Haaga-Helia
ammattikorkeakoulu Oy

Luottamuksenarvoisen tekoälyn edellytykset

Sanna Virtanen

Master opinnäytetyö
Liiketoiminnan teknologiat
2020

Haaga-Helia
ammattikorkeakoulu Oy

Tiivistelmä
30.11.2020

Tekijä(t)
Sanna Virtanen

Suuntautuminen
Digitaalisen liiketoiminnan mahdollisuudet

Opinnäytetyön nimi
Luottamuksenarvoisen tekoälyn edellytykset

Sivumäärä + liitesivumäärä
106+7

Elinkeinoministeri Mika Lintilän toimeksiannosta vuonna 2017 Suomessa käynnistettiin Tekoälyaika-niminen toimenpideohjelma, joka muun muassa haastoi suomalaisia yrityksiä pohtimaan tekoälyn etiikkaa ja käyttöä eri näkökulmista ja luomaan omat tekoälyn periaatteet. Sen tavoitteena oli varmistaa, että Suomessa tekoälyn käyttö ja kehitys olisi ihmiskeskeistä ja luotettavaa.

Samasta syystä opinnäytetyöntilaaja halusi selvittää, miten tekoälyä käytetään luottamuksenarvoisesti. Työssä selvitettiin, millaisia käsitteitä ja teemoja liittyy luottamuksenarvoiseen tekoälyn hyödyntämiseen: mitä tarkoittavat esimerkiksi eettinen, oikeudenmukainen, läpinäkyvä tai selitettävä tekoäly. Lisäksi selvitettiin, mitkä tekijät jo ohjaavat tekoälyn käyttöä, millaisia haasteita tekoälyn hyödyntämisessä voi ilmetä luottamuksenarvoisuuden näkökulmasta, miten haasteita on ratkottu ja miten ja kenelle luottamuksenarvoisuudesta tulee viestiä. Tavoitteena oli luoda työeläketoimialalla toimivalle työntilaajalle selvitys siitä, miten luottamuksenarvoisuutta rakennetaan, onko eettisille periaatteille tarvetta, ja millaisin toimin työtä voitaisiin jatkaa.

Lähestymistavaksi valikoitui tapaustutkimus, joka ei pyri tilastolliseen yleistettävyyteen vaan ajallisen, paikallisen ja yksityiskohtaisen tiedon tuottamiseen ja kohteen syvälliseen ymmärtämiseen. Työ toteutettiin tutkimuksellisen kehittämisen mallin mukaisesti. Tutkimusaineisto koostui sekä aiheeseen liittyvien eri alojen asiantuntijoiden että työntilaajan edustajien teemahaastatteluista. Ulkopuolisten asiantuntijoiden haastatteluilla varmennettiin tietoperustan kattavuutta ja eri näkökulmien riittävyyttä ja työntilaajan haastatteluilla selvitettiin työntilaajan nyky- ja tavoitetilaa. Sekundäärisenä aineistonkeruumenetelmänä käytettiin dokumenttianalyysiä, jonka avulla hyödynnettiin työntilaajan olemassa olevaa aineistoa. Analyysimenetelminä käytettiin teemoittelua ja sisältöanalyysiä. Tietoperustan mallina käytettiin oivalluttavaa-perinteistä mallia, joka mahdollisti referoinnin lisäksi myös opinnäytetyöntekijän oman ajattelun esille tuonnin ja lähteiden välisen vuoropuhelun.

Tutkimusaineiston analyysin tuloksena löydettiin useita tekoälyn luottamuksenarvoisuuteen liittyviä jaettuja haasteita sekä ajankohtaisia ja käytännöllisiä ratkaisuvaihtoehtoja, mutta ennen kaikkea ymmärrys, että tekoälyn etiikassa on kysymys dialogista, jota on käytävä jatkuvasti ja joka edellyttää myös tietoisia valintoja ja kompromissejä. Varsinaisten uusien periaatteiden luomisen sijaan tulisi keskittyä tekoälyn etiikan normatiiviseen ytimeen ja käytännön luottamuksenarvoisuuden rakentamiseen läpi tekoälyratkaisun elinkaaren. Opinnäytetyön tuloksena syntyi ajankohtainen selvitys aiheesta sekä ehdotus jatkotoimenpiteistä, joilla edistetään työntilaajan eettistä ja vastuullista liiketoimintaa.

