TURUN YLIOPISTO

Koneoppimisalgoritmien tekemä syrjintä rekrytoinnissa

Tietojenkäsittelytiede
Tietotekniikan laitos, Teknillinen tiedekunta
Kandidaatintutkielma

Laatija:
Petra Kuisma

Tammikuu 2025

Kandidaatintutkielma
Tietotekniikan laitos, Teknillinen tiedekunta
Turun yliopisto

Tutkinto-ohjelma: Tietojenkäsittelytiede
Tekijä: Petra Kuisma
Otsikko: Koneoppimisalgoritmien tekemä syrjintä rekrytoinnissa
Sivumäärä: 31 sivua
Päivämäärä: Tammikuu 2025

Koneoppimisalgoritmien käyttö rekrytoinnissa on yleistynyt viime vuosien aikana. Tekoäly tarjoaa rekrytoinnissa mahdollisuuksia prosessien tehostamiseen ja automatisointiin, mutta tuo samalla mukanaan eettisiä haasteita. Tässä tutkielmassa tarkastellaan koneoppimisalgoritmien käyttöä rekrytoinnissa. Ensimmäisenä tavoitteena on tarkastella sitä, millaisia ongelmia algoritmien käyttö on aiheuttanut rekrytoinnin kontekstissa. Toisena tavoitteena on tarkastella, millaisilla tekniikoilla koneoppimisalgoritmien aiheuttamaa syrjintää voidaan rekrytoinnissa vähentää. Tutkielma toteutettiin kirjallisuuskatsauksena.

Tutkielmassa havaitaan, että koneoppimisalgoritmien käyttö rekrytoinnissa voi aiheuttaa ongelmia koulutusaineiston epäedustavuuden, algoritmien suunnittelun vinoumien sekä päätöksenteon läpinäkymättömyyden takia. Näillä ongelmilla on erityisen merkittäviä vaikutuksia työnhakijoiden esikarsinta- ja valintavaiheissa. Lisäksi tuloksissa havaitaan, että algoritmien tekemää syrjintää voidaan vähentää kolmella keskeisellä tekniikalla: esikäsittelyllä, prosessoinninaikaisella käsittelyllä ja jälkikäsittelyllä. Esikäsittelyssä korostuvat aineiston muokkaus ja sensitiivisten piirteiden poistaminen ja piilottaminen, kun taas prosessoinninaikaisessa käsittelyssä algoritmin toimintalogiikkaa pyritään muokkaamaan oikeudenmukaisempaan suuntaan. Jälkikäsittelyssä joko tasapainotetaan algoritmin tuottamia tuloksia oikeudenmukaisuusmittarien avulla tai lisätään algoritmien selitettävyyttä ja läpinäkyvyyttä.

Syrjinnän vähentämisen tekniikoilla on saatu muokattua algoritmeja oikeudenmukaisempaan suuntaan, mutta tästä huolimatta syrjinnän täydellinen poistaminen algoritmeista on haastavaa ja käytetyt tekniikat voivat vaikuttaa haitallisesti algoritmien ennustetarkkuuteen. Tutkielman tulosten perusteella algoritmien käyttö rekrytoinnin karsinta- ja valintavaiheissa vaatii kriittistä tarkastelua, sillä algoritmien toiminta voi toisintaa tai jopa vahvistaa yhteiskunnallista epätasa-arvoisuutta.

Asiasanat: tekoäly, koneoppiminen, rekrytointi, syrjintä, oikeudenmukaisuus

Sisällysluettelo

1 Johdanto

Tekoäly ja sen keskeinen osa-alue koneoppiminen ovat tuoneet merkittäviä toimintatapojen muutoksia monelle toimialalle, kuten rekrytointiin. Rekrytoinnissa tekoälyä on alettu hyödyntää muun muassa lupaavien työnhakijoiden etsimiseen, työhakemusten läpikäyntiin ja videohaastattelujen analysointiin. Koneoppimisalgoritmit pystyvät analysoimaan suuria tietomääriä, mikä on auttanut yrityksiä tehostamaan hakemuksien käsittelyä. Prosessien sujuvoittamisen lisäksi algoritmien systemaattisuuden on odotettu lisäävän rekrytoinnin objektiivisuutta ja läpinäkyvyyttä, sillä algoritmien toiminta perustuu opittuihin malleihin pikemminkin kuin inhimillisiin arvioihin. [1] Vaikka tekoälyn avulla voidaan saavuttaa järjestelmällisempiä rekrytointimenetelmiä, on kuitenkin huomattu, että tekoälytyökalut voivat toisintaa tai jopa vahvistaa rekrytoinnissa valmiiksi olemassa olevia syrjinnän muotoja [2], [3], mikä herättää tarpeen pohtia tekoälyn käytön eettisyyttä rekrytoinnissa. Rekrytoinnin kontekstissa on keskitytty muun muassa sukupuoleen, ikään ja etnisyyteen kohdistuvaan syrjintään [1], [2], [3], [4], [5], [6]. Eettiset kysymykset liittyvät paitsi syrjinnän mahdollisuuteen myös siihen, miten tekoälytyökalujen päätöksenteon avoimuus ja selitettävyys voidaan taata [4]. Automaattisten järjestelmien käyttöä säädellään lisäksi hallinnollisesti, minkä takia työkalujen syrjintään ja avoimuuteen tulee kiinnittää huomiota. Esimerkiksi Euroopan unionin tietosuoja-asetus (GDPR) säätää selitysoikeudesta, jonka mukaan kansalaisilla on oikeus pyytää selitys heistä tehtyisin algoritmisiin päätöksiin ja vaatii myös toimenpiteitä syrjivien vaikutusten estämiseksi arkaluonteista tietoa käsiteltäessä [7].

Tämän tutkielman tavoitteena on tarkastella koneoppimisalgoritmien tekemää syrjintää rekrytoinnissa. Tutkielmassa tarkastellaan ensinnäkin sitä, miten koneoppimista on aikaisemmin hyödynnetty rekrytoinnissa ja millaisia ongelmia siitä on seurannut. Toiseksi tutkielmassa tarkastellaan koneoppimismallien tekemän syrjinnän vähentämisen tekniikoita. Tarkastelun keskiössä on algoritmien mahdollisesti aiheuttama epäsuora institutionaalinen ja rakenteellinen syrjintä, jossa tekoälyn voi ajatella uusintavan kehittäjänsä, käyttäjänsä ja laajemmin yhteiskunnassa vaikuttavia oletuksia reiluudesta ja oikeudenmukaisuudesta. Tutkielmassa ei kuitenkaan käsitellä kaikkia tekoälyn ja koneoppimisen eettisiä ulottuvuuksia. Esimerkiksi juridiset ja sääntelyyn liittyvät kysymykset sekä tietoturva ja yksityisyys rajataan tutkielman ulkopuolelle. Myöskään tekoälyn käytön taloudellisia näkökulmia ei käsitellä. Tutkielman tavoitteena on keskittyä erityisesti syrjinnän ja oikeudenmukaisuuden näkökulmiin.

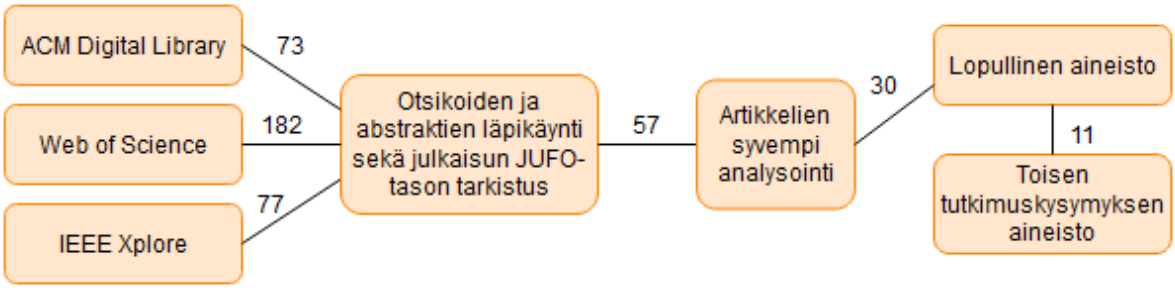
Tutkimusongelmaa lähestytään kahden tutkimuskysymyksen kautta:

TK1 Mitä ongelmia koneoppimisalgoritmien käytöstä on ilmennyt rekrytoinnissa?

TK2 Millä tekniikoilla rekrytoinnissa ilmenevää koneoppimisalgoritmien tekemää syrjintää voidaan vähentää?

Tutkielma toteutettiin kirjallisuuskatsauksena. Tiedonhaussa käytetyt tietokannat olivat ACM Digital Library, Web of Science ja IEEE Xplore. Tutkielman tiedonhaun kulkua on havainnollistettu kuvassa 1. Jokaisessa tietokannassa hakukierroksia tehtiin kaksi. Ensimmäisellä kierroksella otsikosta haettiin tietoa hakulausekkeella ("machine learning" OR "artificial intelligence" OR AI OR algorithm) AND (recruit* OR hiring OR employ* OR resume OR interview) ja abstraktista lisäksi lausekkeella (bias OR *fair*). Toisella hakukierroksella otsikosta haettiin tietoa hakulausekkeella (recruit* OR hiring OR employ* OR resume OR interview) AND (bias OR *fair*) ja abstraktista hakulausekkeella ("machine learning" OR "artificial intelligence" OR AI OR algorithm). Rekrytointia kuvaavien hakutermien rajaamisella otsikkoon haluttiin varmistaa, että hakutuloksissa keskityttiin rekrytointiin eikä syrjintään yleisellä tasolla.

Haun tekemisen jälkeen tuloksista luettiin ensin läpi otsikot ja abstraktit. Kriteereinä tutkimuksen valitsemiselle jatkoon oli se, että tutkimuksessa käsiteltiin koneoppimisalgoritmien tekemää syrjintää rekrytoinnissa. Toisen tutkimuskysymyksen kohdalla kriteerinä oli, että tutkimuksessa sovellettiin jotain syrjinnän vähentämisen algoritmista menetelmää rekrytoinnin kontekstissa. Sopivista tutkimuksista otettiin mukaan ne, jotka oli julkaistu JUFO-tasoltaan vähintään tason 1 julkaisussa. Artikkelien syvemmässä analysoinnissa lopullisen aineiston määräksi saatiin yhteensä 30 tutkimusartikkelia tai seminaarijulkaisua, joista toisen tutkimuskysymyksen aineistoksi valikoitui 11 tutkimusta.



Kuva 1. Tiedonhakuprosessi.

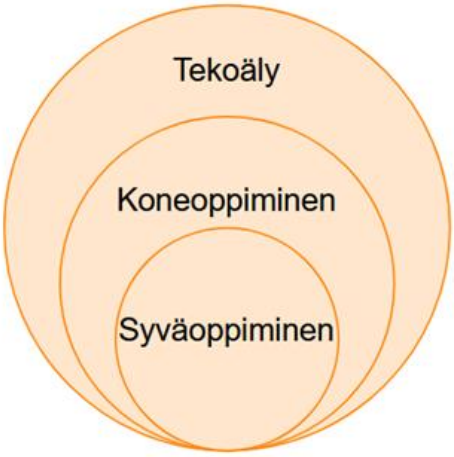
Tutkielma rakentuu johdannon lisäksi kuudesta pääluvusta ja yhteenvedosta. Luvussa 2 esitellään ensin koneoppimista tekoälyn osa-alueena sekä käydään läpi tutkielman kannalta olennaiset koneoppimisen käsitteet. Tämän jälkeen luvuissa 3 ja 4 tarkastellaan syrjinnän käsitettä sekä syrjintään ja oikeudenmukaisuuteen liittyviä eettisiä näkökulmia. Luvussa 5 keskitytään tekoälyavusteisen rekrytoinnin tuomiin haasteisiin, minkä jälkeen luvussa 6 tarkastellaan koneoppimisalgoritmien tekemän syrjinnän vähentämiseen käytettyjä tekniikoita rekrytoinnissa. Luvussa 7 pohditaan tutkielman tuloksia ja luvussa 8 on tutkielman yhteenveto sekä vastaukset tutkimuskysymyksiin.

2 Mitä tekoäly on?

Tekoälystä on 2010-luvulta lähtien tullut yhä merkittävämpi osa teknologista kehitystä koneiden nopeutumisen ja datan määrän valtavan kasvun seurauksena [8, s. 2]. Laveasti miellettynä tekoälyn historia yltää jopa varhaismoderniin 1600- ja 1700-lukujen mekaniikkaan ja erityisesti laskukoneisiin ja automaatteihin sekä niiden herättämiin filosofisiin pohdintoihin. Modernina tieteellisenä projektina tekoäly alkaa kuitenkin konkretisoitua 1900-luvun puolessa välissä. [9]

Vaikka tekoälyyn perustuvia tai sitä käyttäviä sovellutuksia esiintyykin nykyään hyvin laajalti, ei sille kuitenkaan ole muodostunut yhtä ainoaa vakiintunutta määritelmää [9]. Kielitoimiston sanakirja määrittelee tekoälyn koostuvan ”tietokoneen toiminnoista, jotka jäljittelevät ihmiselle tyypillisiä älykkyyttä vaativia toimintoja” [10]. Toisaalta tekoälyn on sanottu olevan itsessään älykästä: yksi esimerkki tällaisesta määritelmästä on, että tekoäly on älyä, joka havainnollistuu algoritmien kautta [11, s. 7]. Tekoälyllä voidaan kuitenkin viitata konkreettisen teknologisen sovelluksen lisäksi myös laajemmin älykkäitä järjestelmiä ja niiden kehittämistä tutkivaan tieteenalaan [12]. Tekoälyä voidaan siis yhtäältä lähestyä käytännöllisessä mielessä apuvälineenä tiettyjen teknologioiden kehittämiseen, toisaalta sitä voidaan ajatella jopa yhtenä älykkyyden muotona. Jälkimmäinen lähestymistapa on herättänyt keskustelua siitä, mitä älykkyys ylipäätään on ja voiko koneella olla älyä, vai onko älykkyys vain ihmisten ja eläinten ominaisuus [11], [12]. Määritelmäeroista riippumatta jokaiseen lähestymistapaan näyttää yhdistyvän koneen kyky suorittaa monimutkaisia tehtäviä ilman ihmisen jatkuvaa valvontaa ja ohjausta sekä kyky oppia menneisyydessä tehdyistä virheistä koneen suorituskykyä parantaen [13, s. 651].

Tekoälyn osa-alueisiin kuuluu koneoppiminen, neuroverkot sekä syväoppiminen, joiden suhteutumista toisiinsa on kuvattu kuvassa 2. Tässä kappaleessa keskitytään näihin osa-alueisiin, ensin kone- ja syväoppimisen ja sitten tekoälyyn tiiviisti liittyvään luonnollisen kielen esittelyyn. Samalla esitellään kirjallisuuskatsauksen tulosten kannalta oleellisia tekoälyn menetelmiä.



