoälyn läpinäkyvyys (eli esimerkiksi sen toiminnan ja tulosteiden ymmärrettävyys ja jäljitettävyys) on merkityksellistä yhdenvertaisuusperiaatteen näkökulmasta. Läpinäkyvyyden puute voi osoittautua ongelmalliseksi sekä a) tekoälyn käyttämän datan ja mallin tasolla varsinkin ohjaamattoman koneoppimisen kohdalla, b) kehitys- ja käyttöprosessin tasolla organisaatioissa, sekä c) kansalaisen tasolla sikäli, kun tekoälyjärjestelmän käytöstä ei informoida, kun sen prosessoinnin logiikka ei ole ymmärrettävissä tai kun sen tuottamia tulosteita ei pystytä selittämään. Kaikkien tasojen tulisi ideaalisti olla mahdollisimman avoimia kuitenkin tieto- ja yksityisyydensuojaan, luottamuksellisuuteen ja järjestelmän turvallisuuteen liittyvät seikat huomioiden.

Tekoälyn tilaaja ja kehittäjä: Kehikossa tehdään jako tekoälyjärjestelmän tilaavaan ja käyttöönottavaan, yleensä julkisen sektorin organisaation, sekä toisaalta tekoälyn tekniseen, usein yksityiseen kehittäjään. Jako havainnollistaa tarvetta näiden tahojen väliseen yhteistyöhön läpi järjestelmän elinkaaren. Yhteistyö on tärkeää myös organisaatioiden sisällä eri tiimien kesken, varsinkin silloin, jos tekoälysovellus kehitetään yhden organisaation toimesta omaan käyttöön.

3.4 Politiikkasuositukset

Tämä luku esittelee hankkeen politiikkasuositukset, jotka on muodostettu kaikkien kolmen osahankkeen tuloksien ja kerätyn aineiston pohjalta. Politiikkasuositusten päätavoitteena on varmistaa arviointikehikon hyödynnettävyys ja käyttö osana julkisen hallinnon toimintaa. Suositusten päätavoite on jaettu kolmeksi osatavoitteeksi, jotka sisältävät sekä lyhyen että pitkän aikavälin tavoitteita ja toimenpiteitä kehikon käyttöönottamiseksi ja sen jatkuvuuden varmistamiseksi. Osatavoitteet ovat:

1. **Algoritmista syrjintää koskevan yleisen tietoisuuden lisääminen**
2. **Eri toimijaryhmien yhteistyön lisääminen tekoälyjärjestelmien vastuullisessa kehittämisessä**
3. **Yhdenvertaisuuden edistämisen mahdollistava sääntely ja työkalut tekoälyn käytössä**

Suositukset eivät sisällä kehikon yksityiskohtia koskevia tarkkoja ohjeistuksia, vaan keskittyvät sääntelyä koskeviin kysymyksiin sekä kehikon hyödynnettävyyteen. Suosituksissa painottuu julkisen hallinnon näkökulma, mutta ne viittaavat myös muihin toimijaryhmiin, kuten kehittäjiin seuraten näin itse kehikon rakennetta.

1. Algoritmista syrjintää koskevan yleisen tietoisuuden lisääminen

Tieto ja tietoisuus algoritmisen syrjinnän erityispiirteistä ja yhteiskunnallisesta merkityksestä on vahvistumassa, mutta tätä trendiä on aktiivisesti tuettava tiedonvälityksen ja vuorovaikutuksen, koulutuksen ja kyvykkyyksien lisäämisen sekä vastuiden tunnistamisen avulla. Yhdenvertaisuuslain mukaisesti viranomaiselle, koulutuksen järjestäjälle ja työnantajalle on määritelty vastuu yhdenvertaisuuden edistämisestä. Poliittisen ohjauksen kautta tulisi varmistaa, että näillä keskeisillä toimijoilla on riittävän hyvä ymmärrys algoritmisen syrjinnän juurisyistä ja riskeistä. Lisäksi näiden ryhmien käsitystä siitä, miten tekoälyä voidaan käyttää yhdenvertaisuuden edistämiseksi, tulisi vahvistaa.

1.1. Yhdenvertaisuuden edistämisen ja tekoälyn hyödyntämisen välistä yhteyttä on vahvistettava. Toistaiseksi keskustelu tekoälyjärjestelmien vinoumista on keskittynyt syrjinnän kieltoon, mutta yhdenvertaisuuden edistämisen näkökulman vahvistaminen on erityisen tärkeää Suomen yhdenvertaisuuslain sekä eurooppalaisen syrjintälainsäädännön näkökulmasta. Tämä on mahdollista esimerkiksi kiinnittämällä huomiota siihen, miten tekoälyä voidaan hyödyntää nykyisten menettelytapojen syrjivyyden havaitsemiseen ja marginalisoitujen ryhmien aseman parantamiseen.

1.2. Yhdenvertaisuuden edistäminen tekoälyn käyttämisessä edellyttää laajaa sosioteknistä näkökulmaa, joka korostaa yhdenvertaisuusvaikutusten arvioinnin paikallisuutta ja tapauskohtaisuutta. Tekoälysovellusten mahdollisten syrjivien vaikutusten realisoitumiseen vaikuttavat useat kontekstisidonnaiset, sosiaaliset, kulttuuriset ja teknologiset tekijät, kuten järjestelmän käyttötarkoitus, kohdeväestön koostumus, käyttäjien toiminta sekä muutokset käyttökontekstissa ajan myötä. Sosioteknisen näkökulman hyödyntäminen merkitsee esimerkiksi monitieteellistä lähestymistä tekoälyjärjestelmän suunnittelun alusta lähtien, jolloin erot yhteiskunnallisissa käyttötapauksissa ja vaikutuksissa eri ihmisryhmiin voidaan paremmin huomioida.

1.3. Tekoälyjärjestelmien vielä suhteellisen alhainen käyttöönotto Suomessa mahdollistaa niiden kehittämisen ohjaamisen syrjimättömyyden ja yhdenvertaisuuden edistämisen huomioivaan suuntaan. Erityisesti vielä alkuvaiheessa oleva tekoälyjärjestelmien hyödyntäminen eri aloilla tulisi nähdä tilaisuutena huomioida syrjimättömyys ja yhdenvertaisuus jo suunnittelu- ja kehityskaaren alkuvaiheessa, jolloin toimenpiteet niiden edistämiseksi ovat tehokkaimpia.

1.4. Suomessa on laadittava kansalliseen kontekstiin soveltuvat tarkemmat ohjeistukset, lisättävä tietoa toimijoiden vastuista, sekä luotava standardeja organisaatiotason kehitykseen algoritmisen syrjinnän osalta. Lainmukaisuusvaatimuksien, vastuiden ja velvollisuuksien täytyy olla selviä organisaatiotasolla. Osaaminen ja resurssit tekoälyjärjestelmien syrjivien vaikutusten huomioimiseen ovat vielä vajavaisia. Tämä korostuu varsinkin ulkomailta hankittujen tekoälyjärjestelmien kohdalla, jotka voivat olla ovat verrattain läpinäkymättömiä, ns. mustia laatikoita. Tästä syystä läpinäkyvyyttä lisäävät ohjaustyökalut, kuten arviointikehikko ovat arvokkaita.

1.5. Teknisiä menetelmiä syrjivien vinoumien ehkäisemiseen tulisi käsitellä osana laajempia tekoälyjärjestelmien auditointi- ja vaikutuksenarviointiprosesseja. Syrjintää on arvioitava tapaus- ja asiakohtaisesti, joten yhdenvertaisuuteen liittyviä haasteita ei voida ratkaista ainoastaan teknisin menetelmin, kuten datan laadunarvioinnilla. Mikäli teknisten oikomismenetelmien käytöstä halutaan tehokas (itse)sääntelyn keino, ne edellyttävät tuekseen selkeitä standardeja ja läpinäkyvää auditointia, sekä syrjinnän yhteiskunnallisen kontekstin huomioimista.

1.6. Neutraalien tai vinoumista vapaiden algoritmien sijaan tulisi tavoitella niiden oikeutettua käyttöä. Tämä korostaa osallistavia ja yhdenvertaisuusperiaatteen toteutumisen varmistavia prosesseja tekoälyn suunnittelussa, kehittämisessä ja käyttöönotossa**.** Täysin ”vinoutumatonta” järjestelmää ei ole olemassa ja lisäksi järjestelmiin voidaan myös rakentaa tarkoituksellisesti positiivisia yhdenvertaisuusvaikutuksia edistäviä vinoumia. Yhdenvertaisuusvaikutusten tehokasta ja vaikuttavaa arvioimista voisikin osaltaan edesauttaa ”teknologian neutraalisuuden” diskurssista luopuminen.