Asiasanat
tekoäly, data, luottamus, eettisyys, vastuullisuus, periaatteet

Sisällys

Lyhenteet

AI	Artificial Intelligence eli tekoäly, koneäly tai keinoäly
AI HLEG	High-Level Expert Group on AI on Euroopan komission nimittämä, riippumattomien korkean tason asiantuntijoiden ryhmä, joka pyrkii tukemaan Euroopan AI-strategiaa ja on tähän mennessä luonut esimerkiksi eettiset periaatteet ja seitsemän vaatimusta tekoälyn käytölle.
ALTAI	AI HLEG:n luotettavan tekoälyn varmistamiseksi tekemä itsearviointilista, joka pohjautuu ryhmän tekemiin eettisiin periaatteisiin ja vaatimuksiin
EK	Elinkeinoelämän keskusliitto
ETK	Eläketurvakeskus on lakisääteinen yhteistyöelin ja yhteisten palvelujen tuottaja työeläketoimialalla.
FCAI	Finnish Center for Artificial Intelligence eli Suomen tekoälykeskus
GDPR	General Data Protection Regulation eli EU:n yleinen tietosuoja-asetus 679/2016
IEEE	Suuri, kansainvälinen tekniikan alan järjestö
IoT	Internet of Things eli esineiden internet tarkoittaa esineiden yhdistämistä internetiin niin, että ne voivat kerätä, jakaa, lähettää ja vastaanottaa tietoa verkon yli.
ISO	Kansainvälinen standardoimisjärjestö
TELA	Työeläkevakuuttajat TELA ry on Suomessa toimivien työeläkevakuuttajien edunvalvontajärjestö.

1 Johdanto

”*Tekoälyn eettinen ulottuvuus ei ole mikään ylimääräinen ominaisuus. Jotta yhteiskuntamme voi hyödyntää teknologioita kaikilta osin, sen täytyy pystyä luottamaan niihin. Eettinen tekoäly hyödyttää kaikkia.*” *(*Euroopan komissio 2019.)

Eettisestä ja vastuullisesta tekoälystä puhutaan paljon. Vuoden 2018 puolivälistä vuoden 2019 puoliväliin asti yli 3 600:n tutkitun, tekoälyyn liittyvän, globaalin uutisartikkelin dominoivimmat aiheet olivat tekoälyn eettisen käytön viitekehykset ja periaatteet, algoritmiset vinoumat, kasvojentunnistuksen käyttö ja suurten teknologiayhtiöiden rooli. Myös tekoälyn käyttöön liittyvä lainsäädännöllinen keskustelu on lisääntynyt merkittävästi kongressien ja komiteoiden raportein ja lausumin ja niin eurooppalaisessa kuin kansallisessakin lainsäädännössä on jo otettu ensimmäisiä askeleita kohti tarkempaa juridista viitekehystä. (Perrault ym. 2019, 7; Saidot 2019.)

Tämä näkyy myös Suomessa. Sanna Marinin hallitusohjelma korostaa eettisen sekä sosiaalisesti ja taloudellisesti kestävän data- ja tekoälypolitiikan ja sääntelyn tärkeyttä sekä seuraa tekoälyn vaikutuksia muun muassa yhdenvertaisuuteen (Valtioneuvosto 2019, 73, 81). Tekoälyaika-ohjelman etiikkahaasteeseen – johon osallistumista tämänkin opinnäytetyöntilaaja pohti – ovat vastanneet jo useat kymmenet suomalaiset yritykset. Haasteen tavoitteena oli muun muassa kyseenalaistaa etiikan ja innovoinnin vastakkainasettelua ja nostaa vastuullisuus tekoälyn käytön normiksi. Yritykset ovat pyrkineet omien eettisten tekoälyä koskevien periaatteiden avulla selventämään *”suhdettaan tekoälyn kehittämiseen ja soveltamiseen”* ja toimeksiannon mukaisesti määrittäneet, *”miten tekoälyä käytetään reilulla ja luottamusta herättävällä tavalla”.* (Työ- ja elinkeinoministeriö 2019, 102; Ollila 2019, 95, 97.)