Kuva 2. Tekoälyn osa-alueet Tuomisen ym. [8] mukaan.

2.1 Koneoppiminen ja syväoppiminen

Koneoppimisessa kone oppii toistuvista malleista ilman ihmisen jatkuvaa puuttumista oppimisprosessiin [8, s. 6]. Koneoppimisalgoritmit oppivat löytämään aineistossa toistuvia rakenteita ja käyttävät niitä ennusteiden tekemiseen ja luokitteluun. Algoritmit toimivat ihmisen antaman tehtävän mukaisesti ja muokkaavat toimintaansa autonomisesti oppimisen edetessä. [11, ss. 83–84] Koneoppimista sovelletaan monella alalla, kuten kasvojentunnistuksessa, markkinoinnissa, opiskelijavalinnoissa ja rekrytoinnissa. Myös esimerkiksi sosiaalisen median mainosten takana on koneoppimismalleja, jotka perustavat suosituksensa käyttäjän selaushistoriaan. Oppimisprosessi perustuu *koulutusaineistoon*, jonka avulla algoritmit oppivat suorittamaan haluttua tehtävää. Koulutusaineiston lisäksi algoritmeille syötetään erillistä *testiaineistoa* arvioimaan algoritmin yleistämiskykyä. [14, s. 179] Algoritmin koulutuksen jälkeen sitä voidaan käyttää halutun tehtävän parissa, kuten edellä mainittuun sosiaalisen median mainosten ja muidenkin sisältöjen räätälöityyn kohdentamiseen.

Koneoppiminen jaetaan oppimisen tyylin perusteella ohjattuun oppimiseen, ohjaamattomaan oppimiseen ja vahvistettuun oppimiseen [8, s. 6]. *Ohjatussa oppimisessa* aineisto sisältää ohjelmoijan valmiiksi luokittelemia syöte–tavoite-pareja ja kone oppii luokittelemaan saamaansa aineistoa samankaltaisiin pareihin [13, s. 653]. Ohjattu oppiminen voi olla menetelmiltään joko luokittelua tai regressiota aineiston luonteen mukaan: diskreetin aineiston tapauksessa käytetään luokittelua ja jatkuvassa aineistossa taas regressiota. Esimerkki luokittelusta on roskapostisuodatin, jonka ohjelmoija kouluttaa aluksi

luokittelemaan tietyn tyyppiset viestit roskapostiksi. Algoritmin koulutuksen edetessä algoritmi oppii päättelemään itse, mitkä viestit ovat roskapostia. Esimerkkejä regressiosta ovat taas lämpötilan tai tuotteen hinnan määrittäminen. [8, s. 13] *Ohjaamattomassa oppimisessa* ohjelmoija ei syötä koneelle valmiiksi luokiteltua aineistoa, vaan kone luokittelee aineistoa itse oman logiikkansa mukaisesti [11, ss. 85–86]. Kone yrittää tällöin löytää aineistosta riippuvuuksia ja samankaltaisuuksia ja ryhmittelee aineistoa klustereiksi löytämiensä tulosten perusteella. Ryhmittelyn tuloksena saman ryhmän sisällä olevilla entiteeteillä pitäisi olla keskenään enemmän samankaltaisia piirteitä kuin toiseen ryhmään kuuluvien entiteettien kanssa. [8, s. 13] Esimerkkinä ohjaamattomasta oppimisesta voi käyttää kuvantunnistusalgoritmia, jossa koneelle syötetään miljoonia kuvia, kone tunnistaa kuvista kissan kaltaisia hahmoja ja muodostaa niistä sitten yhden klusterin [13, s. 653]. Kolmannessa kategoriassa, *vahvistusoppimisessa,* koneelle annetaan palautetta sen tuottaman tuloksen mukaan. Esimerkiksi shakinpeluualgoritmissa koneelle voitaisiin antaa positiivista palautetta voiton jälkeen ja negatiivista palautetta häviön yhteydessä. Tällöin kone oppii muuttamaan voittoa tai häviötä edeltäneitä toimintoja saamansa palautteen perusteella. [13, s. 653]

Oppimisen tyyli ja valittu algoritmi valitaan käsillä olevan ongelman mukaan. Päätökseen vaikuttavat muun muassa käsiteltävän datan laatu ja määrä. [8, s. 14] Suuren aineistomäärän käsittelyyn soveltuvat algoritmit perustuvat erityisesti *syväoppiville neuroverkoille*. Neuroverkot ovat ihmisen ja eläinten aivoissa olevia hermosolujen muodostamia verkostoja [14, s. 245]. Tekoälyn kontekstissa neuroverkot ovat näitä verkostoja jäljitteleviä teknisiä rakenteita. Neuroverkot ovat hyödyllinen väline ison aineistomäärän hallitsemiseen ja neuroverkot ovatkin yleistyneet 2010-luvulta lähtien aineiston määrän kasvun seurauksena. [8, ss. 6–7] Neuroverkot rakentuvat eri tehtäviä suorittavista kerroksista. Moderneissa neuroverkkoteknologioissa kerroksia on useita, jolloin niistä muodostuva kuvio on syvä ja monikerroksinen. Neuroverkkojen lähestymistapaa tekoälyteknologioihin kutsutaankin syväoppimiseksi. [15, ss. 1–2]

Yksinkertaisin neuroverkko, *perseptroni*, koostuu kahdesta kerroksesta, syöte- ja ulostulokerroksesta. Monimutkaisemmissa neuroverkoissa on piilokerroksia syöte- ja ulostulokerrosten välissä ja näitä verkkoja kutsutaan monikerroksisiksi perseptroneiksi (engl. multilayer perceptron, MLP). [15, s. 5] Jokaisella kerroksella sijaitsee useita kerrosten matemaattiset operaatiot suorittavia neuroneita. Neuroverkon tekemät matemaattiset operaatiot perustuvat yhdistettyyn laskentaan, jossa syötekerros ottaa ensin vastaan neuroverkolle syötettävän aineiston. Piilo- ja ulostulokerroksen neuroneissa lasketaan

syötteiden painotettu summa, johon lisätään lopuksi neuronin vakiotermi. Summa viedään lopuksi aktivointifunktioon, josta saatu tulos lähetetään verkossa eteenpäin seuraavalle neuronille. [8, s. 24] Neuroverkot oppivat koulutuksen yhteydessä syötteiden ja tulosteiden välisen suhteen, minkä jälkeen ne pystyvät yleistämään tietoja ja tuottamaan aineistosta ennusteita [16, ss. 24–25].

Monikerroksisessa neuroverkossa syöte etenee kerros kerrokselta kohti ulostulokerrosta siten, että edellisen kerroksen neuronien tulosteet toimivat toisen neuronikerroksen neuronien syötteinä. Monikerroksiset neuroverkot koulutetaan kolmivaiheisella takaisinkytkentäalgoritmilla (engl. back propagation algorithm). Kun syötteet ovat kulkeneet verkon läpi, ulostulokerroksen tulosta verrataan haluttuun lopputulokseen virheen laskemiseksi. Koulutuksen toisessa vaiheessa verkon painoja säädetään virheen perusteella taaksepäin edeten. Näitä vaiheita toistetaan, kunnes verkko alkaa tuottaa haluttuja tuloksia. [16, ss. 56–58] Monikerroksisia neuroverkkoja hyödynnetään nykyään laajasti monissa eri sovelluksissa, kuten luonnollisen kielen käsittelyä vaativissa tehtävissä, joissa koneen tulee oppia käsittelemään kielen rakenteita ja tekemään niistä ennusteita [13, s. 856]. Seuraavassa alaluvussa kuvaillaan luonnollisen kielen käsittelyn keskeisiä piirteitä.

2.2 Luonnollisen kielen käsittely

Luonnollisen kielen käsittely (engl. Natural Language Processing, NLP) on monitieteinen tutkimusala, joka yhdistää kielitietieteen, tietojenkäsittelytieteen, psykologian, kognitiotieteen ja neurotieteen tekstiaineistojen käsittelyssä. Luonnollisen kielen käsittely voidaan määritellä monin eri tavoin [8, s. 64], mutta tässä tutkielmassa luonnollisen kielen käsittelyllä tarkoitetaan ihmiskielten automaattista käsittelyä ja yhdistelyä. Käsittely voi tarkoittaa yksinkertaisempia, säännöllisillä lausekkeilla tehtyjä merkkijonojen mallintamisalgoritmeja tai kehittyneempiä neuroverkkoja toiminnassaan hyödyntäviä järjestelmiä [17, s. 3]. Tekoälyn yhteydessä luonnollisen kielen käsittely liittyy pyrkimykseen opettaa koneita suorittamaan erilaisia kieleen liittyviä tehtäviä ja oppimaan luonnollisen kielen rakenteen syväoppivien neuroverkkojen avulla [13]. Koneoppimismenetelmien käyttö luonnollisen kielen käsittelyssä on ollut hallitseva lähestymistapa 2000-luvulta lähtien, kun tätä ennen käytettiin lähinnä diskreettejä tilastollisia menetelmiä [17, s. 4].

Luonnollisen kielen käsittelyssä neuroverkon syötteenä käytetään sanojen vektoripohjaisia esityksiä ja neuroverkkoa koulutetaan taaksepäin suuntautuvan takaisinkytkennän avulla. Vektorit voidaan muodostaa joko laskemalla vektorin arvo manuaalisesti (tekstin koodaus,

engl. text encoding) tai kouluttamalla neuroverkkoa (sanan vektorointi, engl. word vectorization). *Tekstin koodauksessa* teksti pyritään muuttamaan numeeriseen muotoon, jotta sitä voidaan käsitellä ja analysoida tietokoneella. Tekstin koodaus voidaan tehdä pienille aineistoille luokittelemalla aineistoa yksitellen sana kerrallaan, kun taas suuremmille aineistoille käytetään sanojen esiintymistiheyteen tai indeksiin perustuvaa koodausta. *Sanan vektoroinnissa* taas jokainen sana esitetään neuroverkon avulla luotuna vektorina eli sanaupotteena (engl. word embedding). [18, ss. 82–101] Sanaupotteita käytetään löytämään usein yhdessä esiintyviä sanoja. Esimerkiksi sanat ”maali”, ”turnaus” ja ”seura” liittyvät urheiluaiheisiin, kun taas ”osake”, ”laina” ja ”toimitusjohtaja” talousaiheisiin. Vektoriavaruusesitystä käyttämällä pyritään arvioimaan sanojen samankaltaisuutta mittaamalla niiden etäisyyksiä vektoriavaruudessa. [17, s. 50]

Nykyään luonnollisen kielen käsittelyyn pohjautuvia tekoälysovelluksia on useita, joista ehkä yleisesti tunnetuimpia ovat chatbotit, kuten ChatGPT, sekä kielen kääntämiseen tarkoitetut koneet, kuten DeepL. Luonnollisen kielen käsittelyyn pohjautuvia tekoälysovelluksia on integroitu niin ihmisen arkeen kuin liiketoimintaan muokaten päätöksentekotapojamme. Esimerkiksi liiketoiminnassa organisaatiot pystyvät käsittelemään paljon aikaisempaa isompia tekstiaineistoja nopeammin ja tehokkaammin, jolloin tiettyjä ennen manuaalisesti tehtyjä analyyseja on ulkoistettu tekoälysovellusten tehtäväksi. [17, s. 3]

3 Syrjintä ja oikeudenmukaisuus

Syrjinnän käsitteelle ei ole yhtä yleisesti omaksuttua ja ilmiön kaikkia aspekteja tyhjentävästi huomioivaa määritelmää [19]. Erään näkemyksen mukaan syrjintää voi lähestyä käytäntönä, toimintana tai politiikkana, joka asettaa syrjittyjä yksilöitä epäedulliseen asemaan suhteessa ei-syrjittyihin sen perusteella, että syrjityt kuuluvat tai heidän katsotaan kuuluvan johonkin merkittävään (engl. salient) sosiaaliseen ryhmään. Tällöin toiminta, käytäntö tai politiikka kohdistuu joihinkin henkilöihin sen perusteella, että he ovat tai heidän arvioidaan olevan osa tällaista ryhmää. [19], [20] Esimerkkejä tällaisista sosiaalisista ryhmistä ovat sukupuoli, etnisyys, uskonto ja seksuaalinen suuntautuminen, mutta relevantteja ryhmien jäsentymisen tapoja voi olla muitakin eikä tutkimuskirjallisuudessa ole yksimielisyyttä siitä, mitkä ryhmistä ovat merkittäviä ja miksi [19]. Ryhmiä voidaan hahmottaa myös suojeltujen perusteiden tai suojattujen ominaisuuksien (engl. protected grounds) käsitteen kautta: tällöin syrjinnän kieltävä lainsäädäntö ikään kuin suojaa tietyt ominaisuudet, joita ei saa käyttää arviointiperusteena esimerkiksi rekrytointitilanteessa [21]. Syrjinnässä kyse ei ole kuitenkaan jaottelusta ryhmiin sinänsä, vaan jaottelusta seuraavasta haitasta, joka kohdistuu tiettyihin yksilöihin ja ryhmiin, mutta ei toisiin. Syrjinnän käsittelyssä vertailu ja suhteellisuus ovatkin merkittävässä asemassa ja ne selittävät, miksi syrjintä on lähellä tasa-arvon käsitettä. Suhteellinen haitta implikoi loogisesti epätasa-arvoa syrjittyjen ja vertailuluokan välillä. [19]

Syrjinnän käsitteen tarkemmaksi tarkastelemiseksi ja syrjinnän muotojen jaottelemiseksi on kehitetty useita erilaisia alakäsitteitä ja näkökulmia. *Suoralla syrjinnällä* tarkoitetaan tarkoituksellista eriarvoista kohtelua yksilön ryhmäjäsenyyden perusteella. Tietyn etnisen ryhmän edustajien pääsyn kieltäminen ravintolaan pelkästään heidän etnisen taustansa vuoksi on suoran syrjinnän selkeä esimerkki. Suorasta syrjinnästä voi kuitenkin olla kyse myös silloin, jos toiminta tai politiikka kohdistuu haitallisesti tiettyyn ryhmään, vaikka ryhmää ei suoranaisesti nimettäisikään. [19] Toisen näkökulman mukaan suorasta syrjinnästä on kyse, kun suojeltujen ominaisuuksien osalta ihmisiä kohdellaan vähemmän suopeasti kuin toisia kohdellaan, on kohdeltu tai kohdeltaisiin vastaavassa tilanteessa [21]. Kirjallisuudessa on myös esitetty, että suoraa syrjintää voi esiintyä silloin, kun toiminnan tosiasiallisena tarkoituksena ei ole syrjiä. Suoraa syrjintää voi tämän näkemyksen mukaan olla myös välinpitämätön asenne mahdollisesti relevanttien ryhmien oikeuksia kohtaan tai yleinen ennakkoluuloisuus, joka aiheuttaa suhteellisen haitan ryhmälle ja sen jäsenille. Tätä kutsutaan myös syrjinnäksi erilaisen kohtelun kautta (engl. disparate treatment discrimination). [19]