1.7. Osana tietoisuuden lisäämistä tulee korostaa organisaatioiden vastuullisten toimintatapojen ja kulttuurien merkitystä tekoälyn syrjimättömässä soveltamisessa. Tietoisuuden lisääminen laajasti eri sektoreilla on osa arviointikehikon menestyksellistä käyttöönottoa. Vastuullinen toimintatapa koskee tässä yhteydessä esimerkiksi dokumentointi- ja arkistointikäytäntöjä, koulutusta, vastuuhenkilöiden rooleja ja resursseja sekä seurantakäytäntöjen juurruttamista osana tekoälyjärjestelmien kehitystä- ja käyttöönottoa.

1.8. Haavoittuvimpien sidosryhmien ja kansalaisten osallistumista jo tekoälysovelluksen suunnitteluvaiheeseen sekä heidän oikeuksiensa toteutumista on tuettava. Erityisesti tämä pätee sektoreilla ja toimialoilla, kuten vakuutusalalla ja turvallisuussektorilla, joiden läpinäkyvyyden ja toiminnan avoimuuden puute vaikeuttaa algoritmisen syrjinnän todentamista ja ehkäisemistä sekä mahdollisten syrjittyjen henkilöiden pääsyä oikeuksiinsa. Syrjinnän ohella algoritmisen päätöksenteon kohdalla on arvioitava mitkä

ovat päätöksenteon kohteiden tosiasialliset mahdollisuudet riitauttaa syrjiviksi havaittuja päätöksiä. Varsinkin näillä aloilla on myös kriittisesti arvioitava tekoälyavusteiselle päätöksenteolle vaihtoehtoisia tai täydentäviä menettelytapoja.

2. Eri toimijaryhmien yhteistyön lisääminen tekoälyjärjestelmien vastuullisessa kehittämisessä

Arviointikehikko ohjaa sen käyttäjiä hahmottamaan ne vastuutahot, jotka elinkaariajattelun mukaisesti osallistuvat tekoälyjärjelmän kehittämiseen ja sen syrjimättömyyden varmistamiseen. Roolien ja vastuiden selkiyttäminen varmistaa, että mikään toteutusvaihe ei jää huomiotta. Sekä järjestelmän tilaajalla että sen kehittäjällä on vastuu yhteistyön sujuvuudesta ja tiedonkulusta eri toimijoiden välillä. Itse kehikon käyttöön liittyvän välittömän yhteistyön lisäksi tarvitaan kuitenkin sen rajat ylittävää koordinointia, jotta arviointikehikon käyttö voidaan yleisesti juurruttaa osaksi tekoälyjärjestelmien kehittämiseen ja hyödyntämiseen liittyviä käytäntöjä.

2.1. Molemminpuolista ymmärrystä tekoälyn kehittäjien ja virkahenkilöiden välillä yhdenvertaisuuden haasteista tulisi edistää. Syrjivät vinoumat voivat ilmetä läpi tekoälyn elinkaaren, liittyen järjestelmän suunnitteluun, dataan, algoritmiin, läpinäkyvyyteen ja käyttöönottoon. Tästä syystä järjestelmien kehityksessä ja käyttöönotossa tarvitaan tekoälyjärjestelmän julkisten tilaajien ja yksityisten kehittäjien saumatonta yhteistyötä.

2.2. Yhdenvertaisuusvaltuutetun roolia sekä ylimpien laillisuusvalvojien välistä yhteistyötä tekoälyn valvonnassa on syytä kasvattaa. Yhdenvertaisuusvaltuutetun toimenkuvaa ja resursseja on vahvistettava tekoälyyn liittyvien syrjintä- ja yhdenvertaisuusasioiden valvonnassa tietosuojavaltuutetun roolin esimerkin suuntaisesti. Yhdenvertaisuusvaltuutetun lisäksi oikeuskansleri, tietosuoja- ja tasa-arvovaltuutetut sekä uusi tiedonhallintalautakunta voivat olla asiassa avainrooleissa.

2.3. Arviointikehikon pilotointia ja soveltamista eri käyttökonteksteihin, kuten koulutukseen, finanssialalle tai kuntatasolle tulee tukea ja edistää. Jotta kehikon käytettävyydestä ja mahdollisista rajoituksista saadaan riittävästi palautetta ja tietoa varhaisessa vaiheessa jatkokehitystä varten, tulisi sen ensimmäiset käyttökerrat suunnitella piloteiksi, joissa sen rajoja ja mahdollisuuksia voidaan testata käytännössä. Myös muutoin kehikon soveltaminen eri konteksteihin on kannustettavaa, esimerkiksi luomalla tarkempiin käyttötapauksiin soveltuvia ohjeistuksia, lisäämällä kysymyksiä tai painottamalla tekoälyn elinkaaren eri osa-alueita. Kehikon soveltamisessa eri käyttöyhteyksiin tulisi hyödyntää poikkitieteellistä yhteistyötä.

2.4. Kehikkoa tulisi soveltaa ja integroida yhteiseksi viitekehykseksi julkisen puolen tekoälyhankkeiden kilpailutuksiin varhaisessa vaiheessa. Näin voidaan myös ohjata yksityisten palveluntuottajien yhdenvertaisuuden huomiointia ja arviointia skaalautuvasti läpi julkisen sektorin tekoälyhankkeiden. Se myös edistäisi julkisen ja yksityisen puolen yhteistyötä ja jaettua ymmärrystä aiheesta. Kehikkoa on mahdollista myös hyödyntää esimerkiksi tekoälyyn liittyvän tutkimus- ja innovaatiorahoituksen kohdentamisessa ja arvioinnissa.

2.5. Osana arviointikehikon käyttöönottoa ja sen vakiinnuttamista tulee tarkastella jo olemassa olevien vaikutustenarviointien ja hankintakäytäntöjen prosesseja ja työkaluja. Erityisesti yleiseen tietosuoja-asetukseen (GDPR) kytkeytyvän tietosuojan vaikutustenarvioinnin vahvistamista arviointikehikon algoritmisten syrjintäriskien arvioinnilla tulisi edistää. Tämä on hyödyllistä myös siksi, että tietosuoja-asetuken korostama läpinäkyvyys on keskeistä myös algoritmisen syrjinnän välttämisessä.

2.6. Eri hallinnon tasoilla on selvennettävä miten vastuut sisäisestä koulutuksesta, resursoinnista ja tiedon hankinnasta tekoälyn yhdenvertaisesta käytöstä järjestetään. Tämä koskee ymmärryksen muodostamista esimerkiksi käyttökokemuksista, sovellusaloista ja sääntelytarpeista**.** Tekoälyjärjestelmien käyttöönotto on poikkihallinnollinen kysymys, joka koskettaa eri hallinnon sektoreita eri tavoin. Keskeistä on ymmärrys siitä, millaista keskitettyä koordinointia työkalujen, kuten arviointikehikon käyttöönotto ja vakiinnuttaminen vaativat ja missä määrin vastuuta voidaan keskittää tai hajauttaa.

2.7. Hallinnon eri sektoreiden sisällä kertyvä alakohtainen tieto algoritmisen syrjinnän riskeistä ja keinoista edistää yhdenvertaisuutta tulisi saada koko julkisen sektorin käyttöön verkostomaisesti. Lisäksi sektorikohtaiset opit ja esimerkkien kautta kertyvät kokemukset tulisivat olla jaetun järjestelmän kautta eri hallinnonalojen saatavilla. Tätä voidaan lisätä poikkihallinnollista yhteistyötä, vuorovaikutusta ja koulutusta vahvistamalla sekä luomalla oppimisen mahdollistavia yhteisöjä ja verkostoja (vrt. hallinnon lohkoketjuverkosto[267]). Lisäksi virkahenkilövaihdot voisivat lisätä vuorovaikutusta ja ymmärrystä eri hallinnonaloista ja -rooleista vaikutustenarvioinnissa.

2.8. Yhteistyötä ja monimuotoisuutta on tekoälyn syrjimättömän käytön varmistamiseksi lisättävä myös organisaatioiden sisällä. Tekoälyjärjestelmiä valmistelevien tiimien sosiokulttuurinen ja tiedollinen monimuotoisuus on tärkeä epäsuora tekijä syrjivien vinoumien synnyssä. Sekä tekoälyn kehittäjien että tilaajien tulisi siis huomioida se

267 ”Hallinnon lohkoketjuverkosto - Valtiovarainministeriö.” https://vm.fi/documents/10623/12501645/Hallinnon+lohkoteknologiaverkostot/f0631cae-0358-ea08-66a7-48e508e8b5d4/Hallinnon+lohkoteknologiaverkostot.pdf. Viitattu 6.6.2022.

oman organisaationsa sisällä. Tämä kattaa esimerkiksi projektijohtajat, datatietielijät, tietoturvan, lakipuolen ja liiketoiminnan asiantuntijat. Erityisesti johdon sitoutuminen organisaatioissa on keskeisessä roolissa. Tämä vaatii organisaatioilta vastuuhenkilöiden nimeämistä ja yhdenvertaisuusvaikutusten arvioinnin resursointia. Myös yhteistyö yliopistojen, tutkimusinstituutioiden ja kansalaisyhteiskunnan kanssa on kannustettavaa.