Puhe ei kuitenkaan riitä. Strategisissa teknologiatrendeissään Gartner toteaa, että vastuullinen tekoäly tarkoittaa ennen kaikkea siirtymistä julistuksista ja periaatteista operatiiviseen toimintaan niin organisaatio- kuin yhteiskuntatasollakin (Burke 2020, 11). Siksi tämän opinnäytetyön ydin on syventää yleistä ymmärrystä siitä, mitä tekoälyn etiikalla tarkoitetaan, millaisia tekoälyn eettisyyteen, laillisuuteen ja turvallisuuteen liittyvät haasteet voivat olla ja miten niitä voidaan ratkoa. Tapaustutkimuksena selvitetään, millaisia haasteet ja eettiset kysymykset ovat työeläketoimialaa edustavan työntilaajan kontekstissa ja miten tekoälyn etiikka ylipäätään suhteutuu yritys- ja yhteiskuntavastuuseen. Iso kysymys on, mitä eettisillä periaatteilla voidaan saavuttaa ja miltä osin lainsäädäntö on puutteellista. Kun on saatu vastaus siihen, *mitä* tekoälyn luottamuksenarvoisuus tarkoittaa, voidaan miettiä, *miten* luottamuksenarvoisuutta voidaan konkreettisesti rakentaa.

Työntilaajana toimii yksi suomalaisista keskinäisistä työeläkevakuutusyhtiöistä. Työeläkeyhtiöiden lakisääteisenä tehtävänä on turvata asiakkaidensa työeläkkeet keräämällä yrittäjän eläkevakuutusmaksuja (YEL) ja työnantajan eläkevakuutusmaksuja (TyEL), sijoittamalla osa kerätyistä varoista nykyisiä ja tulevia eläkkeitä varten ja maksamalla kerääntyneet eläkkeet eläkkeensaajille eläkepäätösten mukaisesti. Vastuullisuus nähdään sisäänkirjoitettuna työntilaajan lakisääteiseen tehtävään ja luottamus ja tehokkuus ovat työeläkejärjestelmän kulmakiviä. Myös työeläkelakien yhdenmukainen toimeenpano on tehokkuuden ohella hajautetun työeläkejärjestelmän keskeinen tavoite.

Julkisen ja yksityisoikeudellisen sektorin väliin jäävän työeläketoimialan osalta kysymys ei liity vain siihen, mitä eettisellä itsesääntelyllä voidaan saavuttaa, vaan myös siihen, miltä osin itsesääntely on edes mahdollista. Opinnäytetyön kirjoitushetkellä oikeusministeriössä pohditaan vielä, miltä osin ja millä edellytyksillä edes tekoälyä sisältämättömät automaattiset hallintopäätökset ovat työeläketoimialalla – mutta myös esimerkiksi Veron, Kelan ja Maahanmuuttoviraston osalta – sallittuja ja laillisia. Kiinnostavaksi kysymykseksi työntilaajan kontekstissa muodostuu siis, miltä osin työeläkeyhtiön toiminta on yksityisoikeudellista ja miltä osin julkisen tehtävän hoitamista ja mitä se tarkoittaa tekoälyn kannalta.

Työntilaajan konteksti edellyttää luottamuksenarvoisen tekoälyn tarkastelua niin yritysten, yhteiskunnan, asiakkaiden kuin kansalaistenkin näkökulmasta, jolloin julkinen opinnäytetyö toimii itsenäisenä kokonaisuutena kaikille luottamuksenarvoisesta tekoälystä kiinnostuneille. Työeläketoimialaa koskeva selvitys tehdään työntilaajalle ja pääosin salataan. Opinnäytetyö toimii osana sen viitekehyksen rakentamista, jonka puitteissa dataan perustuvia hankkeita voidaan edistää työntilaajan arvojen mukaisesti huomioiden tekoälyn mukanaan tuomat uudet haasteet ja mahdollisuudet. Lisäksi työ edesauttaa valmistautumista mahdollisiin velvoittaviin säännöksiin.

Tekoälyn periaatteista puhuttaessa puhutaan usein tekoälyn *eettisistä* periaatteista. Toisaalta esimerkiksi Montrealin julistus Montréal Declaration for Responsible AI puhuu *vastuullisesta* ja Euroopan komission nimittämä, korkean tason asiantuntijaryhmä AI HLEG *luotettavasta* tekoälystä ja sen kolmesta edellytyksestä: laillisuudesta, eettisyydestä ja teknisestä ja sosiaalisesta luotettavuudesta. Tässä opinnäytetyössä käytetään termiä *luottamuksenarvoisuus* kattamaan nämä toisiinsa nivoutuvat edellytykset ja korostamaan luottamuksenarvoisuuden osoittamista; aitoa ja aktiivista toimijuutta ja yhteiskunnallista hyväksyttävyyttä.