Konkreettinen esimerkki syrjinnästä erilaisen kohtelun kautta on tilanne, jossa työnantaja hylkää naispuolisen hakijan insinöörin tehtävään, koska uskoo, etteivät naiset sovi teknisiin ammatteihin. [22, s. 13] *Epäsuorassa syrjinnässä* syrjiviä käytäntöjä ei ole suunnattu suoraan tiettyä ryhmää ja sen jäseniä vastaan, mutta toimi tai valitut käytännöt vaikuttavat silti suhteettomasti tähän ryhmään – vaikka tahallisuutta, ongelmallisia ennakkoluuloja tai välinpitämättömyyttä ei olisikaan läsnä [19]. Toisen näkökulman mukaan epäsuoraa syrjintää esiintyy tilanteessa, jossa näennäisesti neutraali järjestely, kriteeri tai toiminta asettaisi ihmiset, joilla on suojeltuja ominaisuuksia, vahingolliseen asemaan suhteessa muihin, ellei tuo järjestely, kriteeri tai toiminta ole oikeutettu objektiivisesti legitiimillä tavoitteella ja keinot ovat myös sekä kohtuullisia että tarpeellisia [21]. Epäsuoraa syrjintää kutsutaan myös syrjinnäksi erilaisten vaikutuksien kautta (engl. disparate impact discrimination). Esimerkki syrjinnästä erilaisten vaikutusten kautta on rekrytoinnissa käytetty kykytesti, joka vaikuttaa ulkoisesti neutraalilta mutta jonka tulokset heijastavat koulutusjärjestelmän eriarvoisuuksia, minkä vuoksi tietty etninen ryhmä menestyy heikommin. Jos testin käytölle ei ole painavaa liiketoiminnallista perustetta tai jos on olemassa vähemmän syrjivä vaihtoehto, on kyse syrjinnästä erilaisten vaikutusten kautta. [22, s. 14]

Syrjintää voi myös seurata kollektiivisen toiminnan myötä eikä vain yksittäisten henkilöiden toimesta. *Organisationaalisesta syrjinnästä* on kyse, kun kollektiiviset toimijat, kuten yritykset, yliopistot tai valtiolliset virastot, toimivat syrjivästi. Organisationaalinen syrjintä voi olla sekä suoraa että epäsuoraa. *Rakenteellisella* tai *institutionaalisella syrjinnällä* pyritään taas kuvaamaan tilannetta, jossa yhteiskunnan lait, säännöt tai tavat johtavat systemaattisesti tiettyjen ryhmien epäedulliseen asemaan. Institutionaalisen syrjinnän muodot liittyvät organisaatioiden ja yhteiskuntarakenteiden luomiin esteisiin, jotka asettavat tietyt ryhmät huonompaan asemaan pitkällä aikavälillä. Vaikka tällaisten syrjivien sääntöjen historiallisena taustana usein saattaakin olla tarkoituksellinen toiminta organisationaalisen syrjinnän muodossa, rakenteellisella syrjinnällä pyritään kuvaamaan korostetusti epäsuoraa syrjintää, joka on seurausta yhteiskunnan vakiintuneiden sääntöjen usein piilevistäkin vaikutuksista ja joka eroaa käsitteellisesti yksilöiden tai organisaatioiden tekemästä suorasta syrjinnästä. [19]

Erillisenä syrjinnän muotona voi pitää myös niin kutsuttua välillistä syrjintää eli syrjintää välittäjän kautta (eng. discrimination by proxy), jonka voi mieltää eräänlaiseksi välitapaukseksi suoran ja epäsuoran syrjinnän välillä. Tällaista syrjintää on aikaisemmin tutkittu muun muassa juuri koneoppimisalgoritmien näkökulmassa. Välillistä syrjintää

tapahtuu, kun algoritmeja hyödyntävä rekrytointisysteemi käyttää näennäisesti neutraalia ominaisuutta korvaajana suojellulle perusteelle ja välittäjän käyttö johtaa tiettyyn vahinkoon niille hakijoille, jotka assosioituvat suojeltuihin perusteisiin. Välillisessä syrjinnässä ongelmana on se, että algoritmit voivat rakentaa välittäjiä, jotka korreloivat ryhmäjäsenyyden kanssa ja tehdä rekrytointipäätöksiä osittain niiden perusteella. Tällöin päätökset voivat johtaa samankaltaisiin lopputuloksiin kuin tilanteessa, jossa päätös olisi perustunut suojeltuun ryhmäjäsenyyteen. Välittäjien havaitseminen on hankalaa koneoppimisalgoritmien läpinäkyvyysongelmien takia, minkä lisäksi systeemi voi myös peittää syrjinnästä kertovia vihjeitä. Tämän lisäksi ihmiset eivät usein pysty identifioimaan itse välittäjiä. [21]

Syrjinnästä aiheutuu monenlaisia seurauksia syrjinnän kohteena oleville yksilöille tai ryhmille. Eriarvoinen kohtelu syrjinnän kohteena oleviin sosiaalisiin ryhmiin vaikuttaa näiden ryhmien jäsenten mahdollisuuksiin toimia yhteiskunnassa tasa-arvoisina jäseninä. Syrjintä rikkoo myös ihmisoikeuksia, sosiaalista ja yhteiskunnallista tasa-arvoa sekä dehumanisoi sen kohteena olevia ryhmiä ja niiden jäseniä. Lisäksi syrjintä rikkoo yksilöiden oikeuksia tulla kohdelluksi tasavertaisesti ilman perusteetonta haittaa tai vähättelyä sekä ylläpitää ja pahentaa eriarvoisuutta yhteiskunnassa, mikä heikentää kohderyhmien jäsenten mahdollisuuksia itsensä toteuttamiseen ja yhteiskunnalliseen osallistumiseen. Syrjintä esimerkiksi estää vähemmistöjä saamasta samoja mahdollisuuksia työpaikkoihin tai koulutukseen. Lisäksi syrjintä alentaa sen kohteena olevien ryhmien jäseniä moraalisesti alempiarvoisiksi ja asettaa heidät alisteiseen asemaan, jossa he ovat alttiina hyväksikäytölle tai syrjäytymiselle. [19]

Tässä tutkielman kannalta merkittävin syrjinnän muoto on epäsuora institutionaalinen ja rakenteellinen syrjintä. Syrjintä ei usein ole suoraa ja tarkoituksellista, vaan se voi olla seurausta yhteiskuntaan syvään juurtuneista tiedostamattomista oletuksista ja rakenteellisista tekijöistä. Tekoälyjärjestelmät, kuten koneoppimisalgoritmit, heijastavat usein kehittäjänsä ja käyttäjänsä oletuksia reiluudesta ja oikeudenmukaisuudesta. Jos näitä oletuksia ei kyseenalaisteta riittävästi, hyvää tarkoittavat henkilöt voivat tulla osaksi syrjivien prosessien uusintamista yhteiskunnassa. [23] Syrjintä on tiiviisti yhteydessä oikeudenmukaisuuteen: syrjintä vähentää oikeudenmukaisuutta sekä sosiaalisella että distributiivisella tasolla, kun taas oikeudenmukaisuus edellyttää syrjinnän välttämistä ja yhteisen moraalisen tasa-arvon kunnioittamista. [19] Oikeudenmukaisuuden kysymykset liittyvät yleisellä tasolla siihen, miten yksilöiden tulisi kohdella toisiaan, millaisia lakeja tulisi säätää ja mikä on paras yhteiskunnallisen järjestymisen tapa [24, s. 10]. Yksiselitteistä määritelmää myöskään oikeudenmukaisuudelle ei ole, vaan käsitettä on lähestytty eri näkökulmista, teorioista ja

filosofisista koulukunnista käsin vuosituhansien ajan [25, s. 11]. Voidaan kuitenkin yleisellä tasolla sanoa, että oikeudenmukaisuutta määriteltäessä määritellään yhteiskunnallisten instituutioiden tarkoitus sekä se, mitä ominaisuuksia instituutiot arvostavat ja palkitsevat sekä millaisia etuja niissä jaetaan [24, s. 234].

Oikeudenmukaisuutta lähestytään usein kolmesta näkökulmasta: distributiivisesta, proseduraalisesta sekä retributiivisesta oikeudenmukaisuusmääritelmästä käsin. *Distributiivisessa oikeudenmukaisuudessa* keskitytään resurssien oikeudenmukaiseen jakamiseen ihmisten kesken. [25, s. 13] Distributiivisessa oikeudenmukaisuudessa oikeudenmukaisuus on toteutunut silloin, kun resurssit on jaettu tasapuolisesti jonkin tietyn kriteerin, kuten tasa-arvon, ansion tai tarpeen, perusteella [26]. Distributiivisessa oikeudenmukaisuudessa pyritään resurssien reiluun jakoon, mutta aina ei ole yksimielisyyttä siitä, millä perustein resurssit tulisi jakaa ja minkälaisia hyveitä tulisi palkita [24, s. 216]. *Proseduraalisessa oikeudenmukaisuudessa* keskitytään taas menettelyyn, jolla hyödykkeiden jakoon päädytään [25, s. 13]. Puhtaasti proseduraalisessa oikeudenmukaisuudessa tuloksen oikeudenmukaisuus perustuu siihen, että noudatetaan reilua prosessia. Tällöin oikeudenmukaisuuden perustana on oikeutettujen menetelmien käyttö eikä lopputuloksella itsellään ole merkitystä. Proseduraalinen oikeudenmukaisuus korostaa siis menettelytapojen reiluutta: oikeudenmukaisuus toteutuu, kun jokaisen päätöksen taustalla on sama läpinäkyvä prosessi. Proseduraalisessa oikeudenmukaisuudessa riskinä on kuitenkin se, ettei oikeudenmukainen prosessi välttämättä johda oikeudenmukaisiin lopputuloksiin. [26] *Retributiivisella oikeudenmukaisuudella* tarkoitetaan taas rikoksiin ja rangaistuksiin liittyvää oikeudenmukaisuutta. [25, s. 14]

Koneoppimisen kontekstissa syrjinnän ja oikeudenmukaisuuden kysymyksiä tarkastellaan tyypillisesti viidestä näkökulmasta: tietoisuuteen perustuva oikeudenmukaisuus (engl. fairness through awareness), demografinen tasapaino (engl. demographic parity), tarkkuuden tasapaino (engl. accuracy parity), mahdollisuuksien tasa-arvo (engl. equality of opportunity) ja tasapainotetut todennäköisyydet (engl. equalized odds). Tietoisuuteen perustuvassa oikeudenmukaisuudessa pyritään vähentämään ennakkoluuloja poistamalla sensitiiviset muuttujat (kuten sukupuoli tai etninen tausta) aineistosta. Demografinen tasapaino puolestaan viittaa mallin tulosten tasapuoliseen jakautumiseen eri ryhmien kesken siten, ettei mallin suorituskyky vaihtele ryhmien välillä. Tarkkuuden tasapaino pyrkii siihen, että algoritmin kokonaisennustetarkkuus on yhtenevä eri ryhmissä. Mahdollisuuksien tasa-arvon periaatteessa mitataan, onko pätevällä henkilöllä yhtäläinen mahdollisuus tulla algoritmin

valitsemaksi riippumatta siitä, mihin ryhmään hän kuuluu. Tasapainotetut todennäköisyydet puolestaan ovat tästä tiukempi versio, sillä ne edellyttävät, että algoritmin ennusteet ovat oikeudenmukaisia sekä oikeiden että väärien positiivisten ja negatiivisten tulosten osalta eri ryhmien välillä. [27] Koneoppimisen reiluusmetriikoissa korostuvat siis yhtäältä distributiivinen, toisaalta lopputuloksen oikeudenmukaisuus.

Syrjintää vähentäviä algoritmeja markkinoidaan usein proseduraalisen oikeudenmukaisuuden näkökulmasta, mutta todellisuudessa algoritmien toiminnan oikeudenmukaisuuden perustana on distributiivinen tai lopputuloksen oikeudenmukaisuus. Lopputulokseen keskittyvä algoritmi ei ota huomioon algoritmin menettelytapojen oikeudenmukaisuutta. [28] Lopputuloksen oikeudenmukaisuuteen keskittyminen saattaakin johtaa syytöksiin siitä, että epäpätevämpi hakija pääsee rekrytointiprosessissa eteenpäin, jos algoritmi on päätynyt valitsemaan hänet vain oikeudenmukaisuusmittarin tuomien vaatimusten takia [28], [29]. Tällöin oikeudenmukaisuuskäsitys on ristiriidassa proseduraalisen oikeudenmukaisuuden näkökulman kanssa, jossa tärkeintä olisi oikeudenmukaiset menettelytavat prosessin aikana.

4 Tekoälyn etiikka

Tekoälyn kehityksen ja tekoälyä koskevien käsitysten yleistymisen myötä on huomattu tekoälyn tuovan mukanaan teknisten edistysaskelien lisäksi uudenlaisia eettisiä ongelmia. Toisaalta tekoäly myös itsessään syventää tai muuttaa luonteeltaan aiempia ongelmakenttiä. [12] Tekoälyn eettisiin ongelmiin lukeutuvat yhtäältä ajankohtaiset tekoälyn kehittämiseen liittyvät kysymykset, kuten miten tekoälyn kehitysprosessissa ja sen tuomien toiminnallisten mahdollisuuksien myötä varmistetaan ihmisten tieto- ja yksityisyydensuoja, mutta toisaalta myös laajemmat, vaikeammin rajattavat tekoälyn tuomat ongelmat yhteiskunnassa, kuten työttömyys prosessien automatisoitumisen seurauksena. Muita tekoälyyn liitettyjä ongelmakenttiä ovat käyttäytymisen manipulointi, tekoälyjärjestelmien läpinäkyvyys, ennakkoluulot päätöksentekojärjestelmissä sekä ihmisen ja robotin vuorovaikutus. [30] Myös esimerkiksi aiemmin mainittuun sosiaalisen median sisältöjen räätälöityyn kohdistamiseen liittyy monia eettisiä huolia: se mahdollistaa poliittisten vaikutusyritysten tehokkaamman kohdentamisen ja räätälöinnin kohdeyleisön mukaisesti, mikä saattaa osaltaan mahdollistaa myös laajasti keskusteltua yhteiskunnallisen kuplautumisen ilmiötä [31]. Tekoälyyn liittyvät huolet eivät siten rajaudu sellaisiin scifi-elokuvista tuttuihin filosofisiin kysymyksiin, kuten onko tehokas tekoäly moraalis-oikeudellinen persoona [12], vaan koskettavat monia nyky-yhteiskunnan perustavia elementtejä, kuten pääomien ja työn jakautumista sekä yhteiskunnallista vakautta [30].