3. Yhdenvertaisuuden edistämisen mahdollistava sääntely ja työkalut tekoälyn käytössä

Samalla kuin tietoisuus algoritmisesta syrjinnästä lisääntyy ja tekoälyjärjestelmien hyödyntäminen yleistyy, digitalisaatiota ja tekoälyä koskeva sääntely on jatkuvasti päivittymässä. Yhdenvertaisuuden edistämisen mahdollistavan sääntelyn yhdistäminen tekoälyn hyödyntämiseen ja tähän liittyvien työkalujen kehittäminen tulee olla kiinteä osa kansallista ja EU-tason politiikkaa tulevina vuosina.

3.1. Ministeriöiden tulisi edistää tekoälyn syrjimätöntä kehittämistä ja käyttöönottoa tukevien työkalujen, kuten arviointikehikon viemistä käytäntöön omalla hallinnonalallaan. Arviointikehikko tulisi ottaa käyttöön vähintäänkin strategiatasolla. Tämä integroisi kehikon osaksi hallinnon ohjausjärjestelmiä, kun sovelluksia suunnitellaan, ja sen tarkempi soveltaminen jäisi virastoille ja muille laitoksille. Erityisesti valtiovarainministeriön rooli avoimen julkisen hallinnon kehittäjänä on asiassa potentiaalisesti tärkeä. Arviointikehikon soveltamista voitaisiin edistää esimerkiksi virastojen tulosohjauksen kautta.

3.2. Tekoälyn yhdenvertaisuusvaikutusten arviointi tulisi vakiinnuttaa ja tehdä pakolliseksi hallinnon tekoälyhankkeissa seuraten Alankomaiden ja Kanadan esimerkkiä. Tekoälysovellusten yhdenvertaisuusarviointi voitaisiin kytkeä esimerkiksi viranomaisten velvollisuuteen tehdä toiminnallista tasa-arvo- ja yhdenvertaisuussuunnittelua, ja kirjata se viraston tai laitoksen strategiaan ja henkilöstön koulutukseen. Tämä mahdollistaisi myös yhtenäisten standardien luomisen asiaan. Arviointikehikon käytöllä ja soveltamisella läpi hallinnonalojen voidaan osaltaan vastata tähän.

3.3. Olemassa olevan sääntelyn soveltuvuus algoritmisen syrjinnän ehkäisyyn ja sovitteluun on arvioitava ja selkiytettävä eri hallinnonalojen tarpeet huomioiden. Tekoälyn laajan soveltamisalan myötä algoritmiseen syrjintään ei välttämättä voida puuttua yhtä hyvin eri käyttökonteksteissa. Tämän selkeyttäminen helpottaisi myös työkalujen, kuten arviointikehikon, istuttamista osaksi olemassaolevaa lainsäädäntöä niiden käyttökelpoisuuden lisäämiseksi.

3.4. Julkisen hallinnon on huomioitava välillinen ja moniperusteinen syrjintä paremmin algoritmien käytön kasvaessa joko lainsäädännölliseti tai muilla tavoin. Esimerkiksi oikeussuojakeinoja voitaisiin kehittää huomioimaan paremmin moniperustainen

algoritminen syrjintä tukipalveluissa ja oikeusprosessissa. Koneoppimisen ja syrjivien korvikemuuttujien myötä välillisen ja moniperusteisen syrjinnän riskit korostuvat algoritmisessa päätöksenteossa läpi sektoreiden ja toimialojen. Välillisen syrjinnän arviointi on kuitenkin hyvin tilannekohtaista, joten tekoälyn käytössä tapahtuva syrjintä on vaikeammin havaittavissa ja näyttävissä toteen.

3.5. Sääntelyn kehittämisessä on keskityttävä laajemmin tekoälyn uusintamaan rakenteelliseen eriavoisuuteen, ei vain oikeudelliseen syrjintään. On huomioitava myös tekoälyavusteinen rakenteellinen syrjintä ja laajempi yhteiskunnallinen ja historiallinen eriarvoisuus, jota sovellukset voivat uusintaa tai jopa vahvistaa. Samalla on huomioitava tekoälyjärjestelmien muunnettavuus eri kieli-ja vähemmistöryhmille ja näin kohtuulliset mukautukset.

3.6. Suomen tulee olla EU:n eettiseen ja luotettavaan digitalisaatioon ja tekoälyyn liittyvän sääntelyn kehittymisessä aktiivisesti ja oikea-aikaisesti mukana. Sääntelyn kehitys yhdenvertaisuuden edistämisen näkökulmasta on keskeistä. Suomi on suuntaa näyttävässä erityisasemassa yhdenvertaisuuslainsäädäntönsä erityisluonteen ansiosta sikäli, että kiellettyjen syrjintäperusteiden listan avoimuus mahdollistaa suhteellisen laajat toimet tosiasiallisen yhdenvertaisuuden edistämiseksi. Tähän liittyen hallinnon kyvykkyyksiä vaikuttaa EU-lainsäädäntöön ja kantoihin vaikuttamiseksi oikea-aikaisesti on edistettävä.

3.7. Sääntelyn ja julkisten työkalujen, kuten arviointikehikon ajantasaisuus on arvioitava ja päivitettävä ajoittain. EU:n tekoälyyn liittyvä lainsäädäntö kehittyy jatkuvasti samaan aikaan, kun tieto algoritmisesta syrjinnästä ja sen vaikutuksista lisääntyy. Algoritmien auditointia tukevien kompetenssien ja alustojen kehittäminen on oletettavasti yksi tulevan sääntelyn tärkeistä kulmakivistä. Työkaluja ja sääntelyä tulisi siis kehittää suhteessa EU:n sääntelyyn, kansalliseen lainsäädäntöön ja muuhun kehittyvään sääntelyyn tai aloitteisiin kansainvälisesti (Esim. Euroopan neuvosto, UNESCO).

3.8. Arviointityökalujen ja sääntelyn ei tulisi lisätä tekoälyn kehittäjien taakkaa, vaan tukea jo olemassa olevaa työtä ja lain asettamien velvoitteiden täyttämistä. Kehittäjiä tulisi motivoida hyödyntämään tekoälyn yhdenvertaisuutta edistävää potentiaalia, esimerkiksi kansallisten tekoälyohjelmien, kuten Tekoäly 4.0 -ohjelman kautta. Myös muut vapaaehtoiset eettiset ohjeet tulisi kytkeä osaksi syrjinnän ehkäisyä.

Liite 1: Sitaatteja haastatteluista

5) Missä tehtävissä käytetään tekoälyä työnteon tukena? Käytetäänkö tekoälyä esimerkiksi päätöksentekoprosessien apuna?

> *"Ainoastaan apuväline, aina virkamies tekee lopullisen päätöksen. Pakkokeinolain rikoksista epäiltyjen henkilöiden tietokannassa voi olla viisikin vuotta vanha kuva. Painoa voi tulla tai kadota tai voi alkoholisoitua tai parta kasvaa – siksi aina ihminen lopuksi mukana. Annettu kattava morphologinen koulutus. Kasvojentunnistustieteessä tarkat kriteerit eri kasvojen osien tunnistamiseen." – P*

9) Implementoidaanko tekoälyratkaisut sisäisesti organisaation itsensä toimesta vai käytetäänkö niiden tuottamiseen ulkoista palveluntarjoajaa?

> *" Ostettu algoritmi, joka on kehitetty omana sovelluksena sopivaksi. Taustalla kaupallinen algoritmi, mutta itse sovellettu omaan tarkoitukseen." – P*

Voiko organisaationne tekoälyjärjestelmien käyttö syrjiä ihmisiä sukupuolen, iän, etnisen alkuperän, seksuaalisen suuntautumisen, poliittisen mielipiteen tai jonkun muun vastaavan syyn perusteella?

> *"Ei, ja tuskin tulevaisuudessakaan. Rikosasiain tietosuojalaki säätää tätä (säädetty 2018); vaikutustenarviointivaatimus, jonka mukaan poliisin on kuultava tietosuojavaltuutettua. Siinä kontrolli, jossa jos vaikka muutetaan biometrisesti tunnistettavaan muotoon kuvia, katsotaan mitä ja miten voi muuntaa. Vaatimukset tietosuojavaltuutetulta tarkkoja, mutta ainahan sitä voi olla riskejä. Riskejä kuitenkin aina, käytetään järjestelmiä tai ei." – P*

3) Mistä tekoälyjärjestelmiin liittyvät syrjivät vaikutukset voisivat syntyä?

> *"On joukko dataa, jolla koulutetaan järjestelmää toisaalta on vasteita joihin reagoidaan. Esim. keskustelupalstan hyväksyttävä/hylättävä kommentti/ viesti. On aiempia, ihmisten tekemiä päätöksiä, joiden pohjalta tekoälyä opetetaan tekemään valintoja.*
>
> *Aineisto harvoin tasapainossa potentiaalisten syrjivien ominaisuksien suhteen. Saattaa ryhtyä korreloimaan asioiden kanssa, jotka ovat em. listalla. Jos esim. rikoksen tekijöitä on paljon jossain alipopulaatiossa, voi olla, että tekoäly oppii helpommin ottamaan juuri näitä tarkkailuun tai valitsemaan heitä." - V*

5) Minkälaisia monimuotoisuutta ja eri näkökulmien ymmärtämistä edistäviä käytäntöjä organisaatiossasi on käytössä erityisesti teknologiapainotteisissa toiminnoissa ja projekteissa?

> *"Aihetta nostetaan esiin diversiteettikeskusteluissa. Koko henkilöstölle pidettäviä tilaisuksia kerran kuussa, joissa nostetaan aiheita usein esiin. Lakisääteiset suunnitelmat miten käydään läpi yhdenvertaisuuteen liittyviä asioita. Kartoitetaan sitä, miten ollaan koettu työssä, onko syrjintää." – V*

1) Minkälaista arviointia mahdollisista syrjivistä vaikutuksista on tehty? (Onko mahdollisten syrjivien vaikutusten löytämiseksi ja minimoimiseksi laadittu strategiaa tai suunnitelmaa? Onko arvioitu syötettävän datan ja algoritmien suunnittelun vaikutuksia, yhdessä tai erikseen?)

> *"Aikalailla rajoittuu tekoälyn testaamiseen. Tuodaan testikeissejä ja katsotaan mitä vasteita tulee; mustaa laatikkoa ei voi "ymmärtää", joten täytyy käyttää todellisia käyttötapauksia joita järjestelmälle voitaisiin syöttää. Yleinen testaussuunnitelma. Tulee sattumaltakin vastaan ongelmakohtia, aina testaus ei riitä. Ihan kaikkea ei voida ottaa ennalta huomioon. Syrjimättömyyden testaussuunnitelmaa ei ole." – V*

4) Minkälaisen (reiluuden) määritelmän mukaan syrjivyyttä tarkastellaan organisaatiossasi? Onko se yleisesti käytetty ja hyväksytty näkökulma riittävän syrjimättömyyden varmistamiseksi? Onko reiluuden, yhdenvertaisuuden ja syrjimättömyyden määrittelystä käyty keskustelua sidosryhmien kanssa, esim. ikäihmisten, kehitysvammaisten tai muiden mahdollisesti haavoittuvassa asemassa olevien tahojen/heitä edustavien tahojen kanssa?

> *"Esim. ihmisten profilointi, jossa halutaan tietää kaikki mahdolliset yksilöivät tiedot, jotta voitaisiin esim. pikavippimainontaa kohdistaa alttiisiin henkilöihin. Kehittäjillä myös aika laaja vapaus ilmoittaa, ettei halua osallistua, jos arveluttava projekti.*
>
> *Ei ole tiedossa, että ainakaan kesken projektin ei ole keskeytetty. Keskeytykset hyvin aikaisessa vaiheessa. Mukana myyjä ja team-liidi (ja tiimi?) jotka keskustelevat siitä, halutaanko lähteä mukaan projektiin vai ei." – V*

11) Yhdenvertaisuusanalyysi: demografista dataa ja sen pohjalta lopputulemat:
a) Onko mallia kehitettäessä tai testattaessa pääsyä demografiseen dataan tai syrjintäperustaiseen dataan?

"Aika usein on. Tämä täytyy kertoa asiakkaalle toisinaan erikseen, että eivät anna pääsyä kaikkeen dataan. Asiakkaatkaan eivät aina tiedä mitä dataa voidaan käyttää tai ei. Joskus näkyvissä nimi+hetu+pankkitiedot; asiakkaalle tulee tällöin kertoa, että ei tällaista dataa voi käyttää.

Tämä kovin uutta. Dataa konemuodossa ei ole kuitenkaan (ainakaan tässä määrin ja tällä tarkkuudella) kovin pitkään ollut, joten siksikään asiakkaat eivät aina tiedä mitä dataa voi jakaa eteenpäin." – V

12) Miten syrjintäriskejä arvioidaan ja pyritään estämään algoritmin tai tilastollisen mallin tasolla? Käytetäänkö esim. tilastollisia testejä ja mittareita? Jos kyllä, millä metodologialla? Jos ei, miksi ei?

"Datalle tehdään kartoitus, jos vaikka olisi muuttuja "sukupuoli", onko aineisto tasapainossa, vai onko se vinoutunutta; tarkistetaan kaikkien yksittäisten muuttujien suhteen. Jos data ei ole tasapainossa, se vaikuttaa helposti päätöksiin, tehdään kaikissa vaiheissa opettamista.

Voidaan generoida uutta, samasta tilastosta olevaa dataa, joka on tasapainoisempaa alkuperäisen datan pohjalta. Mallissa voi myös olla painokertoimia, joilla saadaan korjattua datan tulkintaa." – V

3) Millaisia tekoälyteknologioita organisaatiossanne käytetään tai aiotaan käyttää?

"Tekoälypohjainen chat, NLP, käytössä. Puhekomponentti myös mukana. Ei ole vielä otettu tuotantoon. Maturiteettia bisnes-sanastoon tarvitaan. Puheesta-tekstiksimuuttokonversio vielä hiukan hakusessa. Tarkoitus viedä myös asiointiin.

Koneoppimista myös skenaario- ja simulaatiorakentamisessa. Miten muuttuvasta datasta voi tehdä näitä tulevaisuuteen. 2016 alussa mallinnettiin turvapaikkahakijoiden tilanteesta. Datan perusteella muuttuva, ei staattinen malli. Oli käytössä.

Päätöksenteossa on kokeiltu, kuinka hyvin tekoälypohjainen malli arvioisi hakemuksen kompleksisuutta. Todennäköisyyksiä sen pohjalta. Vasta testivaiheessa. 4 RPA-tyyppistä automaatiota ja käsittelyä. Syrjivät vaikutukset –teeman alla perinteinenkin ohjelmistorobotiikka myös relevanttia, vaikka ei ole tekoälyä, vaan perinteisiä algoritmeja."
-M

5) Missä tehtävissä käytetään tekoälyä työnteon tukena? Käytetäänkö tekoälyä esimerkiksi päätöksentekoprosessien apuna?

> *"Ei päätöksentekoprosessien apuna "puhtaimmillaan ymmärrettynä". Avustavina palveluina on tulossa. Tulkinta ja johtopäätökset jäävät aina ihmiselle. Automaattista päätöksentekoa ei ole, eikä lain mukaan voi tullakaan. Tehdään automaattinen päätöksentekopohja, jonka ihminen käy läpi ja hyväksyy/hylkää." - M*

1) Voiko organisaationne tekoälyjärjestelmien käyttö syrjiä ihmisiä sukupuolen, iän, etnisen alkuperän, seksuaalisen suuntautumisen, poliittisen mielipiteen tai jonkun muun vastaavan syyn perusteella?

> *"Periaatteessa ei, koska lainsäädäntö on syrjimätöntä, ja Migri noudattaa lainsäädäntöä. Lakia tarkastellaan erittäin tarkasti, ja Migrin järjestelmät noudattavat näitä. Kielteisen päätöksen saaminen voidaan kokea syrjiväksi, vaikka olisi lainmukainen. Jos on, tekoäly on rakennettu lainsäädännön viitekehyksen ulkopuolelle." -M*

7) Mistä organisaationne käyttämien tekoälyjärjestelmien riski syrjiä muodostuu? Mihin syrjinnän riskit liittyvät?

> *"Itseoppivia kun ei ole, niin vähenee. Riski piilee tekoälyn kouluttamisessa. Vastuu mistä datasta itse oppii, on ihmisen vastuulla. Ei voi sanoa että "nyt meidän tekoäly vaan oppi näin"." -M*

1) Automaattisen päätöksenteon lainsäädäntöhanke. Minkälaisia rajoja automaation käytölle ko. organisaation kohdalla he toivoisivat, että tulee tarkoittamaan? Kuinka paljon ja mitä kaikkea lainsäädännön tulisi mahdollistaa automaattisen päätöksenteon osalta? Minkälaiset rajoitukset ovat perusteltuja?

> *"Olisi hyvä luoda yleinen viitekehys sille, mitä organisaatiolta odotetaan kun se hyödyntää automatisaatiota. Sopivalla tarkkuudella julkista, jotta asiakkaat tietävät miten asioita käsitellään. Suostumus pitää olla. Monopoliasema vaikuttaa tähän – joka julkishallinnossa usein on, täälläkin.*
>
> *Pitää olla realistisia reunaehtoja. Automatisaation hyödyntäminen ei saa vaatia enemmän efforttia kuin mitä siitä saadaan hyötyjä.*
>
> *Päätös pitäisi voida tehdä organisaation nimissä, ja virkavastuu rakennetaan taustalla jollain tavalla." - M*

8) Mitä organisaatiosi tehnyt pitääkseen huolta, että tekoälyn käyttö olisi mahdollisimman läpinäkyvää järjestelmän sidosryhmille?