Etiikalla tarkoitetaan yleisesti filosofian osa-aluetta, joka tutkii ihmisen toimintaa hyvän ja pahan tai moraalisesti oikean ja moraalisesti väärän näkökulmasta. Etiikassa on useita vaihtoehtoisia teorioita, joissa pohditaan kysymyksiä muun muassa hyvästä ja pahasta, oikeudenmukaisuudesta, arvoista ja siitä, miten ihmisten tulisi toimia. [32] Tekoälyn etiikan alalla tutkitaan tekoälyyn liittyviä eettisiä periaatteita, ohjeita ja säädöksiä [33]. Tekoälyn etiikassa, kuten muussakaan teknologian etiikassa, tekoäly ei pelkisty vain teknologiaan, vaan huomiota kiinnitetään siihen, mitä ihmiset tekevät tekoälyllä, miten he käyttävät sitä ja miten tekoäly integroidaan laajempaan sosio-tekniseen ympäristöön [11, s. 80]. Toisin sanoen tekoälyn etiikassa pohditaan, miten ja millaisin mahdollisin seurauksin tekoäly asettuu ihmisten elämään ja miten arvioida tähän liittyviä kysymyksiä ja ehdotettuja vastauksia. Tekoälyn etiikka liittyykin tiiviisti laajempiin keskusteluihin, kuten teknologian filosofiaan [34], [35]. On esitetty, että tekoälysovellusten toimintaa tulee arvioida samalla eettisellä mittaristolla kuin ihmisen toimintaa. Huomioon tulisi tällöin ottaa oikeudenmukaisuuden,

reiluuden ja syrjinnän vähentämisen, autonomian kunnioittamisen, vahingon tuottamisen kiellon, toisten auttamisen ja hyväntekeväisyyden vaatimukset [36]. Tekoälyn etiikkaan liittyy siis monia erilaisia aspekteja ja ongelmia. Tässä luvussa keskitytään tutkimuskysymysten kannalta keskeisiin osa-alueisiin selitettävyyteen, algoritmiseen syrjintään sekä vastuun kysymyksiin.

Tekoälyn selitettävyyden – tai pikemminkin selittämättömyyden – ongelma liittyy syväoppivien koneoppimisalgoritmien rakenteeseen. Syväoppiviin neuroverkkoihin perustuvat tekoälysovellukset ovat nykyään niin monimutkaisia, ettei niiden toimintaa pysty selittämään täysin edes neuroverkkoa koodannut ohjelmoija [36]. Asiantuntija tuntee algoritmin käytännön toimintalogiikan, muttei pysty selittämään, miten algoritmi tarkalleen ottaen päätyy antamaansa tulokseen [11, ss. 116–117]. Puhutaankin niin kutsutusta tekoälyn mustasta laatikosta, jonka avaaminen on tekoälytutkijoiden haasteena [11, s. 123], [36]. Selitettävän tekoälyn ala (XAI, Explainable AI) pyrkii vastaamaan mustan laatikon ongelmiin ja tekemään syväoppivista neuroverkoista selitettävämpiä alan asiantuntijoille sekä laajemmalle ihmisjoukolle [36].

Tekoälyn tekemien päätösten avoimuus ja selitettävyys onkin entistä tärkeämpää, kun tekoälyä käytetään päätöksenteon tukena ja varsinkin silloin, kun päätöksentekoon epäillään liittyvän vinoumia tai syrjintää [36]. *Koneoppimisalgoritmien tekemällä syrjinnällä* tarkoitetaan tiettyä ihmisryhmää tai yksilöitä epätasa-arvoisesti käsittelevää koneoppimismallia. Algoritmi voi syrjiä esimerkiksi iän, sukupuolen, asuinpaikan tai etnisyyden perusteella ja käsitellä kahta muiden ominaisuuksien perusteella saman profiilin edustajaa eri tavoin. [11, ss. 125–129] Syrjintä ilmenee tyypillisesti silloin, kun päätökseen on vaikuttanut ominaisuus, jolla ei todellisuudessa ole merkitystä käsiteltävänä olevan asian kannalta. Syrjintä ei usein ole tietoista, vaan se on opittu kognitiivinen ominaisuus, josta päätöksentekijä ei aina itsekään ole tietoinen. [30]

Syitä koneoppimismallien tekemälle syrjinnälle on useita. Syrjintä voi johtua esimerkiksi siitä, että mallin kouluttamiseen käytetty aineisto on valmiiksi syrjivä tai muuten ongelmallinen. Aineisto voi esimerkiksi olla kooltaan liian pieni tai se voi sisältää vinoutuneen otoksen ihmisistä, kuten vain länsimaisten maiden ihmisiä. Aineisto voi myös olla syrjivä siksi, että syrjintä on juurtunut osaksi laajempia yhteiskunnallisia rakenteita, joita koneoppimisalgoritmi sitten vahvistaa. [11, ss. 125–126] Tekoälyjärjestelmät koulutetaan aina historiallisella aineistolla, jonka taustalla olevia oletuksia ja malleja voi olla hankala tai jopa

mahdotonta poistaa. Käytetyn aineiston ajallisten ja paikallisten erityispiirteiden voidaan ajatella jäljentyvän tekoälyjärjestelmien toimintaan ja toisintuvan niiden käytön myötä. [37] Lisäksi syrjintä voi johtua yksinkertaisesti siitä, että ihmiset luottavat tekoälyn tuottamiin tuloksiin liikaa. Syrjintä on usein tiedostamatonta eivätkä neuroverkkojen ohjelmoijat tai tekoälyn käyttäjät osaa yleensä ennakoida syrjivän koneoppimismallin käyttämisestä aiheutuvia vaikutuksia. Syrjintää voi tapahtua myös silloin, kun tekoälyn rakentajat tai sitä käyttävän sovelluksen suunnittelijat eivät ota tätä ongelmakenttää huomioon jo algoritmin suunnitteluvaiheessa. [11, ss. 125–127]

Rekrytoinnin kontekstissa yksi esimerkki syrjivästä koneoppimisalgoritmista on Amazonin vuosina 2014–2015 rekrytointiinsa kehittämä tekoälysovellus, jonka huomattiin syrjivän naishakijoita. Tekoälysovelluksen tarkoitus oli seuloa hakijoiden joukosta ansioluetteloiden perusteella parhaat hakijat ja palkata heidät. Sovellus koulutettiin Amazonin aikaisempiin rekrytointeihin käytetyillä ansioluetteloilla sekä niihin liittyvillä henkilöstövalinnoilla. Koulutuksen jälkeen tekoälysovellus alkoi suosia miespuolisia hakijoita tekemissään valinnoissa, vaikka hakemuksissa ei viitattu suoraan sukupuoleen. Selvisi, että sovellus etsi ansioluetteloista aikaisemmin menestyneiden mieshakijoiden perusteella tiettyjä kielellisiä yhdistelmiä, minkä seurauksena esimerkiksi ”naisten shakkikerhon kapteeni” hylättiin, mutta ”shakkikerhon kapteeni” ei. Sovellusta yritettiin kehittää vähemmän syrjiväksi, mutta Amazon päätyi luopumaan sovelluksen käytöstä kokonaan vuonna 2015. [37, s. 10]

Amazonin rekrytointitapaus on hyvä esimerkki siitä, että syrjivää koneoppimisalgoritmia on hankalaa korjata neutraalimmaksi. Koneoppimisalgoritmeja kehitettäessä luodaan yhden osan todellisuudesta sisältävä abstraktio, johon algoritmin kehittäjä on valinnut tiettyjä osia todellisuudesta ja jättänyt tiettyjä osia pois. Valintaa liittyy myös siihen, minkälaisella aineistolla algoritmit koulutetaan. [11, s. 91] Koneoppimisalgoritmit ovat alttiita samoille stereotypioille ja syrjiville rakenteille kuin ihmisetkin, sillä algoritmit koulutetaan oikeasta maailmasta löytyvällä aineistolla [2]. Oikeasta maailmasta löytyvä aineisto on usein valmiiksi vinoutunutta, eli esimerkiksi sukupuoleen, ikään ja etnisyyteen liittyvää syrjintää tapahtuu usein ilmankin tekoälyä [37]. Koneoppimisalgoritmit eivät ole koskaan neutraaleja kuvauksia todellisuudesta ja niiden suunnitteluun sekä niiden tuottamiin tuloksiin tuleekin suhtautua kriittisesti. Ihmisen harkinnalla on edelleen suuri rooli ja jos teknologiaan luotetaan liikaa, tuloksena voi olla syrjivä lopputulos. [11, ss. 125–127]

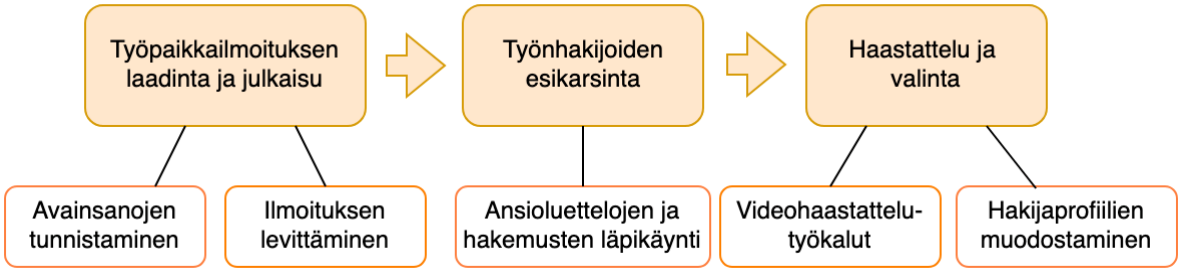
Koneoppimismallien tekemää syrjintää on pyritty vähentämään monin eri tavoin, mutta se on osoittautunut haastavaksi. On esimerkiksi huomattu, että muokkauksen seurauksena koneoppimismalli saattaa tehdä huonompia ennusteita tai sen tehokkuus voi kärsiä. [38], [39] On jopa esitetty, että tekoälyalgoritmit tulisivat aina olemaan vinoutuneita johonkin suuntaan: esimerkiksi rekrytoinnin kontekstissa algoritmi, joka käy läpi ansioluetteloita, etsii aineistosta tiettyjä piirteitä. Algoritmin voi näin sanoa olevan syrjivä niiden eduksi, joilta löytyy nämä etsityt piirteet. [11, ss. 131–132] Toisaalta on myös esitetty varovaisen positiivisia avauksia tekoälyn mahdollisesta käytöstä nimenomaan epäoikeudenmukaisuuden ja syrjinnän tunnistamisen työkaluna. Tekoälyn tehokkuutta kaavojen ja toistuvien rakenteiden tunnistamisessa voitaisiin siis parhaassa tapauksessa hyödyntää myös syrjintäongelman ratkaisemiseen. [36], [37] Rekrytoinnissa tekoälyn mahdollisuuksia syrjinnän torjunnassa on myös pyritty kehittämään tekoälypohjaisten rekrytointityökalujen avulla. Rekrytointiprosessia on pyritty kehittämään reilummaksi esimerkiksi käyttämällä rekrytointityökaluja apuna inhimillisten ennakkoasenteiden ja syrjinnän lieventämiseen. [40]

Kolmas tekoälyn etiikasta puhuttaessa ajankohtainen kysymys on se, kenellä on *moraalinen vastuu tekoälyn tekemistä päätöksistä*. Moraalisella vastuulla tarkoitetaan sitä, että henkilö tai toimija on vastuussa päätöksistään moraalisesta näkökulmasta. Kyky tällaiseen harkintaan on moraalisilla toimijoilla, eli moraalisia seikkoja päätöksenteossaan huomioivia yksilöillä tai entiteeteillä. [41] Näin ollen moraalinen toimijuus perustuu kykyyn tehdä tietoisia, vapaaehtoisia päätöksiä ja kantaa vastuu niiden seurauksista. Tämä sisältää myös selitysvelvollisuuden: moraaliselta toimijalta odotetaan, että hän pystyy perustelemaan tekonsa ja päätöksensä. Tekoälyn kontekstissa moraalinen toimijuus on haastava käsite, sillä tekoäly voi tehdä ihmisiin vaikuttavia päätöksiä kykenemättä moraaliseen pohdintaan. Tekoälyn toimijuutta voisikin kuvailla ennemmin tekniseksi kuin moraaliseksi. Tämä tarkoittaa, että vastuu tekoälyn toiminnasta jää lopulta ihmisille, jotka suunnittelevat ja käyttävät tekoälypohjaisia työkaluja. Vastuun ottaminen vaatii joidenkin näkökulmien mukaan tiedon siitä, mitä on tapahtunut, mutta tämän selvittäminen on vaikeaa tai mahdotonta, kun vastassa on tekoälyn musta laatikko. Ongelma ei palaudu luottamuksen kysymykseen eli siihen, miten ihmiset saataisiin luottamaan tekoälypohjaisiin työkaluihin paremmin. Selityskyky on paitsi moraalinen velvoite myös välttämätön edellytys vastuulliselle päätöksenteolle. [11, ss. 109–123]

5 Rekrytoinnin koneoppimisalgoritmit ja syrjintä

Koneoppimisalgoritmeja on käytetty rekrytoinnissa jo 1900-luvun lopusta lähtien. Aluksi käyttö perustui ansioluetteloiden seulontaan ja hakemusten lajittelun automatisointiin, kuten tiettyjen avainsanojen etsimiseen ansioluetteloista. Tekoälypohjaiset sovellukset kehittyivät tämän jälkeen nopeasti ja jo 2010-luvun lopussa rekrytoinnissa käytettiin kasvojen ilmeitä ja kehonkieltä analysoivia videohaastatteluohjelmistoja. Rekrytoinnissa käytettävät tekoälymenetelmät perustuvat pitkälle syväoppimisen ja luonnollisen kielen käsittelyn tekniikoihin, joita käyttäen rekrytointiprosesseja on saatu automatisoitua vähentäen HR-ammattilaisten manuaalisen työn määrää. Nykyään tekoälymenetelmiä käytetään rekrytoinnissa laajasti muun muassa ansioluetteloiden ja hakemusten läpikäynnissä, hakijoiden persoonallisuuden arvioinnissa sekä kielimuurien ylittämisessä. [2]

Tekoälyä voi käyttää rekrytoinnin joka vaiheessa: työpaikkailmoituksen laadinnassa ja julkaisussa, työpaikkojen etsimisen ja hakemisen apuna, työntekijöiden valinnassa ja arvioinnissa sekä työnhakijoiden avustamisessa työnhaussa [42]. Tekoälyn käyttöä rekrytoinnissa on havainnollistettu kuvassa 3.



Kuva 3. Tekoälyn käyttö rekrytoinnin eri vaiheissa.

Työpaikkailmoitusten laadinnan vaiheessa tekoälyn avulla luodaan tiettyjä avainsanoja sisältävä ilmoitus, jolla pyritään kiinnittämään sopivien hakijoiden huomio. Julkaistua ilmoitusta levitetään tekoälyn avulla sopivissa medioissa mainostamaan ilmoitusta työpaikan profiiliin sopiville käyttäjille. [4] Työntekijöiden karsinta- ja valintavaiheissa tekoälysovelluksia on useita. Yksi yleinen käyttökohde on hakijoiden ansioluettelojen ja hakemusten läpikäynti tekoälyn avulla. Ansioluetteloista ja hakemuksista voidaan etsiä tiettyjä avainsanoja, aikaisempaa työkokemusta tai tietynlaista osaamista luonnollisen kielen käsittelyyn pohjautuvan algoritmin avulla. Tämän läpikäynnin seurauksena algoritmi suosittelee työnantajaa valitsemaan tietyt työnhakijat jatkoon ja karsimaan osan pois.

Haastatteluvaiheessa taas käytetään tekoälypohjaisia, hakijoiden ilmeitä ja kehonkieltä analysoivia videohaastattelutyökaluja. Valintavaiheessa hakijoiden soveltuvuutta arvioidaan tekoälyavusteisesti muun muassa muodostamalla hakijoista profiileja heidän sosiaalisen median käyttäjätiliensä ja muun internetaktiivisuutensa perusteella. [42] Näiden käyttötapojen lisäksi tekoälypohjaisia chatbotteja käytetään läpi rekrytointiprosessin hakijoiden neuvomisessa sekä luomaan työnantajasta helposti lähestyttävä kuva [4].

Tekoälypohjaisten rekrytointisovellusten käyttöönotosta on huomattu olevan paljon hyviä seurauksia rekrytoinnissa. Koneoppimisalgoritmien käytön on esimerkiksi huomattu nopeuttavan rekrytointiprosessia ja vapauttavan HR-kentällä toimivien ihmisten työpanosta pois rutiinitehtävistä [2]. Koneoppimisalgoritmien käytön on lisäksi väitetty lisäävän objektiivisuutta ja oikeudenmukaisuutta rekrytoinnissa, sillä tekoälyn ei nähdä tekevän samanlaista rakenteellista syrjintää kuin ihminen tiedostamattaan tekee [28]. Ihmisen tekemällä syrjinnällä tarkoitetaan muun muassa päätöksiä, joihin ovat vaikuttaneet esimerkiksi tiettyjä ryhmiä syrjivät ihmisen (tiedostamattomat) asenteet ja stereotypiat. Näiden minimoimiseksi on kehitetty esimerkiksi HR-ammattilaisten koulutuksia ja arviointikriteereitä, käytetty anonyymia rekrytointia ja tehty rekrytointitiimeistä monimuotoisempia. [2] Näistä toimista huolimatta ihminen voi tehdä tiedostamattaan syrjiviä päätöksiä rekrytoinnissa ja tähän ongelmaan on esitetty ratkaisuksi tekoälyä. On kuitenkin osoitettu, että koneoppimismallit toisintavat yhteiskunnassa valmiiksi olemassa olevia syrjiviä rakenteita. [11, ss. 129–130], [43]

Koska koneoppimismalleihin perustuvien ratkaisujen käyttö on yleistynyt rekrytointiprosesseissa, mallien kehittämiseen kiinnitetään yhä enemmän huomiota. Tästä huolimatta tekoälypohjaiset algoritmit sisältävät yhä syrjintään johtavia heikkouksia sekä algoritmisia vinoumia. [2] Algoritmisella vinoumalla tarkoitetaan koneoppimisalgoritmien tekemiä järjestelmällisiä, syrjintään johtavia virheitä. Virheet voivat kohdistua lailla suojattuihin ominaisuuksiin, kuten etnisyyteen tai sukupuoleen. Kun koneoppimisalgoritmien tekemät arvioinnit toistuvasti ali- tai yliarvioivat tietyn ryhmän tuloksia, voidaan puhua vinoutuneesta algoritmista. [1] Algoritmien toimintaan liittyvät ongelmat ilmenevät algoritmin koulutusaineistossa, algoritmin toimintamekanismeissa ja päätöksentekovaiheessa [43]. Eri vaiheissa esiintyviä algoritmin toimintaan liittyviä ongelmia on koottu taulukkoon 1.

Taulukko 1. Koneoppimisalgoritmien käytössä ilmenevät ongelmat.

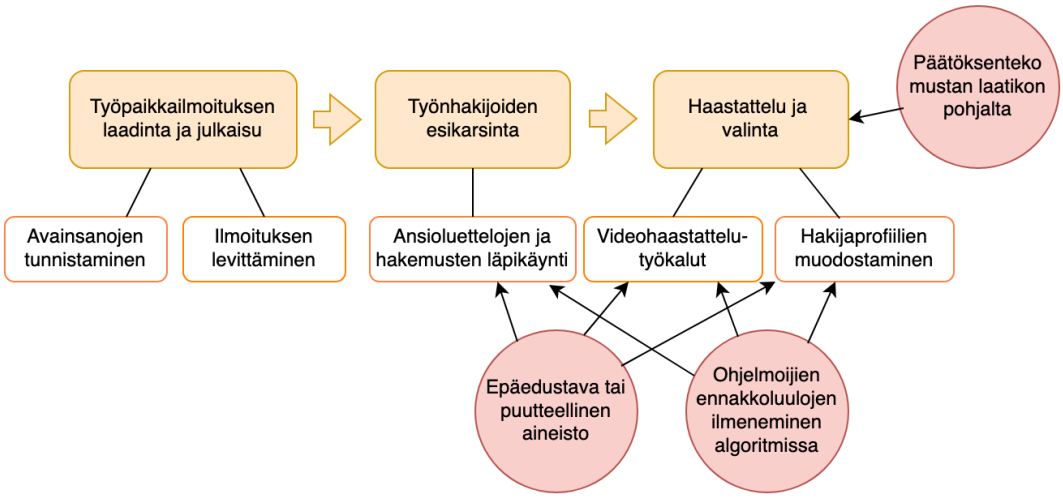
Tekoälyavusteisen rekrytoinnin vaihe	**Algoritmin toimintaan liittyvä haaste**	**Algoritmin käytöstä seuraavat ongelmat**
Koneoppimisalgoritmin koulutus	Epäedustava tai yksipuolinen aineisto lailla suojattujen ominaisuuksien suhteen	Algoritmi tuottaa syrjiviä tuloksia
	Puutteellinen aineisto työpaikkaan vaadittujen ominaisuuksien suhteen	Algoritmin tekemät päätökset eivät perustu henkilön osaamiseen ja kokemukseen
Koneoppimisalgoritmin toiminta	Ohjelmoijien ennakkoluulojen ilmeneminen algoritmissa	Algoritmi tuottaa syrjiviä tuloksia
Päätöksenteko	Päätöksen tekeminen mustan laatikon pohjalta	Päätöksentekijä ei tiedä, millä perusteella algoritmi on tuottanut tuloksensa Vastuun kysymykset

Koulutusaineistoon liittyvät ongelmat johtuvat liian pienistä tai epäedustavista aineistoista. Koneoppimisalgoritmin kouluttamiseen käytettävä aineisto toimii koneoppimisalgoritmin perustana. Jos algoritmin koulutukseen käytettävä aineisto on liian pieni, puutteellinen tai epäedustava, johtaa se usein algoritmin puolueellisuuteen. [1] Esimerkki epäedustavan aineiston käyttämisestä rekrytointisovelluksen yhteydessä on aikaisemmassa luvussa esitelty Amazonin tapaus: vaikka algoritmin koulutusaineisto oli suuri ja sisälsi dataa kymmenen vuoden ajalta, sisälsi algoritmin koulutusaineisto enemmän esimerkkejä työpaikalle palkatuista miehistä ja heidän profiileistaan kuin naisista, minkä seurauksena algoritmi päätteli miesten soveltuvan paremmin työhön [44]. Edustusvinoumaan perustuvat päätökset johtavat helposti yli- tai aliedustettuihin ryhmiin kohdistuvaan syrjintään [45]. Epäedustavan koulutusaineiston lisäksi koneoppimisalgoritmi voi johtaa syrjimään tiettyjä hakijoita, jos algoritmin koulutukseen käytettävästä aineistosta puuttuu työtehtävän kannalta olennaisia muuttujia. Tällöin algoritmin tekemät päätökset perustuvat helposti muihin tekijöihin kuin henkilön osaamiseen ja kokemukseen. [2]

Koneoppimisalgoritmin toimintaan liittyvät ongelmat johtuvat itse algoritmin toiminnasta. Algoritmin suunnitteleva henkilö valitsee ja koodaa rekrytointiin sopivan mallin ominaisuudet ja jos nämä asetetaan puolueellisesti, tuottaa algoritmi vinoutuneita tuloksia. [1] Algoritmit voivat siten ilmentää ja toisintaa algoritmia koodanneen henkilön ennakkoluuloja [46]. *Päätöksentekovaiheeseen liittyvät ongelmat* ilmenevät ihmisen tekemien,

koneoppimisalgoritmin tulokseen perustuvien päätösten tekemisen yhteydessä. Koneoppimisalgoritmien käytöstä aiheutuva syrjintä ei johdu pelkästään siitä, kuinka hyvin koulutukseen käytetty aineisto edustaa kohdeväestöä tai kuinka hyvin algoritmit on suunniteltu, vaan se on seurausta myös ihmisten tekemistä päätöksistä [47]. Rekrytoinnin parissa toimivien henkilöiden ymmärrys algoritmien tekemästä syrjinnästä onkin havaittu puutteelliseksi [2]. On esimerkiksi huomattu, että jotkut HR-ammattilaiset odottavat tekoälypohjaisilta työkaluilta nopeita ja yksinkertaisia ratkaisuja monimutkaisiin ongelmiin [40]. Toisaalta tekoälysovellukset voidaan kokea hankaliksi ymmärtää ja tämän takia rekrytointipäätösten pohjaaminen tekoälypohjaisen työkalun ehdotuksiin saattaa herättää rekrytoijissa epäilyksiä [48]. Myös yhteiskunnalliset muutokset tai poikkeavat kulttuuriset arvot voivat aiheuttaa ongelmia päätöksentekovaiheessa: jos algoritmi on suunniteltu tietyin, loppukäyttäjien arvoista poikkeavin oletuksin, voi se aiheuttaa ennakkoluuloa tekoälysovelluksen käyttämistä kohtaan [45].

Tekoälyn käytöstä seuraavien ongelmien sijoittumista rekrytointiprosessin eri vaiheisiin on havainnollistettu kuvassa 4. Kuvasta huomataan, että tutkijoiden esiin nostamat ongelmat korostuvat erityisesti työnhakijoiden karsinta- ja valintavaiheissa. Epäedustava tai puutteellinen koulutusaineisto voi vaikuttaa siihen, minkälaisia ansioluetteloja algoritmi suosii, minkälaisia eleitä se etsii videohaastatteluista ja minkälaisia profiileja se tekee hakijoista. Ohjelmoijien ennakkoluulojen heijastuminen algoritmin toimintaan saattaa samaan tapaan vaikuttaa haitallisesti erityisesti työnhakijoiden menestymiseen esikarsinta- ja haastatteluvaiheissa. Algoritmin selitettävyyteen liittyvät ongelmat konkretisoituvat erityisesti valintavaiheessa, jossa päätös rekrytoitavasta henkilöstä tehdään.



Kuva 4. Tekoälyn käyttö ja sen ongelmat rekrytoinnin eri vaiheissa.

6 Koneoppimisalgoritmien tekemän syrjinnän vähentämisen tekniikat rekrytoinnissa

Koneoppimisalgoritmien tekemän syrjinnän vähentämiseen on kolme keskeistä tapaa: esikäsittely, prosessoinninaikainen käsittely sekä jälkikäsittely [38], [39]. Kirjallisuuskatsauksen artikkeleissa käytetyt syrjinnän vähentämisen tekniikat on esitelty taulukossa 2.

Taulukko 2. Syrjinnän vähentämiseen käytetyt tekniikat.

Käsittelyvaihe	**Ongelma, joka halutaan ratkaista**	**Syrjinnän vähentämisen tekniikka**	**Tekniikkaa hyödyntäneet tutkimukset**
Esikäsittely	Epäedustava tai yksipuolinen aineisto lailla suojattujen ominaisuuksien suhteen	Piirteiden peittäminen tai poistaminen aineistosta	[44], [47]
		Aineiston optimointi monimuotoisemmaksi	[49], [50]
Prosessoinninaikainen käsittely	Syrjiviä tuloksia tuottavan algoritmin tapa tuottaa ennusteita	Algoritmin toimintalogiikan muokkaaminen	[29], [51], [52]
Jälkikäsittely	Musta laatikko	Selitettävyyden lisääminen	[53], [54], [55]
	Algoritmin tuottamat syrjivät ennusteet	Tulosten tasapainottaminen algoritmin toiminnan jälkeen	[56]

6.1 Esikäsittely

Esikäsittelymenetelmät keskittyvät aineiston muokkaamiseen ennen sen käyttämistä koneoppimisalgoritmin koulutukseen [39]. Esikäsittelymenetelmillä on pyritty vastaamaan koulutusaineiston epäedustavuuden tuomiin haasteisiin ja oletuksena on, että tiettyjen muuttujien jakaumat ovat epätasapainossa algoritmin koulutusaineistossa [38]. Esikäsittelyssä aineiston ominaisuuksia ja luokkia muokataan ennalta määriteltyjen oikeudenmukaisuuskriteerien mukaisesti tavoitteena laajennettu ja monipuolisempi koulutusaineisto. Esikäsittelytekniikoiden syrjinnän vähentäminen perustuu neutraalimpaan sanojen ja ominaisuuksien esitystapaan [2] ja niitä käytetään erityisesti silloin, kun algoritmin toimintaa ei ole mahdollista muokata [39].

Kirjallisuuskatsauksessa esikäsittelymenetelmistä selvästi useimmin käytetty oli luonnollisen kielen käsittelyn tf-idf-algoritmi. Tf-idf-algoritmin toiminta perustuu sanojen esiintymistiheyteen käsiteltävässä dokumentissa ja aineistossa yleisesti. Algoritmi laskee ensin yksittäisten sanojen esiintymiskertoja dokumentissa, minkä jälkeen se vertaa näiden sanojen esiintymistiheyttä koko aineiston tasolla. Tf-idf-arvoja käytetään lopuksi asiakirjojen vertailuun. [49] Rekrytoinnin kontekstissa tämä voi tarkoittaa esimerkiksi ansioluettelon ja hakemuksen vertaamista haetun työpaikan vaatimuksiin.