"Asiakaspalvelurajapinta läpinäkyvä kun chatbotin kanssa puhuu, muuten ei relevantti. Kun automaattista päätöksentekoa ei tehdä, niin ei olla loppuun asti viety. Automaatiokokonaisuus avataan julkisesti sitten kun/jos tätä voidaan tehdä. On täysin välttämätöntä, että kuvataan, avataan, mutta millä tasolla, on vielä auki. Liian detaljitasoinen avaaminen voi vaikuttaa siihen, että järjestelmää voidaan käyttää hyväksi." -M

b) Onko mallia kehitettäessä tai testattaessa pääsyä dataan "todellisista lopputulemista" (ns. ground truth data)? Jos ei ole pääsyä, mistä tämä johtuu? Onko tälle oikeudellisia, organisatorisia tai teknisiä esteitä?

"Oikealla datalla kun simulaatioita tehdään, tulee "todellisia lopputulemia", mutta simulaatio tekee sitten mallia tulevaisuuteen. Dynaamisissa ilmiöissä käytettävä." -M

4) Mikä on tekoälysovelluksen käytön päätarkoitus?

"Ideana on helpottaa vahinkokäsittelijöiden työtä; tekoäly osaa luokitella ja katsoa mitä kuuluu mihinkin; jos jotain vaikka tiedoista puuttuu, niin pystyisi poimimaan puuttuvia osia.

Vakuutusten suosittelussa myynnin lisäämistä ja myös kohdentamista (tälläkin myyntiä). Myös sääntöpohjaisia algoritmeja, jotka saattavat kaivaa mm. sähköposteista lisätietoa."-Va

2) Miten näet teidän tai vastaavien organisaatioiden roolin tasa-arvon ja yhdenvertaisuuden edistämisessä?

"Kaikkien vastuulla tehdä tasa-arvon ja yhdenvertaisuuden edistämistä. Korvausten puolella erityisesti pitää olla tarkkana. Siellä data voisi antaa vääristyneen kuvan helpommin. Jos data ei huomioi jotain, miten se vaikuttaa malliin – ja miten sen voisi korjata." - Va

3) Mistä tekoälyjärjestelmiin liittyvät syrjivät vaikutukset voisivat syntyä?

"Meistä, yhteiskunnasta. Tarvittavaan dataan ei ole pääsyä. Erityisesti tietysti potilastiedot. Datan sisältö itsessään myös voi olla vinoutunutta." – Va

7) Mistä organisaationne käyttämien tekoälyjärjestelmien riski syrjiä muodostuu? Mihin syrjinnän riskit liittyvät?

> *"Tekoälyn käytön ei pitäisi vaikuttaa vahinkokorvauspäätöksiin. Virheitä tai puutteita tiedoissa voi olla. Esim. korvaus matkapuhelimesta viisi kertaa vuoden sisään, voi tulla kysely että miksi näin monta? Voidaan laskea "syrjinnäksi" vaikka olisi esim. iso perhe, ja päävakuutettu yksi henkilö kotivakuutuksessa.*
>
> *Suurin osa vapaata tekstiä, tekoäly tulkitsee sitä. Lyhytsanaisia helpommin voitaisiin syrjiä, koska tekoäly ei saa irti samaa tietoa kuin pitemmistä selvityksistä.*
>
> *Voi olla, että pienemmät korvaukset pääsevät helpommin läpi.*
>
> *Voi olla mahdollisuus ei-suomen- tai ruotsinkielisten käsittelyn hankaluudesta. Menee käsittelijälle, jos ei ymmärrä." - Va*

3) Mainitut lainsäädäntöhankkeet: Uskotteko, että mainitut hankkeet voivat auttaa syrjimättömyyden takaamisessa?

> *"Toivottavasti, muuten kuulostaa turhalta. Huoli, että Suomessa on niin paljon regulaatiota kaikesta, että Kiinat ja vastaavat kehittävät sitten meistä riippumattomasti. Pitää olla tarkkana, ettei tule vain sieltä." -Va*

2) Minkälaisia prosesseja, työkaluja ja kehikoita syrjivyyden minimoimiseksi on käytetty tai suunnitellaan käytettävän? (Minkälaisia testaus- tai monitorointikeinoja käytetään mahdollisten syrjivien vaikutusten löytämiseksi ja arvioimiseksi? Onko prosessit suunniteltu syrjivyyden arvioimiseksi tekoälyjärjestelmien koko elinkaaren ajaksi?)

> *"Ei tietoinen tästä näkökulmasta; pitää käydä haastattelemassa kehittäjiä aiheesta!" -Va*

7) Kuinka tietoisia kehitysprosesseihin osallistuvat henkilöt ovat syrjimättömyyden tematiikasta?

> *"Uskon, että ovat tietoisia, mutta "tällaista dataa meillä on, sitä nyt vaan on käytettävä"." -Va*
>
> *"Algoritmin muodostuksessa pyritty imitoimaan mahdollisimman diskriminoimatta oikean maailman tilannetta. Algoritmi seuloo potilaiden*

kokonaismassaa ja muodostamiensa algoritmien perusteella etsii osumia. Tekoäly ei kuitenkaan (saa) päät(tä)ä mitään. Asiantuntijaa edelleen kuitenkin vielä tarvitaan perusteluun. Ihminen kumminkin vielä fiksumpi, vaikka raaka laskentakyky tekoälyllä onkin kova. Intuitiota kuitenkin edelleen tarvitaan. Algoritmi: laskentakyky, seulontakyky; Asiantuntija: tarkistus ja päätös. Sitten Finndatan nimeämä asiantuntija arvioi hälytyksen, sitten konsultoi THL:n nimeämää kahta asiantuntijaa. Jos kaikki nämä viisi tahoa ovat samaa mieltä, lähtee hälyte potilaan alueellisesti hoidosta vastaavalle lääkärille, joka lähettää potilaalle tiedon nopeutetusti mahdollisesta riskistä. Sitten potilas tekee päätöksen." - H

"Ellei käyttö koske koko väestöä, voi olla syrjiviä vaikutuksia. Opt-out -järjestelmä tästä syystä." - H

"Microsoft Azure -ympäristössä, anonymisoituna, josta ne voidaan tupla-anonymisoida. Tiukasti rajoitettu pääsy. Anonymisointia ei pureta ellei tule viranomaismääräystä." - H

"1) Miten ymmärrät / miten organisaatiossasi ymmärretään tekoälyn eettisen tai vastuullisen kehittämisen yleisesti? (Yleisempi kysymys AI-etiikasta; mitä siitä keskustellaan organisaatiossa? Mitä AI-etiikan teemoja vastaaja nostaa esille?)

"Tekoälyn eettinen hyödyntäminen tulevaisuuden kannalta, kyllä mä sen, että se on vastuullinen toiminnan kulmakivi. Eli näen, että meillä ei ole vaihtoehtoa. Tehtiin pari vuotta sitten isomman organisaatiomuutoksen yhteydessä (jolloin innovaatiotoiminta myös käynnistettiin) niin ensimmäinen asia ennen kuin yhtään tekoälyratkaisua tehtiin, niin luotiin nuo tekoälyn eettiset periaatteet." - K

2) Minkälaisia prosesseja, työkaluja ja kehikoita syrjivyyden minimoimiseksi on käytetty tai suunnitellaan käytettävän? (Minkälaisia testaus- tai monitorointikeinoja käytetään mahdollisten syrjivien vaikutusten löytämiseksi ja arvioimiseksi? Onko prosessit suunniteltu syrjivyyden arvioimiseksi tekoälyjärjestelmien koko elinkaaren ajaksi?)

"Meillä on tuollainen etiikka-alusta/-kehys Saidot.ai:n kanssa tehtynä, ja sitä sovelletaan soveltuvin osin. Ja ylätasolla eettiset periaatteet, johon liittyy myös erilaisten vastuiden hahmottamista. Nämä muodostavat jotakuinkin työkalun/prosessit arviointiin." -K

Liite 2: Haastattelukysymykset

1 Yleinen osuus

Haastateltavan tiedot ja kuvaus organisaation tekoälyjärjestelmistä

Toimialue ja työnkuva

1. Voitko kertoa organisaationne päätoiminnoista?
2. Mikä on roolisi ja työtehtäväsi organisaatiossanne?

Tekoälyn käyttötarkoitus ja pääsovellusalue

3. Millaisia tekoälyteknologioita organisaatiossanne käytetään tai aiotaan käyttää?
4. Mikä on tekoälysovelluksen käytön päätarkoitus?
5. Missä tehtävissä käytetään tekoälyä työnteon tukena? Käytetäänkö tekoälyä esimerkiksi päätöksentekoprosessien apuna?
6. Mitkä tehtävät tekoäly tulee suorittamaan itsenäisesti?
7. Mikä on tekoälyjärjestelmien autonomisuuden taso (perinteiset algoritmit, oppivat algoritmit, koneoppiminen)?
8. Voitaisiinko nämä tehtävät suorittaa ilman tekoälyn apua? Mitä hyötyä tekoälyn käytöstä on?
9. Implementoidaanko tekoälyratkaisut sisäisesti organisaation itsensä toimesta vai käytetäänkö niiden tuottamiseen ulkoista palveluntarjoajaa?