Esikäsittelyä tehtiin kirjallisuuskatsauksen artikkeleissa sekä poistamalla ja piilottamalla piirteitä aineistosta että optimoimalla aineistoa monimuotoisemmaksi. Ensimmäistä tapaa käytettiin Gatandeepin ym. [44] ja Boothin ym. [47] tutkimuksissa. Gagandeep ym. [44] hyödynsivät algoritmia sukupuoleen liittyvien sanojen ja termien piilottamiseen ansioluetteloita sisältävästä aineistosta. Tutkijat eivät poistaneet sukupuolittuneita termejä aineistosta kokonaan, vaan sen sijaan piilottivat niitä, jotta algoritmi pystyisi huomioimaan termien ympäröivää kontekstia paremmin ja vähentämään siten piiloyhteyksiä aineistossa (esimerkiksi ”he is a software engineer” -> ”!&# is a software engineer”). Tutkimuksessa hyödynnettiin tf-idf-algoritmia tarkastelemaan sitä, miten hyvin sukupuolittuneiden termien piilottaminen aineistosta poisti sukupuoleen perustuvaa syrjintää. [44] Booth ym. [47] keskittyivät taas syrjinnän vähentämiseen koneoppimiseen perustuvissa videohaastattelutyökaluissa, joissa koneoppimisalgoritmit arvioivat videoidussa haastattelussa esiintyvän henkilön rekrytoitavuutta henkilön vastausten ja persoonallisuudesta muodostettujen päätelmien perusteella. Sukupuoleen liittyviä piirteitä normalisoitiin ensin tilastotieteellisillä menetelmillä, minkä lisäksi koulutusaineistosta poistettiin ne sukupuoleen liittyvät piirteet, joista algoritmi todennäköisimmin tunnisti sukupuolen. [47]

Aineiston optimointiin keskityttiin taas Deshpanden ym. [49] sekä McCareyn ja McTavishin [50] tutkimuksissa. Deshpande ym. [49] keskittyivät ansioluetteloiden kirjoitustyyliin ja syrjintään johtaviin sosio-lingvistisiin tekijöihin. Esimerkiksi ansioluetteloiden kirjoitustyyli ja sanavalinnat voivat liittyä hakijan taustaan, jotka algoritmi voi oppia yhdistämään tiettyihin ryhmiin. Tutkijat painottivat aikaisemmin lasketut tf-idf-arvot uudelleen lisäämällä niihin ylimääräisen oikeudenmukaisuusattribuutin tarkoituksenaan tehdä piirteistä vähemmän demografisesti syrjiviä. [49] McCarey ja McTavish [50] taas muokkasivat työhakemuksista muodostettuja tf-idf-arvoja geneettisillä algoritmeilla. Geneettiset algoritmit ovat evoluution periaatteita jäljitteleviä optimointimenetelmiä, jotka vertailevat ratkaisujen soveltuvuutta,

valitsevat parhaat ratkaisut jatkoon ja tekevät niistä risteytyksiä ja mutaatioita tuottaakseen parempia tuloksia. [50]

Esikäsittelymenetelmien on huomattu tasa-arvoistavan koneoppimisalgoritmien tekemiä ennusteita, mutta vähentävän myös algoritmin suorituskykyä ja tarkkuutta. Esimerkiksi Booth ym. havaitsivat tutkimuksessaan, että sukupuolitietoja sisältävien piirteiden poistaminen koneoppimismallista voi parantaa mallien oikeudenmukaisuutta, mutta samalla vähentää mallin tehokkuutta ja tarkkuutta [47]. Koneoppimismallin tarkkuudella viitataan siihen, miten hyvin malli pystyy vastaamaan käsillä olevaan tehtävään: rekrytoinnin kontekstissa tämä tarkoittaa parhaiten työpaikkailmoitukseen vastaavaa hakijaa. Esikäsittelymenetelmien haasteena on myös se, että tekniikat voivat muuttaa sanojen merkitystä, mikä vaikuttaa algoritmin tekemien ennustusten oikeellisuuteen [2]. Lisäksi on esitetty, ettei esikäsittelymenetelmillä pystytä käsittelemään tilanteita, joissa yksilöiden kokemaan syrjintään vaikuttavat useat eri tekijät, sillä menetelmissä keskitytään vain yksittäisiin ominaisuuksiin [2], [43]. Esimerkiksi Tilmes [57] tarkasteli vammaisuuden huomioimista koneoppimisalgoritmien suunnittelussa ja toiminnassa ja havaitsi, että vammaisuus koodataan koneoppimismalleissa usein yhdeksi muuttujaksi, jolloin sosiaalinen ja ympäristöllinen konteksti jäävät huomiotta. Tutkimuksen mukaan vammaisuuden kaltaista attribuuttia ei pysty mallintamaan staattisena ja binäärisenä muuttujana ilman, että mallintaminen vahvistaisi ableismia. [57]

6.2 Prosessoinninaikainen käsittely

Prosessoinninaikaisella käsittelyllä pyritään muuttamaan algoritmin toimintalogiikkaa sisällyttämällä algoritmiin syrjinnän vähentämiseen tarkoitettuja oikeudenmukaisuusmittareita [29]. Oikeudenmukaisuusmittareilla pyritään takaaman se, etteivät algoritmit etsi vain koulutusaineiston kautta oppimallaan tavalla tehtävän kannalta parhaita tuloksia, vaan huomioon otetaan myös oikeudenmukaisuusmittarin tuomat vaatimukset [38]. Oikeudenmukaisuusmittarit voivat liittyä esimerkiksi sukupuolten väliseen tasapuolisuuteen tai yksittäisten ryhmien suosimisen välttämiseen [39].

Kirjallisuuskatsauksen prosessoinninaikaista käsittelyä hyödyntävät tutkimukset käyttivät syrjinnän tunnistamiseen ja vähentämiseen *adversiaaliseksi oppimiseksi* kutsuttua menetelmää. Adversiaalinen oppiminen on syväoppimisen menetelmä, jossa kaksi neuroverkkoa kilpailee keskenään. Ensimmäinen neuroverkko yrittää tehdä mahdollisimman tarkkoja ennusteita, kun taas toinen neuroverkko pyrkii erottamaan ensimmäisen neuroverkon

luomat tiedot oikeista tiedoista. Ensimmäinen neuroverkko pyrkii parantamaan tulostensa laatua, kun taas toinen neuroverkko pyrkii parantamaan kykyään tunnistaa väärät tiedot. Jatkuva kilpailu parantaa molempien mallien suorituskykyä. [58]

Adversiaalisella oppimisella pyrittiin tutkimuksissa vähentämään vektoriavaruudesta sensitiivisiä, syrjinnän mahdollistavia attribuutteja. Peña ym. [51] käyttivät syrjinnän vähentämiseen adversiaaliseen oppimiseen perustuvaa SensitiveNets-algoritmia, joka käyttää aineistosta muodostettuja sanaupotevektoreita ja pyrkii poistamaan syrjintään johtavia sensitiivisiä tietoja vektoreista algoritmin toiminnan aikana. [58] Kim ym. [29] pyrkivät taas vähentämään videohaastattelujen arviointijärjestelmien syrjintää hyödyntämällä adversiaalista oppimista ja tilastollisia menetelmiä (Wasserstein-etäisyys, ks. esim. [59]). Tutkimuksessa käytettiin video-, ääni- ja tekstiaineistoja ennustamaan hakijoiden haastattelupisteitä ja adversiaalisen oppimisen malli koulutettiin tunnistamaan arkaluonteiset tiedot, joiden vaikutus pyrittiin poistamaan mallin ennusteista. [29] Putra ym. [52] hyödynsivät adversiaalista oppimista vähentämään syrjintää videohaastattelujen arviointijärjestelmässä hyödyntäen MAG-BERT-ARL-nimistä mallia. Malli yhdisti tekstiä, ääntä ja kuvaa käsitteleviä algoritmeja ja sen adversiaalinen komponentti tunnisti kohdat, joissa ryhmiä kohdeltiin epätasa-arvoisesti. Näitä kohtia painotettiin uudelleen, jotta malli oppisi paremmin käsittelemään samankaltaisia attribuutteja oikeudenmukaisemmin. [52]

Vaikka prosessoinninaikaisten menetelmien avulla on saatu tutkijoiden mukaan poistettua algoritmeista syrjintään johtavia komponentteja, on menetelmissä havaittu myös haasteita. Prosessoinninaikaiset käsittelytekniikat saattavat heikentää algoritmin tarkkuutta, sillä algoritmi voi alkaa syrjiä oikeudenmukaisuusmittarien ulkopuolella olevia ryhmiä eikä algoritmi siten aina tuota tehtävän kannalta tarkimpia tuloksia [29]. Lisäksi prosessoinninaikaisen käsittelyn optimointimenetelmissä käytetyt oikeudenmukaisuusmittarit ovat saaneet kritiikkiä. Algoritmit optimoidaan suorituskyvyn mukaan suhteessa johonkin tavoitteeseen (esimerkiksi ennustetarkkuus, herkkyys tai spesifisyys), jonka perusteella algoritmit tuottavat ennusteita. On huomautettu, etteivät algoritmit siten ole objektiivisia, vaan algoritmin kehittäjän valitsema oikeudenmukaisuusmittari vaikuttaa siihen, millaisia arvoja algoritmi toiminnassaan toteuttaa. [60]

6.3 Jälkikäsittely

Jälkikäsittelymenetelmät muokkaavat algoritmin tuottamaa lopputulosta oikeudenmukaisuuskriteerien mukaiseksi tai tekevät algoritmin toiminnasta läpinäkyvämpää

[38]. Jälkikäsittelyä voidaan tehdä joko tasapainottamalla eri ryhmien saamia ennusteita tai lisäämällä algoritmin selitettävyyttä [29]. Kirjallisuuskatsaukseen valikoituneissa artikkeleissa selitettävyyden lisääminen oli yleisempi tapa.

Sogancioglu ym. [53] keskittyivät selitettävyyden lisäämiseen hyödyntämällä SHAP-nimistä algoritmia, joka arvioi piirteiden vaikutusta koneoppimisalgoritmin ennusteisiin. SHAP antaa jokaiselle aineiston piirteelle sen suhteellista vaikutusta mallin tulokseen kuvaavan tärkeysarvon. Tutkimuksessa pyrittiin tunnistamaan ennusteisiin vaikuttaneet herkkien attribuuttien, kuten sukupuolen, iän tai etnisyyden, välillisenä edustajana toimivat piirteet. Ortega ym. [54] tutkivat taas, kuinka tehokkaasti induktiiviseen logiikkaohjelmointiin (ILP) kuuluva LFIT-niminen tekniikka pystyy selittämään koneoppimisalgoritmien toimintaa. ILP yleistää sille annettuja esimerkkejä ja muodostaa niistä loogisia sääntöjä algoritmin toiminnan selittämiseen. LFIT on taas ILP:n tekniikka, joka pyrkii selittämään algoritmien toimintaa tarkkailemalla järjestelmän tilasiirtymiä ja oppii siten loogiset säännöt, joilla algoritmi päätyy tarjoamiinsa tuloksiin. Ortegan ym. tutkimuksessa LFIT:ia testattiin neuroverkolla, joka pisteytti ansioluetteloja eri attribuuttien, kuten sukupuoleen liittyvien arvojen perusteella. Näin LFITillä pyrittiin lisäämään mallin läpinäkyvyyttä tarjoamalla selityksiä algoritmin päätöksistä. Russo ym. [55] pyrkivät lisäämään algoritmin läpinäkyvyyttä käyttämällä LIME-nimistä algoritmia koneoppimismallien yksittäisten ennusteiden selittämiseen. LIME tarkastelee mallin käyttäytymistä paikallisesti tietyn ennusteen ympärillä ja luo muunnoksia alkuperäisestä datapisteestä selittäen, mitkä attribuutit vaikuttivat ennusteeseen. [61] Russon ym. tutkimuksessa menetelmän tuottamat tulokset esitettiin tietämysgraafina, jolla pyrittiin havainnollistamaan, miten ja miksi koneoppimisalgoritmi päätyi tiettyyn lopputulokseen [55].

Delecrazin ym. [56] käyttämä jälkikäsittelytekniikka pyrki syrjinnän vähentämiseen selitettävyyden lisäämisen sijaan koneoppimismallin tuloksia tasapainottamalla. Tutkijat käyttivät tähän Microsoftin kehittämää FairLearn-työkalua, jonka algoritmit muuttavat koneoppimismallin tarjoamia ennusteita täyttämään tutkijan valitsemat oikeudenmukaisuuskriteerit. Työkalun algoritmit uudelleenpainottavat datapisteitä toistuvasti niin kauan, että haluttu oikeudenmukaisuuden taso saavutetaan. [62] Delecrazin ym. tutkimuksessa todettiin, että ulkomaalaiset ja oleskeluluvattomat henkilöt olivat aliedustettuina työpaikalle palkatuissa henkilöissä suhteutettuna hakemusten määrään. Tutkijat kehittivät oikeudenmukaisuuden mittareita ja käyttivät Fairlearn-työkalun algoritmeja havaittujen vinoumien vähentämiseen. [56]

Myös jälkikäsittelytekniikat ovat saaneet kritiikkiä. On esimerkiksi huomautettu, että jälkikäsittely erityisesti ryhmien saamia ennusteita tasapainottamalla voi muiden syrjinnän vähentämisen tekniikoiden tavoin heikentää mallin suorituskykyä ja tarkkuutta [2]. Lisäksi käytetyt oikeudenmukaisuusmittarit riippuvat jälkikäsittelijän subjektiivisista valinnoista [60]. Selitettävyyden lisäämiseen pyrkivät mallit ovat taas saaneet kritiikkiä muun muassa siitä, että mallien tuottamat selitykset saattavat olla epäluotettavia eikä tekniikoissa ole standardisoitua lähestymistapaa selitettävyyden mittaamiselle ja arvioinnille [63].

7 Pohdinta

Koneoppimisalgoritmien yleistyvä hyödyntäminen rekrytoinnissa ja sen aiheuttamat ongelmat ovat luoneet tarpeen tarkastella algoritmeissa ilmenevän syrjinnän vähentämisen tekniikoita. Syrjinnän vähentämiseen on kehitetty useita menetelmiä, joiden on myös käytettyjen syrjintämittareiden valossa huomattu vähentävän algoritmien tekemää syrjintää. Näillä tekniikoilla on kuitenkin rajoitteensa. On esimerkiksi osoitettu, että arkaluonteisen tiedon poistaminen aineistosta kokonaan on lähes mahdotonta, sillä aineistoon jäävät sanat voivat toimia epäsuorina viittauksina arkaluonteisiin attribuutteihin. Algoritmi voi esimerkiksi yhdistää tietyt harrastukset sukupuoleen tai tietyn asuinalueen etnisyyteen. Koska algoritmit löytävät tehokkaasti yhteyksiä eri tietojen välillä, saattavat ne edelleen syrjiä tiettyjä ryhmiä piiloyhteyksien takia. [51] Lisäksi tilanteet, joissa yksilöiden kokemaan syrjintään vaikuttavat samanaikaisesti eri tekijät, kuten sukupuoli, etnisyys ja sosioekonominen asema, eivät ole mallinnettavissa algoritmisin menetelmin [2], [43]. Prosessinaikaisten käsittelytekniikoiden on taas todettu vähentävän algoritmin tuottamien ennusteiden tarkkuutta ja oikeellisuutta [29], minkä lisäksi mallien objektiivisuutta on kyseenalaistettu, sillä algoritmin ohjelmoija tekee subjektiivisia valintoja käytettyjen oikeudenmukaisuusmittareiden suhteen [60]. Samaan tapaan myös algoritmin tuloksia optimoivien jälkikäsittelymenetelmien on huomattu vähentävän mallin tuottamien tulosten tarkkuutta [2].

Vaikka syrjinnän vähentämisen tekniikoilla on saatu tutkijoiden mukaan vähennettyä algoritmien tekemää syrjintää, herättää tekniikoiden käyttö rekrytoinnissa myös kysymyksiä. Syrjinnän vähentämiseen tarkoitettujen menetelmien käyttäminen saattaa johtaa kompromisseihin algoritmin tarkkuuden ja oikeudenmukaisuuden välillä. Meritokraattisessa rekrytoinnissa, jossa työpaikat jaetaan ansioiden, kuten koulutuksen ja kokemuksen perusteella, kompromissit voivat johtaa siihen, etteivät algoritmin tuottamat tulokset ole optimaalisia sen enempää tarkkuuden kuin oikeudenmukaisuudenkaan kannalta. Tämän lisäksi myös algoritmien objektiivisuutta voi kyseenalaistaa. Algoritmit pystyvät käsittelemään tietoa systemaattisesti ilman ihmisen välitöntä vaikutusta, mutta mallien kehittäminen sisältää aina algoritmin toimintaan vaikuttavia ihmisen tekemiä subjektiivisia valintoja. Tämä herättää kysymyksen siitä, mitkä arvot pitäisi asettaa etusijalle tai toteuttaa algoritmeissa ja millaiset puolueellisuudet – jos sellaisia sallitaan – ovat hyväksyttäviä [60]. Algoritmit eivät ole sen objektiivisempia kuin niitä kehittävät ja käyttävät ihmisetkään ja kuten ihmisetkin, on tekoäly aina syntynyt jossain kontekstissa. Tämän takia voisi sanoa, että

rekrytoinnissa reiluuden ja oikeudenmukaisuuden ongelmat eivät vähene käyttämällä tekoälyä, vaan tarkastelemalla näitä ongelmia. Lisäksi syrjinnän vähentämisen tekniikoiden haasteiden valossa voidaan pohtia, johtaako algoritmien käyttö toisen syrjinnän muodon vahvistamiseen toisen poistamisen kustannuksella. Bursellin ym. [64] tutkimus havainnollistaa tilannetta: tutkijat vertailivat algoritmien ja ihmisten tekemiä valintoja rekrytoinnissa ja havaitsivat, että suhteessa lähtöaineistoon algoritmin tekemissä esikarsivissa valinnoissa naiset olivat yliedustettuina, kun taas ihmisrekrytoijien algoritmien suosituksiin perustuvissa valinnoissa yliedustettuina olivat hakijat, joiden nimet vaikuttivat eurooppalaisilta. Herääkin kysymys, kuinka pitkälle algoritmien tulisi mennä eri ryhmien tasapainottamisessa ja minkä kustannuksella? Tämä ei ole pelkästään teknologinen, vaan myös eettinen kysymys, joka vaatii jatkuvaa arviointia ja mahdollisten kompromissien ymmärtämistä.

Koneoppimisessa käytetyt syrjinnän vähentämisen tekniikat ovat saaneet lisäksi yleisesti kritiikkiä siitä, ettei niissä kiinnitetä huomiota oikeudenmukaisuuden määritelmään, vaan keskustelussa nojataan hiljaiseksi jäävään oletukseen yhteisestä ja jaetusta oikeudenmukaisuuskäsityksestä [65], [66]. Keskustelussa olettamuksena näyttää olevan, että taloudellisten ja sosiaalisten esteiden poistamisen jälkeen ihmiset ovat tasa-arvoisessa asemassa valinnan suhteen. Ei kuitenkaan ole yksimielisyyttä edes siitä, että tällainen tilanne johtaisi oikeudenmukaiseen valintaan. John Rawls [67] korostaa oikeudenmukaisuusteoriassaan tietämättömyyden verho -ajatuskokeen avulla sitä, miten monet yksilöiden kyvyt ja saavutukset riippuvat olosuhteista, joihin heillä ei ole ollut vaikutusvaltaa ja jotka näyttävät pikemminkin sattumanvaraisilta: esimerkiksi peritty älykkyys tai lapsuuden kasvutilanne vaikuttavat merkittävästi niiden ominaisuuksien osaamiseen, joihin meritokraattisessa valinnassa keskitytään. Rawlsilaisesta näkökulmasta meritokraattisen rekrytoinnin oikeudenmukaisuus näyttäytyy siksi kyseenalaisena. Toisistaan merkittävällä tavalla poikkeavat oikeudenmukaisuuden teoriat osoittavat, että oikeudenmukaisuuden määritelmästä ei ole päästy yhteiskunnallisella tasolla yksimielisyyteen. Tällöin voi olla kyseenalaista olettaa, että algoritmi voisi myöskään koskaan tuottaa kaikkien mielestä täysin oikeudenmukaisia tuloksia.

8 Yhteenveto

Tässä tutkielmassa on tarkasteltu koneoppimismallien käyttöä rekrytoinnissa keskittyen erityisesti syrjintään ja oikeudenmukaisuuteen liittyviin haasteisiin. Koneoppimismallit tarjoavat merkittäviä mahdollisuuksia rekrytointiprosessien tehostamiseen ja automatisointiin, mutta samalla niiden käytössä on ilmennyt eettisiä ongelmia.

Tutkielman ensimmäinen tutkimuskysymys (TK1) on, millaisia ongelmia koneoppimisalgoritmien käyttö on aiheuttanut rekrytoinnissa. Tutkielmassa havaitaan, että keskeisiä ongelmia ovat algoritmin koulutusaineiston epäedustavuus tai puutteellisuus, jotka voivat johtaa syrjiviin lopputuloksiin. Lisäksi tekoälytyökalut voivat heijastaa suunnittelijoidensa ennakkoluuloja ja toisintaa tai jopa vahvistaa ympäröivää yhteiskunnallista epätasa-arvoisuutta. Algoritmien monimutkaisuus ja heikko läpinäkyvyys taas aiheuttavat haasteita päätöksen selitettävyydelle, oikeuttamiselle ja vastuun ottamiseen päätöksestä. Nämä ongelmat korostuvat erityisesti työnhakijoiden esikarsinta- ja valintavaiheissa, joissa tekoälypohjaiset työkalut vaikuttavat hakijoiden etenemiseen rekrytointiprosessissa. Koneoppimisalgoritmit vaikuttavat sopivan hyvin rekrytoinnin manuaalisen ja teknisen työn automatisointiin, kuten työpaikkailmoituksen laatimiseen ja levittämiseen sopivalle yleisölle ehkäpä siksi, että näissä mahdollisen syrjinnän seuraukset ovat pienemmät. Sen sijaan ihmisten inhimillisiä ominaisuuksia ja kykyjä vaativassa arvioinnissa algoritmien käytön on havaittu aiheuttavan eettisiä ongelmia.

Toinen tutkimuskysymys (TK2) keskittyy koneoppimisalgoritmien tekemän syrjinnän vähentämisen tekniikoihin rekrytoinnissa. Algoritmien tekemää syrjintää voidaan vähentää esikäsittelyn, prosessoinninaikaisen käsittelyn sekä jälkikäsittelyn menetelmillä. Esikäsittelyssä aineistoa muokataan ennen algoritmin koulutusta esimerkiksi poistamalla ja piilottamalla sensitiivisiä piirteitä tai lisäämällä aineiston monimuotoisuutta. Prosessoinninaikaisessa käsittelyssä muutetaan algoritmien toimintaa lisäämällä algoritmeihin syväoppimisen tekniikoihin perustuvia oikeudenmukaisuusmittareita. Jälkikäsittelyssä taas pyritään muokkaamaan algoritmin tuottamaa lopputulosta oikeudenmukaisemmaksi tai lisäämään mallin läpinäkyvyyttä. Tutkielman tulosten valossa näyttää kuitenkin siltä, että koneoppimisalgoritmien tekemää syrjintää on haastavaa poistaa algoritmeista kokonaan. Syrjinnän vähentämistekniikat voivat myös tuoda mukanaan toisia syrjinnän muotoja tai heikentää mallien ennustekykyä.

Tutkielman havaintojen perusteella herää kysymys siitä, onko rekrytoinnin arviointia koskevat vaiheet optimaalisia sovelluskohteita koneoppimisalgoritmeille. Rekrytointipäätöksillä voi olla kauaskantoisetkin vaikutukset yksilöiden elämään ja myös laajemmin yhteiskunnan tasolla rekrytoinnin premissit ja päätökset heijastavat ja toisintavat yhteiskunnallisia arvostuksia. Jos rekrytointipäätökset tehdään mahdollisesti syrjivien algoritmien ja mustan laatikon pohjalta, ei tulos vaikuta oikeudenmukaiselta niin distributiivisesta kuin proseduraalisestakaan oikeudenmukaisuusnäkökulmasta käsin. Algoritmien käyttö rekrytointiprosessissa ei ole vain teknologinen kysymys vaan myös eettinen ja yhteiskunnallinen haaste, joka edellyttää tarkkaa harkintaa. Tämän tutkielman valossa algoritmien hyödyntämistä rekrytoinnin arviointi- ja karsintavaiheissa tulisi tarkastella kriittisesti, jotta voidaan varmistaa rekrytointiprosessin oikeudenmukaisuus, läpinäkyvyys ja yhdenvertaisuus.

Jatkotutkimuksessa voitaisiin syventää oikeudenmukaisuusteorioiden soveltamista algoritmiseen arviointiin keskittymällä siihen, miten eri oikeudenmukaisuusteoriat voivat toimia kehyksinä arvioitaessa koneoppimisalgoritmien käyttöä. Tutkimusaihetta voisi myös laajentaa rekrytoinnin ulkopuolelle ja vertailla sitä, miltä algoritmien tekemä syrjintä näyttäytyy toisessa kontekstissa, kuten opiskelijavalinnoissa. Jatkotutkimus syventäisi ymmärrystä siitä, miten koneoppimismalleja voidaan hyödyntää oikeudenmukaisesti ja millaisia eettisiä periaatteita ja käytännön ratkaisuja tarvitaan, jotta teknologia ei johda syrjintään tai epäoikeudenmukaisiin päätöksiin.

Lähteet

[1] Z. Chen, ”Ethics and discrimination in artificial intelligence-enabled recruitment practices”, *Humanit Soc Sci Commun*, vsk. 10, nro 1, s. 567, 2023, doi: 10.1057/s41599-023-02079-x.

[2] E. Albaroudi, T. Mansouri ja A. Alameer, ”A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring”, *AI*, vsk. 5, nro 1, ss. 383–404, 2024, doi: 10.3390/ai5010019.

[3] M. Raghavan, S. Barocas, J. Kleinberg ja K. Levy, ”Mitigating bias in algorithmic hiring: evaluating claims and practices”, teoksessa *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, New York: Association for Computing Machinery, 2020, ss. 469–481. doi: 10.1145/3351095.3372828.

[4] A. Fabris, N. Baranowska, M. J. Dennis, D. Graus, P. Hacker, J. Saldivar, F. Zuiderveen Borgesius ja A. J. Biega, ”Fairness and Bias in Algorithmic Hiring: A Multidisciplinary Survey”, *ACM Trans. Intell. Syst. Technol.*, 2024, doi: 10.1145/3696457.

[5] L. Alexander III, Q. C. Song, L. Hickman ja H. J. Shin, ”Sourcing algorithms: Rethinking fairness in hiring in the era of algorithmic recruitment”, *Int. J. Sel. Assess.*, 2024, doi: 10.1111/ijsa.12499.

[6] J. D’souza, V. Kadam, P. Shinde ja K. Saxena, ”The Quest for Fairness: A Comparative Study of Accuracy in AI Hiring Systems”, teoksessa *2023 3rd Asian Conference on Innovation in Technology (ASIANCON)*, 2023, ss. 1–6. doi: 10.1109/ASIANCON58793.2023.10269895.

[7] B. Goodman ja S. Flaxman, ”European Union Regulations on Algorithmic Decision Making and a “Right to Explanation””, *AI Magazine*, vsk. 38, nro 3, ss. 50–57, 2017, doi: 10.1609/aimag.v38i3.2741.

[8] H. Tuominen, P. Neittaanmäki, E. Niinimäki, I. Pölönen, I. Rautiainen, S. Äyrämö, T. Ruohonen, R. Nyrhinen, A. Ojalainen, P. Vähäkainu ja S-M. Äyrämö, *Tekoälyn perusteita ja sovelluksia*. 2019. Viitattu: 23. syyskuuta 2024. [Verkossa]. Saatavissa: https://jyx.jyu.fi/handle/123456789/64975

[9] S. Bringsjord ja N. S. Govindarajulu, ”Artificial Intelligence”, teoksessa *The Stanford Encyclopedia of Philosophy*, Fall 2024., E. N. Zalta ja U. Nodelman, Toim., Metaphysics Research Lab, Stanford University, 2024. Viitattu: 28. lokakuuta 2024. [Verkossa]. Saatavissa: https://plato.stanford.edu/archives/fall2024/entries/artificial-intelligence/

[10] *Kielitoimiston sanakirja*. Helsinki: Kotimaisten kielten keskus, 2024. Viitattu: 28. lokakuuta 2024. [Verkossa]. Saatavissa: URN:NBN:fi:kotus-201433. Verkkojulkaisu HTML. Päivitettävä julkaisu. Päivitetty 19.3.2024

[11] Mark Coeckelbergh, *AI Ethics*. teoksessa MIT Press Essential Knowledge Series. Cambridge, Massachusetts: The MIT Press, 2020. Viitattu: 20. syyskuuta 2024. [Verkossa]. Saatavissa: https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=2399415&site=ehost-live&scope=site

[12] P. Raatikainen, ”Tekoäly, ihminen ja yhteiskunta - johdatusta teemaan”, teoksessa *Tekoäly, ihminen ja yhteiskunta. Filosofisia näkökulmia.*, P. Raatikainen, Toim., Helsinki: Gaudeamus, 2021, ss. 7–20.