2 Tekoälyjärjestelmien mahdolliset syrjivät vaikutukset

Yleisluontoisempi katsaus siihen, miten haastateltava näkee tekoälyteknologioihin ja -järjestelmiin liittyvät syrjinnän riskit.

1. Miten ymmärrät / miten organisaatiossasi ymmärretään tekoälyn eettisen tai vastuullisen kehittämisen yleisesti? (Yleisempi kysymys AI-etiikasta; mitä siitä keskustellaan organisaatiossa? Mitä AI-etiikan teemoja vastaaja nostaa esille?)
2. Voiko tekoälyn käyttö johtaa ihmisten syrjintään sukupuolen, iän, etnisen alkuperän, seksuaalisen suuntautumisen, poliittisen mielipiteen tai jonkun muun vastaavan syyn perusteella?
3. Mistä tekoälyjärjestelmiin liittyvät syrjivät vaikutukset voisivat syntyä?

3 Haastateltavan organisaation tekoälyjärjestelmien mahdolliset syrjinnät riskit

Syvennytään yleisessä osuudessa haastateltavan listaamiseen ko. organisaation käyttämiin tai suunnittelemiin tekoälyjärjestelmiin sekä niihin liittyviin syrjinnän riskeihin.

1. Voiko organisaationne tekoälyjärjestelmien käyttö syrjiä ihmisiä sukupuolen, iän, etnisen alkuperän, seksuaalisen suuntautumisen, poliittisen mielipiteen tai jonkun muun vastaavan syyn perusteella?
2. Miten näet teidän tai vastaavien organisaatioiden roolin tasa-arvon ja yhdenvertaisuuden edistämisessä?
3. Onko organisaatiossanne yleisesti selvää, mitkä ovat kiellettyjä syrjintäperusteita?
4. Minkälainen on organisaationne vastuu ja näkökulma syrjimättömyyteen liittyen?
5. Minkälaisia monimuotoisuutta ja eri näkökulmien ymmärtämistä edistäviä käytäntöjä organisaatiossasi on käytössä erityisesti teknologiapainotteisissa toiminnoissa ja projekteissa?
6. Kuinka hyvin monimuotoiset näkökulmat näkyvät tekoälyjärjestelmien kehityksessä?
7. Mistä organisaationne käyttämien tekoälyjärjestelmien riski syrjiä muodostuu? Mihin syrjinnän riskit liittyvät?
8. Missä kohtaa järjestelmän kehitystä ja käyttöä syrjintään liittyviä riskejä tunnistetaan tai niitä pitäisi tunnistaa? (Missä vaiheessa tuotanto- tai hankintaketjua?)

4 Lainsäädännöllinen tilanne

Miten tietoisia sen vaikutuksesta tekoälyjärjestelmien syrjimättömyyteen liittyen (kansallinen ja EU- lainsäädäntö).

1. Automaattisen päätöksenteon lainsäädäntöhanke. Minkälaisia rajoja automaation käytölle ko. organisaation kohdalla he toivoisivat, että tulee tarkoittamaan? Kuinka paljon ja mitä kaikkea lainsäädännön tulisi mahdollistaa automaattisen päätöksenteon osalta? Minkälaiset rajoitukset ovat perusteltuja?
2. Oletteko ehtineet tutustua Euroopan komission tekoälyteknologioita koskevaan lainsäädäntöehdotukseen? Ehdotus asettaa paljon vaatimuksia tekoälyn soveltamiseen; miten organisaatiossanne on suhtauduttu näin tiukkaan ja monimutkaiseen lainsäädäntökokonaisuuteen?
3. Mainitut lainsäädäntöhankkeet: Uskotteko, että mainitut hankkeet voivat auttaa syrjimättömyyden takaamisessa?

5 Keinot syrjivyyden minimoimiseksi

Paneudutaan siihen, mitä ko. organisaatio on tehnyt tai on tekemässä syrjimättömyyden ja yhdenvertaisuuden edistämisessä tekoälyjärjestelmien kehittämisen ja käytön kontekstissa.

1. Minkälaista arviointia mahdollisista syrjivistä vaikutuksista on tehty? (Onko mahdollisten syrjivien vaikutusten löytämiseksi ja minimoimiseksi laadittu strategiaa tai suunnitelmaa? Onko arvioitu syötettävän datan ja algoritmien suunnittelun vaikutuksia, yhdessä tai erikseen?)
2. Minkälaisia prosesseja, työkaluja ja kehikoita syrjivyyden minimoimiseksi on käytetty tai suunnitellaan käytettävän? (Minkälaisia testaus- tai monitorointikeinoja käytetään mahdollisten syrjivien vaikutusten löytämiseksi ja arvioimiseksi? Onko prosessit suunniteltu syrjivyyden arvioimiseksi tekoälyjärjestelmien koko elinkaaren ajaksi?)
3. Kuka/mikä osasto organisaatiossanne vastaa mahdolliseen syrjintään liittyvistä asioista? Mikä on työnjako eri rooleissa toimivien työntekijöiden välillä: johto, syrjintään liittyvistä asioista vastaava(t) henkilö(t), hankinta, teknologiatiimi(t)?
4. Minkälaisen (reiluuden) määritelmän mukaan syrjivyyttä tarkastellaan organisaatiossasi? Onko se yleisesti käytetty ja hyväksytty näkökulma riittävän syrjimättömyyden varmistamiseksi? Onko reiluuden, yhdenvertaisuuden ja syrjimättömyyden määrittelystä käyty keskustelua sidosryhmien kanssa, esim. ikäihmisten, kehitysvammaisten tai muiden mahdollisesti haavoittuvassa asemassa olevien tahojen/heitä edustavien tahojen kanssa?
5. Millä keinoin eri ihmisryhmät, joihin tekoälyjärjestelmän käyttö saattaa vaikuttaa, identifioitiin?
6. Kuinka paljon ja minkälaista sidosryhmätyöskentelyä (ideointi-, suunnittelu-, toteutusvaiheessa) ja - kuulemista tehtiin tekoälyjärjestelmiä kehitettäessä?
7. Kuinka tietoisia kehitysprosesseihin osallistuvat henkilöt ovat syrjimättömyyden tematiikasta?
8. Mitä organisaatiosi tehnyt pitääkseen huolta, että tekoälyn käyttö olisi mahdollisimman läpinäkyvää järjestelmän sidosryhmille?
9. Minkälaisia valitusprosesseja ja -kanavia käyttäen mahdollisista syrjivistä vaikutuksista / syrjintäepäilyistä voi ottaa yhteyttä? Kuka käsittelee valitukset ja palautteet syrjivyyteen liittyen? Minkälaisiin toimenpiteisiin tämän kaltaiset valitukset johtaisivat?
10. Voisivatko viranomaiset avustaa organisaatiotanne syrjimättömyyden varmistamisen osalta enemmän? Jos, miten? Jos ei, miksi ei?

6 Lisäosuus

Osuus, joka käydään läpi aikarajoituksesta ja haastateltavan syventyväisyyden tasosta riippuen

11. Yhdenvertaisuusanalyysi: demografista dataa ja sen pohjalta lopputulemat:

    a) Onko mallia kehitettäessä tai testattaessa pääsyä demografiseen dataan tai syrjintäperustaiseen dataan? b) Onko mallia kehitettäessä tai testattaessa pääsyä dataan ”todellisista lopputulemista” (ns. ground truth data)? Jos ei ole pääsyä, mistä tämä johtuu? Onko tälle oikeudellisia, organisatorisia tai teknisiä esteitä?

12. Miten syrjintäriskejä arvioidaan ja pyritään estämään algoritmin tai tilastollisen mallin tasolla? Käytetäänkö esim. tilastollisia testejä ja mittareita? Jos kyllä, millä metodologialla? Jos ei, miksi ei?

Liite 3: Osatehtävä 2. kartoituksen aineisto

Artificial Intelligence and Gender Inequality: Key Findings of UNESCO's Global Dialogue. (2020). UNESCO.

Artificial intelligence in hiring. Assessing impacts on equality. (2020). Institute for the Future of Work.

Algorithmic decision-making. (2018). Algo:aware.

Algorithm-driven Hiring Tools: Innovative Recruitment or Expedited Disability Discrimination?. (2020). Center for Democracy & Technology.

Algorithmic Bias Explained: How Automated Decision-making Becomes Automated Discrimination. (2021). The Greenlining Institute.

Algorithmic discrimination in Europe. Challenges and opportunities for gender equality and non-discrimination law. (2021). Gerards, J., Xenidis, R. European network of legal experts in gender equality and non-discrimination.

Algorithms: Please Mind the Bias!. (2020). Institut Montaigne.

An Intelligence in Our Image The Risks of Bias and Errors in Artificial Intelligence. (2017). Osonde, O. & Welser W., RAND Corporation.

Beyond Debiasing. *Regulating AI and its inequalities.* (2021). European Digital Rights.

Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. (2016). Executive Office of the President. The White House, Washington.

Data Capitalism + Algorithmic Racism. (2021). Data for Black Lives & Demos.

Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights. (2019). Euroopan unionin perusoikeusvirasto (European Union Agency for Fundamental Rights).

Discrimination, artificial intelligence, and algorithmic decision-making. (2018). F. Z. Borgesius.

Gender and Intersectional Bias in Artificial Intelligence. (2020). Euroopan komissio.

GETTING THE FUTURE RIGHT: ARTIFICIAL INTELLIGENCE AND FUNDAMENTAL RIGHTS. (2020). Euroopan unionin perusoikeusvirasto (European Union Agency for Fundamental Rights).

How to Prevent Discriminatory Outcomes in Machine Learning. (2018). World Economic Forum [White paper].

Racial Bias in Natural Language Processing. (2019). Shearer, E., et al., Oxford Insights [Research report].

Racial discrimination and emerging digital technologies: a human rights analysis. Report of the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance. (2020). Achiume, E. T, YK:n yleiskokous (Human rights council).

REGULATING FOR AN EQUAL AI: A NEW ROLE FOR EQUALITY BODIES. Meeting the new challenges to equality and non-discrimination from increased digitisation and the use of Artificial Intelligence. (2020). Allen QC, R. & Masters, D. Equinet.

Review into bias in algorithmic decision-making. (2020a). Centre for Data Ethics and Innovation.

AI Barometer. (2020b). Centre for Data Ethics and Innovation.

Risks of Discrimination through the Use of Algorithms A study compiled with a grant from the Federal Anti-Discrimination Agency. (2020). Orwat, C., Institute for Technology Assessment and Systems Analysis (ITAS), Karlsruhe Institute of Technology (KIT).

Using artificial intelligence to make decisions: Addressing the problem of algorithmic bias. (2020). Australian ihmisoikeuskomissio.

Liite 4: Osatehtävä 3. vertailevan analyysin aineisto

Seuraavat kehikot ja raportit tunnistettiin kehitettävälle arviointekehikolle relevanteiksi resursseiksi ja mittatikuiksi:

Actionable Intelligence for Social Policy AISP. (2020). A Toolkit for Centering Racial Equity Throughout Data Integration.

Ada Lovelace Institute, AI Now Institute & OGP. (2021). Algorithmic accountability for the public sector.

AI Now Institute (2018). Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability.

American Council for Technology-Industry Advisory Council (ACT-IAC). (2020). ETH-ICAL Application of Artificial Intelligence Framework. ACT-IAC White paper.

Australian Human Rights Commission (2020). Using artificial intelligence to make decisions: Addressing the problem of algorithmic bias.

Bertelsmann Stiftung. (2020). From Principles to Practice: An interdisciplinary framework to operationalise AI ethics.

BSA | The Software Alliance. (2021). Confronting Bias: BSA's Framework to Build Trust in AI.

Centre for Data Ethics and Innovation. (2021). Ethics, Transparency and Accountability Framework for Automated Decision-Making.

Central Digital and Data Office and Office for Artificial Intelligence. (2019). A guide to using artificial intelligence in the public sector.

Central Digital and Data Office. (2020). Data ethics framework.

Chicago Booth. (2021). Algorithmic Bias Playbook.

ECP: IIA - Artificial Intelligence Impact Assessment (2019).

European Commission, AI HLEG. (2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI).

European Parliament Research Services. (2019). A governance framework for algorithmic accountability and transparency.

Government of Canada: Algorithmic Impact Assessment tool (2019).

Government of Netherlands. (2022). Fundamental Rights and Algorithms Impact Assessment (FRAIA).

GovEx, City of San Francisco, Harvard DataSmart & Data Community DC. (2018). Ethics & Algorithms Toolkit.

HAAS Berkeley. (2020). Mitigating Bias in Artificial Intelligence An Equity Fluent Leadership Playbook.

National Institute of Standards and Technology NIST. (2021). A Proposal for Identifying and Managing Bias in Artificial Intelligence

National Institute of Standards and Technology NIST. (2022). Towards a Standard for Identifying and Managing Bias in Artificial Intelligence

National Institute of Standards and Technology NIST. (2022). AI Risk Management Framework: Initial Draft.

Open Loop. (2021). AI Impact Assessment: A Policy Prototyping Experiment.

Organisation for Economic Co-operation and Development OECD. (2022). Framework for Classification of AI Systems.

Recruitment and Employment Confederation REC. (2021).: Data-driven tools in recruitment guidance.

Tilburg University. (2021). Handbook on non-discriminating algorithms.

The Alan Turing Institute. (2022). Human Rights, Democracy, and the Rule of Law Assurance Framework for AI Systems: A proposal prepared for the Council of Europe's Ad hoc Committee on Artificial Intelligence

U.S. Government Accountability Office (GAO). (2021.) Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities.

World Economic Forum. (2020). Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations.

World Economic Forum. (2020). AI Procurement in a Box: Workbook

Liite 5: Hankkeen asiantuntijaryhmä ja työpajat

Hankkeen kansainvälinen asiantuntijaryhmä tapasi kolme kertaa hankkeen aikana sen tutkimuksellisen laadun varmistamiseksi. Siihen osallistuivat seuraavat henkilöt:

- Renata Ávila, CEO at Open Knowledge Foundation & Affiliate at Stanford HAI.
- Antti Poikola, Leader of Finnish AI Accelerator, Technology Industries of Finland / MyData Global founder
- Jarmo Eskelinen, Director of the Data-Driven Innovation, University of Edinburgh.
- Tiina Valonen, Lawyer at Non-Discrimination Ombudsman Finland.
- Laurens Naudts, Researcher at the KU Leuven Centre for IT & IP Law.
- Marina Dimova, Governance Chief Technical Specialist at UNDP.
- Raphaële Xenidis, Lecturer in EU Law at Edinburgh Law School.
- Janneke Gerards, Professor of fundamental rights law at Utrecht University.
- Teemu Roos, Professor of Computer science and lead instructor Elements of AI at University of Helsinki.

Hankkeen kahteen avoimeen yhteiskehittämistyöpajaan osallistui asiantuntijoita muun muassa suomalaisista yliopistoista (Helsinki, Turku, Tampere, Jyväskylä, Oulu, Aalto), ministeriöistä (VNK, TEM, OM, LVM, STM), virastoista (Digi- ja väestötietovirasto, Tilastokeskus, Terveyden ja hyvinvoinnin laitos, Kela, Poliisi, Business Finland), kaupungeista (Helsinki, Jyväskylä), sekä yksittäisistä yrityksistä ja kansalaisyhteiskunnan järjestöistä.

Lähteet

Ada Lovelace Institute, AI Now Institute and Open Government Partnership. (2021). Algorithmic Accountability for the Public Sector. https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/

Ailisto, H., Neuvonen, A., Nyman, H., Halén, M., & Seppälä, T. (2019). Tekoälyn kokonaiskuva ja kansallinen osaamiskartoitus–loppuraportti. Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja 4/2019.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. Conference on Fairness, Accountability and Transparency, s. 149-159. PMLR.

Binns, R. (2020). On the apparent conflict between individual and group fairness. Proceedings of the 2020 conference on fairness, accountability, and transparency, s. 514-524.

Borgesius, F.Z. (2018). Discrimination, artificial intelligence, and algorithmic decision-making. Strasbourg: Council of Europe.

Borgesius, F.Z. (2020.) Strengthening legal protection against discrimination by algorithms and artificial intelligence. The International Journal of Human Rights. Vol 24: 10. 1572-1593.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Conference on fairness, accountability and transparency, s. 77-91. PMLR.

Calvo, R. A., Peters, D., & Cave, S. (2020). Advancing impact assessment for intelligent systems. Nature Machine Intelligence, 2(2), s. 89-91.

Chhabra, A., Masalkovaitė, K., & Mohapatra, P. (2021). An Overview of Fairness in Clustering. IEEE Access.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data, 5(2), 153-163.

Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. Washington Law Review, 89. https://scholarship.law.bu.edu/faculty_scholarship/618?utm_source=scholarship.law.bu.edu%2Ffaculty_scholarship%2F618&utm_medium=PDF&utm_campaign=PDFCoverPages.

Danks, D., & London, A. J. (2017). Algorithmic Bias in Autonomous Systems. IJCAI, Vol. 17, s. 4691-4697.

Datainspektionen. (2021). Beslut 2019-08-20: Tillsyn enligt EU:s dataskyddsförordning 2016/679 – ansiktsigenkänning för närvarokontroll av elever.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science advances, 4(1). https://doi.org/10.1126/sciadv.aao5580.

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. Communications of the ACM, 63(1), 68-77.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference (pp. 214-226).

Elahi, M., Abdollahpouri, H., Mansoury, M., & Torkamaan, H. (2021). Beyond Algorithmic Fairness in Recommender Systems. Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, s. 41-46.