[13] S. Russel ja P. Norvig, *Artificial intelligence: a modern approach*. Harlow: Pearson, 2020.

[14] W. Ertel, *Introduction to artificial intelligence*, Third edition. teoksessa Undergraduate topics in computer science. Cham, Switzerland: Springer, 2020.

[15] I. Goodfellow, Y. Bengio ja A. Courville, *Deep Learning*. MIT Press, 2016.

[16] I. N. Da Silva, D. Hernane Spatti, R. Andrade Flauzino, L. H. B. Liboni ja S. F. Dos Reis Alves, *Artificial Neural Networks*. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-43162-8.

[17] Y. Zhang ja Z. Teng, ”Natural Language Processing: A Machine Learning Perspective”, Higher Education from Cambridge University Press. Viitattu: 28. syyskuuta 2024. [Verkossa]. Saatavissa: https://www.cambridge.org/highereducation/books/natural-language-processing/2DFF9BBE02CFB388E3EA93105E2DBCFB/introduction/F8765EDEEF2E065B93EB34543312D35B

[18] O. Campesato, *Natural Language Processing Fundamentals for Developers*. Bloomfield, UNITED STATES: Mercury Learning & Information, 2021. Viitattu: 28. syyskuuta 2024. [Verkossa]. Saatavissa: http://ebookcentral.proquest.com/lib/kutu/detail.action?docID=6647713
[19] A. Altman, ”Discrimination”, teoksessa *The Stanford Encyclopedia of Philosophy*, Winter 2020., E. N. Zalta, Toim., Metaphysics Research Lab, Stanford University, 2020. Viitattu: 4. marraskuuta 2024. [Verkossa]. Saatavissa: https://plato.stanford.edu/archives/win2020/entries/discrimination/
[20] K. Lippert-rasmussen, ”The badness of discrimination”, *Ethic Theory Moral Prac*, vsk. 9, nro 2, ss. 167–185, 2006, doi: 10.1007/s10677-006-9014-x.
[21] H. Parviainen, ”Challenges of Direct Discrimination in Algorithmic Recruitment: Insuperable or Not?*”, *IJCL*, vsk. 40, nro Issue 4, ss. 437–466, joulu 2024, doi: 10.54648/IJCL2024017.
[22] S. Barocas, M. Hardt ja A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.
[23] D. Gotterbarn ja D. Kreps, ”Being a data professional: give voice to value in a data driven society”, *AI Ethics*, vsk. 1, nro 2, ss. 195–203, touko 2021, doi: 10.1007/s43681-020-00027-y.
[24] M. J. Sandel, *Oikeudenmukaisuus*. Helsinki: HS kirjat, 2012.
[25] K. Herne, *Mitä oikeudenmukaisuus on?* Helsinki: Gaudeamus, 2012.
[26] D. Miller, ”Justice”, teoksessa *The Stanford Encyclopedia of Philosophy*, Fall 2023., E. N. Zalta ja U. Nodelman, Toim., Metaphysics Research Lab, Stanford University, 2023. Viitattu: 12. marraskuuta 2024. [Verkossa]. Saatavissa: https://plato.stanford.edu/archives/fall2023/entries/justice/
[27] L. Morse, M. H. M. Teodorescu, Y. Awwad ja G. C. Kane, ”Do the Ends Justify the Means? Variation in the Distributive and Procedural Fairness of Machine Learning Algorithms.”, *Journal of Business Ethics*, vsk. 181, nro 4, ss. 1083–1095, 2022, doi: 10.1007/s10551-021-04939-5.
[28] P. Seppälä ja M. Malecka, ”AI and discriminative decisions in recruitment: Challenging the core assumptions”, *Big Data Soc.*, vsk. 11, nro 1, s. 20539517241235872, 2024, doi: 10.1177/20539517241235872.
[29] C. Kim, J. Choi, J. Yoon, D. Yoo ja W. Lee, ”Fairness-Aware Multimodal Learning in Automatic Video Interview Assessment”, *IEEE Access*, vsk. 11, ss. 122677–122693, 2023, doi: 10.1109/ACCESS.2023.3325891.
[30] V. C. Müller, ”Ethics of Artificial Intelligence and Robotics”, teoksessa *The Stanford Encyclopedia of Philosophy*, Fall 2023., E. N. Zalta ja U. Nodelman, Toim., Metaphysics Research Lab, Stanford University, 2023. Viitattu: 15. lokakuuta 2024. [Verkossa]. Saatavissa: https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/
[31] A. Kantola, S. Aaltonen, L. Haikkola, L. Junnilainen, E. Luhtakallio, P. Patana, J. Timonen ja P. Tuominen, *Kahdeksan kuplan Suomi: yhteiskunnan muutosten syvät tarinat*. Helsinki: Gaudeamus, 2022.
[32] C. Bartneck, C. Lütge, A. Wagner ja S. Welsh, *An Introduction to Ethics in Robotics and AI*. Springer Nature, 2021. doi: 10.1007/978-3-030-51110-4.
[33] K. Siau ja W. Wang, ”Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI”, *Journal of Database Management (JDM)*, vsk. 31, nro 2, ss. 74–87, 2020, doi: 10.4018/JDM.2020040105.
[34] M. Franssen, G.-J. Lokhorst ja I. van de Poel, ”Philosophy of Technology”, teoksessa *The Stanford Encyclopedia of Philosophy*, Fall 2024., E. N. Zalta ja U. Nodelman, Toim., Metaphysics Research Lab, Stanford University, 2024. Viitattu: 15. lokakuuta 2024. [Verkossa]. Saatavissa: https://plato.stanford.edu/archives/fall2024/entriesechnology/
[35] I. Niiniluoto, *Tekniikan filosofia*. Helsinki: Gaudeamus, 2020.
[36] A. Laitinen, ”Mustan laatikon ongelma ja oikeus selityksen saamiseen”, teoksessa *Tekoäly, ihminen ja yhteiskunta. Filosofisia näkökulmia.*, P. Raatikainen, Toim., Helsinki: Gaudeamus, 2021, ss. 181–195.
[37] B. C. Stahl, D. Schroeder ja R. Rodrigues, *Ethics of Artificial Intelligence : Case Studies and Options for Addressing Ethical Challenges*. Springer Nature, 2023. Viitattu: 10. lokakuuta 2024. [Verkossa]. Saatavissa: https://directory.doabooks.org/handle/20.500.12854/93981
[38] S. Caton ja C. Haas, ”Fairness in Machine Learning: A Survey”, *ACM Comput. Surv.*, vsk. 56, nro 7, ss. 1–38, 2024, doi: 10.1145/3616865.

[39] D. F. Mujtaba ja N. R. Mahapatra, ”Ethical Considerations in AI-Based Recruitment”, teoksessa *2019 IEEE International Symposium on Technology and Society (ISTAS)*, 2019, ss. 1–7. doi: 10.1109/ISTAS48451.2019.8937920.
[40] J. Hsu, ”Can AI hiring systems be made antiracist? Makers and users of AI-assisted recruiting software reexamine the tools’ development and how they’re used - [News]”, *IEEE Spectrum*, vsk. 57, nro 9, ss. 9–11, 2020, doi: 10.1109/MSPEC.2020.9173891.
[41] A. Kauppinen, ”Osaammeko rakentaa moraalisia toimijoita?”, teoksessa *Tekoäly, ihminen ja yhteiskunta. Filosofisia näkökulmia.*, P. Raatikainen, Toim., Helsinki: Gaudeamus, 2021, ss. 131–158.
[42] Z. Chen, ”Collaboration among recruiters and artificial intelligence: removing human prejudices in employment”, *Cogn Tech Work*, vsk. 25, nro 1, ss. 135–149, 2023, doi: 10.1007/s10111-022-00716-0.
[43] E. K. Kelan, ”Algorithmic inclusion: Shaping the predictive algorithms of artificial intelligence in hiring”, *Hum. Resour. Manag. J.*, vsk. 34, nro 3, ss. 694–707, 2024, doi: 10.1111/1748-8583.12511.
[44] Gagandeep, J. Kaur, S. Mathur, S. Kaur, A. Nayyar, S. P. Singh ja M. Mathur, ”Evaluating and mitigating gender bias in machine learning based resume filtering”, *Multimed Tools Appl*, vsk. 83, nro 9, ss. 26599–26619, 2024, doi: 10.1007/s11042-023-16552-x.
[45] A. Köchling ja M. C. Wehner, ”Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development”, *Bus Res*, vsk. 13, nro 3, ss. 795–848, 2020, doi: 10.1007/s40685-020-00134-w.
[46] S. Njoto, M. Cheong, R. Lederman, A. McLoughney, L. Ruppanner ja A. Wirth, ”Gender Bias in AI Recruitment Systems: A Sociological-and Data Science-based Case Study”, teoksessa *2022 IEEE International Symposium on Technology and Society (ISTAS)*, 2022, ss. 1–7. doi: 10.1109/ISTAS55053.2022.10227106.
[47] B. M. Booth, L. Hickman, S. K. Subburaj, L. Tay, S. E. Woo ja S. K. D’Mello, ”Bias and Fairness in Multimodal Machine Learning: A Case Study of Automated Video Interviews”, teoksessa *Proceedings of the 2021 International Conference on Multimodal Interaction*, Montréal QC Canada: ACM, 2021, ss. 268–277. doi: 10.1145/3462244.3479897.
[48] C. Lee ja K. Cha, ”FAT-CAT-Explainability and augmentation for an AI system: A case study on AI recruitment-system adoption”, *Int. J. Hum.-Comput. Stud.*, vsk. 171, s. 102976, 2023, doi: 10.1016/j.ijhcs.2022.102976.
[49] K. V. Deshpande, S. Pan ja J. R. Foulds, ”Mitigating Demographic Bias in AI-based Resume Filtering”, teoksessa *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, teoksessa UMAP ’20 Adjunct. New York, NY, USA: Association for Computing Machinery, 2020, ss. 268–275. doi: 10.1145/3386392.3399569.
[50] L. R. McCarey ja T. S. McTavish, ”Optimizing sample diversity with fairness constraints on imbalanced, sparse, hiring data”, teoksessa *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, teoksessa GECCO ’22. New York, NY, USA: Association for Computing Machinery, 2022, ss. 727–730. doi: 10.1145/3520304.3529028.
[51] A. Peña, I. Serna, A. Morales ja J. Fierrez, ”Bias in Multimodal AI: Testbed for Fair Automatic Recruitment”, teoksessa *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, kesä 2020, ss. 129–137. doi: 10.1109/CVPRW50498.2020.00022.
[52] B. Putra, K. Azizah, C. Olivia Mawalim, I. Akmal Hanif, S. Sakti, C. Wee Leong ja S. Okada, ”MAG-BERT-ARL for Fair Automated Video Interview Assessment”, *IEEE Access*, vsk. 12, ss. 145188–145205, 2024, doi: 10.1109/ACCESS.2024.3473314.
[53] G. Sogancioglu, H. Kaya ja A. A. Salah, ”Using Explainability for Bias Mitigation: A Case Study for Fair Recruitment Assessment”, teoksessa *Proceedings of the 25th International Conference on Multimodal Interaction*, teoksessa ICMI ’23. New York, NY, USA: Association for Computing Machinery, 2023, ss. 631–639. doi: 10.1145/3577190.3614170.
[54] A. Ortega, J. Fierrez, A. Morales, Z. Wang ja T. Ribeiro, ”Symbolic AI for XAI: Evaluating LFIT Inductive Programming for Fair and Explainable Automatic Recruitment”, teoksessa *2021*

IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW), 2021, ss. 78–87. doi: 10.1109/WACVW52041.2021.00013.

[55] M. Russo, Y. Chudasama, D. Purohit, S. Sawischa ja M.-E. Vidal, ”Employing Hybrid AI Systems to Trace and Document Bias in ML Pipelines”, *IEEE Access*, vsk. 12, ss. 96821–96847, 2024, doi: 10.1109/ACCESS.2024.3427388.

[56] S. Delecraz, L. Eltarr, M. Becuwe, H. Bouxin, N. Boutin ja O. Oullier, ”Making Recruitment More Inclusive: Unfairness Monitoring With A Job Matching Machine-Learning Algorithm”, teoksessa *2022 IEEE/ACM International Workshop on Equitable Data & Technology (FairWare)*, 2022, ss. 34–41. doi: 10.1145/3524491.3527309.

[57] N. Tilmes, ”Disability, fairness, and algorithmic bias in AI recruitment”, *Ethics Inf Technol*, vsk. 24, nro 2, s. 21, 2022, doi: 10.1007/s10676-022-09633-2.

[58] A. Morales, J. Fierrez, R. Vera-Rodriguez ja R. Tolosana, ”SensitiveNets: Learning Agnostic Representations with Application to Face Images”, 2020, *arXiv*: arXiv:1902.00334. doi: 10.48550/arXiv.1902.00334.

[59] X. Zhao, S. Fabbrizz, P. R. Lobo, S. Ghodsi, K. Broelemann, S. Staab ja G. Kasneci, ”Adversarial Reweighting Guided by Wasserstein Distance for Bias Mitigation”, 2023, *arXiv*: arXiv:2311.12684. Viitattu: 13. marraskuuta 2024. [Verkossa]. Saatavissa: http://arxiv.org/abs/2311.12684

[60] S. Fazelpour ja D. Danks, ”Algorithmic bias: Senses, sources, solutions”, *Philosophy Compass*, vsk. 16, nro 8, s. e12760, 2021, doi: 10.1111/phc3.12760.

[61] M. T. Ribeiro, S. Singh ja C. Guestrin, ”’Why Should I Trust You?’: Explaining the Predictions of Any Classifier”, teoksessa *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, 2016, ss. 1135–1144. doi: 10.1145/2939672.2939778.

[62] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, K. Walker ja A. Design, ”Fairlearn: A toolkit for assessing and improving fairness in AI”, 2020, [Verkossa]. Saatavissa: https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf

[63] W. Saeed ja C. Omlin, ”Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities”, *Knowledge-Based Systems*, vsk. 263, s. 110273, 2023, doi: 10.1016/j.knosys.2023.110273.

[64] M. Bursell ja L. Roumbanis, ”After the algorithms: A study of meta-algorithmic judgments and diversity in the hiring process at a large multisite company”, *Big Data Soc.*, vsk. 11, nro 1, s. 20539517231221758, 2024, doi: 10.1177/20539517231221758.

[65] S. A. Friedler, C. Scheidegger ja S. Venkatasubramanian, ”On the (im)possibility of fairness”, 2016, *arXiv*: arXiv:1609.07236.

[66] R. Binns, ”Fairness in Machine Learning: Lessons from Political Philosophy”, arXiv, 2018. doi: 10.48550/arXiv.1712.03586.

[67] J. Rawls, *A theory of justice*. Cambridge, Massachusetts: Harvard University Press, 2005.