Euroopan komissio, AI HLEG (2019). Ethics guidelines for trustworthy AI. Brussels.

Euroopan komissio. (2020). European enterprise survey on the use of technologies based on artificial intelligence. Luxembourg.

Euroopan komissio. (2021). COM/2021/206 final.

Euroopan parlamentti ja Euroopan unionin neuvosto. (2016). Regulation (EU) 2016/679.

Euroopan unionin perusoikeusvirasto. (2020). Getting the future right: Artificial intelligence and fundamental rights. ISBN 978-92-9474-860-7.

Fazelpour, S., Lipton, Z. C., & Danks, D. (2021). Algorithmic Fairness and the Situated Dynamics of Justice. Canadian Journal of Philosophy, 1-17.

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. Philosophy Compass, 16(8). https://doi.org/10.1111/phc3.12760.

Finland's AI Accelerator. (2020). State of AI in Finland report.

Garattini, C., Raffle, J., Aisyah, D. N., Sartain, F., & Kozlakidis, Z. (2019). Big data analytics, infectious diseases and associated ethical impacts. Philosophy & technology, 32(1), 69-85.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), s. 86-92.

Gerards, J., & Xenidis, R. (2021). Algorithmic discrimination in Europe: Challenges and opportunities for gender equality and non-discrimination law. Publications Office of the European Union.

Green, B., & Hu, L. (2018). The myth in the methodology: Towards a recontextualization of fairness in machine learning. Proceedings of the machine learning: the debates workshop.

Green, B. And Chen, Y. (2019). Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments'. FAT* '19: Conference on Fairness, Accountability, and Transparency (FAT* '19).

Green, B., & Chen, Y. (2021). Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), s. 1-33.

Green, B. (2021a). Escaping the" Impossibility of Fairness": From Formal to Substantive Algorithmic Fairness. arXiv preprint arXiv:2107.04642.

Green, B. (2021b). The Flaws of Policies Requiring Human Oversight of Government Algorithms. https://dx.doi.org/10.2139/ssrn.3921216.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Proceedings of the 30th International Conference on Neural Information Processing Systems, 3323-3331.

Hoch, H., Hertweck, C., Loi, M., & Tamò, A. (2021). Discrimination for the Sake of Fairness: Fairness by Design and Its Legal Framework. https://ssrn.com/abstract=3773766.

Hokkanen, T. (2021). Kasvojentunnistusteknologia ja sen riskit. Haaga-Helia ammattikorkeakoulu, Opinnäytetyö Tietojenkäsittely.

Holland, S., Hosny, A., & Newman, S. (2020). The dataset nutrition label. Data Protection and Privacy: Data Protection and Democracy, 1.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need?. Proceedings of the 2019 CHI conference on human factors in computing systems, 1-16.

Hutchinson, B., & Mitchell, M. (2019). 50 years of test (un) fairness: Lessons for machine learning. Proceedings of the Conference on Fairness, Accountability, and Transparency, 49-58. https://doi.org/10.1145/3287560.3287600.

Institute for the Future of Work. (2021). Algorithmic Impact Assessments: Building a systematic framework of accountability for algorithmic decision making. Policy Briefing.

Jin, Z., Xu, M., Sun, C., Asudeh, A., & Jagadish, H. V. (2020). Mithracoverage: a system for investigating population bias for intersectional fairness. Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, 2721-2724.

Kaminski, M. E., & Malgieri, G. (2020). Multi-layered explanations from algorithmic impact assessments in the GDPR. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, s. 68-79.

Kaminski, M., Malgieri, G. (2021). Algorithmic impact assessments under the GDPR: producing multi-layered explanations, International Data Privacy Law, Volume 11, Issue 2. Pages 125–144, https://doi.org/10.1093/idpl/ipaa020.

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1-33.

Kamiran, F., Karim, A., & Zhang, X. (2012). Decision theory for discrimination-aware classification. 2012 IEEE 12th International Conference on Data Mining, 924-929. 10.1109/ICDM.2012.45;

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 35-50. 10.1109/ICDMW.2011.83;

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. International Conference on Machine Learning, 2564-2572. PMLR.

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2019). An empirical study of rich subgroup fairness for machine learning. Proceedings of the Conference on Fairness, Accountability, and Transparency, 100-109.

Keinänen, A. ja Pajuoja, J. (2020), Miten vaikutusten arviointia voitaisiin parantaa? Vaikutusarviointi ja sen kehittämistarpeet suomalaisessa lainvalmistelussa, Eduskunnan tarkastusvaliokunnan julkaisu 1/2020.

Kim, P. (2017). Auditing Algorithms for Discrimination, 166 U. Pa. L. Rev. Online

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. The quarterly journal of economics, 133(1), 237-293.

Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Business Research, 1-54.

Lee, M. S. A., Floridi, L., & Singh, J. (2021). Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. AI and Ethics, 1-16.

Lee, M. S. A., & Singh, J. (2021). The landscape and gaps in open source fairness toolkits. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1-13.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. Entropy, 23(1), 18.

Mann, M., & Matzner, T. (2019). Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to emergent discrimination. Big Data & Society, 6(2), 2053951719895805.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the conference on fairness, accountability, and transparency, s. 220-229.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. Annual Review of Statistics and Its Application, 8, 141-163.

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable machine learning–a brief history, state-of-the-art and challenges. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, s. 417-431. Springer, Cham.

Morley, J., Machado, C. C., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: A mapping review. Social Science & Medicine, 113172.

Moss, E., Watkins, E., Singh, R., Elish, M., & Metcalf, J. (2021). Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. https://ssrn.com/abstract=3877437

Narayanan, A. (2018). 21 Fairness Definitions and Their Politics. Conference on Fairness, Accountability, and Transparency [Demonstraatio].

Nepelski, D. and Sobolewski, M. (2020). Estimating investments in General Purpose Technologies: The case of AI Investments in Europe, EUR 30072 EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-76-10233-5 (online), doi:10.2760/506947 (online), JRC118953

Ntoutsi, E., Fafalios, P., Gadiraju, U., et al. Bias in data-driven artificial intelligence systems—An introductory survey. WIREs Data Mining Knowl Discov. 2020; 10:e1356.

Oxford Insights (2020). AI Readiness Index 2020.

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. Frontiers in Big Data, 2, 13

Passi, S., & Barocas, S. (2019). Problem formulation and fairness. Proceedings of the Conference on Fairness, Accountability, and Transparency, 39-48.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2020). Data and its (dis) contents: A survey of dataset development and use in machine learning research. ArXiv E-Prints, arXiv:2012.05345. https://arxiv.org/abs/2012.05345

Peng, K., Mathur, A. & Narayanan, A. (2021). Mitigating dataset harms requires stewardship: Lessons from 1000 papers. https://arxiv.org/abs/2108.02922v1.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). Algorithmic impact assessments: A practical framework for public agency accountability. AI Now Institute, 1-22.

Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. NYUL Rev. Online, 94, 15.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), s. 206-215.

Sahlgren, O. (2021). The politics and reciprocal (re) configuration of accountability and fairness in data-driven education. Learning, Media and Technology, 1-14.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the conference on fairness, accountability, and transparency, 59-68.

Selbst, A. D. (2021). An Institutional View Of Algorithmic Impact Assessments. 35 Harvard Journal of Law & Technology 117, UCLA School of Law, Public Law Research Paper No. 21-25, https://ssrn.com/abstract=3867634

Stoyanovich, J., & Howe, B. (2019). Nutritional labels for data and models. A Quarterly bulletin of the Computer Society of the IEEE Technical Committee on Data Engineering, 42(3).

Suresh, H., & Guttag, J. V. (2020). A framework for understanding unintended consequences of machine learning. ArXiv E-Prints, arXiv:1901.10002. https://arxiv.org/abs/1901.10002

The Alan Turing Institute. (2021). Human Rights, Democracy, and the Rule of Law Assurance Framework for AI Systems: A proposal prepared for the Council of Europe's Ad hoc Committee on Artificial Intelligence. https://rm.coe.int/huderaf-coe-final-1-2752-6741-5300-v-1/1680a3f688

Verma, S. & Rubin, J. (2018). Fairness Definitions Explained. 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), pp. 1-7, doi: 10.1145/3194770.3194776.

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).

Wachter, S., Mittelstadt, B., & Russell, C. (2020). Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. W. Va. L. Rev., 123, 735.

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. Computer Law & Security Review, 41. https://doi.org/10.1016/j.clsr.2021.105567.

Xiang, A., & Raji, I. D. (2019). On the legal compatibility of fairness definitions. arXiv preprint arXiv:1912.00761.

Zarsky, T. Z. (2014). Understanding discrimination in the scored society. Washington Law Review, 89.

Zhang, B. H., Lemoine, L., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 335-340. https://doi.org/10.1145/3278721.3278779;

tietokayttoon.fi

ISBN PDF 978-952-383-404-0
ISSN PDF 2342-6799