2 Tavoitteet

Opinnäytetyön tarkoituksena on selvittää ja koota, millaisia käsitteitä ja teemoja luottamuksenarvoiseen tekoälyyn liittyy, mitkä tekijät tekoälyn luottamuksenarvoisuutta uhkaavat, mikä tekoälyn käytössä edellyttää erityistä huomiota, millaisten periaatteiden tulisi ohjata luottamuksenarvoista tekoälyn käyttöä ja miten periaatteet ilmenevät käytännön työssä.

2.1 Odotetut tulokset

Työn tavoitteena on toimittaa työntilaajalle

- selvitys tekoälyn luottamuksenarvoiseen käyttöön liittyvistä haasteista ja toimista ratkoa niitä
- jatkokehitysehdotus, mitä työntilaajan kontekstissa on huomioitava, ja sekä mahdolliset alustavat eettiset periaatteet

2.2 Tutkimuskysymykset

Tutkimuskysymysten avulla pyritään selvittämään, mitä luottamuksenarvoinen tekoäly edellyttää ja onko periaatteellinen lähestymistapa tekoälyn etiikkaan oikea.

- TK1. Mitkä tekijät jo ohjaavat tekoälyn käyttöä?
- TK2. Mitä tekoälyn eettisillä periaatteilla voidaan ratkoa?
- TK3. Miten tekoälyä käytetään luottamuksenarvoisesti?

2.3 Rajaukset

Työssä sivutaan sitä, mitä luottamuksenarvoisuuden edistämiseksi on jo tehty sekä EU-tasolla että kansallisella tasolla ja mitä toisaalta on vielä tekemättä. Työssä käydään läpi esimerkiksi Euroopan komission nimittämän huippuasiantuntijoiden työryhmän, AI HLEG:n, luotettavan tekoälyn periaatteita, vaatimuksia ja niihin liittyvää alustavaa itsearviointilistaa (ALTAI), sekä aihetta sivuavia suomalaisia esiselvityksiä ja lainvalmisteluja. Työhön kuuluu myös tutustuminen eri tahojen periaatteisiin, niitä sivuaviin julkaisuihin sekä tekoälyn käyttöön.

Työhön ei kuulu sen arviointi, millaisia ovat työntilaajan nykyiset valmiudet hyödyntää tekoälyä, vaan tekoälyn liittyvän maturiteettitason mittaamiseen on olemassa omat työkalunsa ja näkökulmansa esimerkiksi täällä: https://ai.digimaturity.vtt.fi/.

Työn tarkoituksena ei ole ottaa kantaa yksittäisiin oikeudellisiin tai eettisiin kysymyksiin, kuten esimerkiksi siihen, onko oikein ottaa tietoinen riski, että yksi asiakkaista saa väärän päätöksen, jos valtaosa asiakkaista saisi siten päätöksensä nopeammin ja oikeudenmukai-

semmin (mikäli lakisääteisen tehtävän hoitamisessa sallittaisiin automaattinen hallintopäätöksenteko ylipäätään) tai että jos asiakas poistattaa henkilötietonsa ja haluaa itsensä unohdettavan, onko hänet unohdettava myös opetusdatasta tai algoritmeista. Tällaisista yksityiskohtaisista kysymyksistä voi lukea lisää esimerkiksi Ida Koskisen pro gradusta (2017). Opinnäytetyössä ei myöskään ratkota mustan laatikon ongelmaa tai muita teknologisia haasteita. Algoritmeistä ja niiden läpinäkyvyyden, reiluuden ja vastuullisen kysymyksistä ja muista yhteiskunnallisista ja sosiaalisista riskeistä voi lukea lisää esimerkiksi Algoritmien valta –Neutraalius ja puolueellisuus koneellisessa päätöksenteossa (https://algoritmitutkimus.fi/).

Opinnäytetyöntekijä pysyy työssään ulkopuolisen tarkkailijan roolissa lukuun ottamatta haastatteluja, jotka edellyttävät osallistuvampaa tutkijan roolia.

2.4 Rakenne

Tietoperusta koostuu kolmannen (3) pääluvun alaluvuista, joista ensimmäisessä (3.1) poraudutaan tämän hetken tekoälyyn teknologioina ja avataan käsitteen monitulkintaisuutta. Toisessa alaluvussa (3.2) tutustutaan työntilaajan kontekstiin: työeläketoimialaan, sen erityispiirteisiin, lainsäädäntöön ja tekoälyn käyttökohteisiin ja kolmannessa (3.3) pyritään kokoamaan tekoälyn luottamuksenarvoisuuteen liittyviä haasteita läpi tekoälyn elinkaaren datankeruusta tekoälyä sisältäviin ratkaisuihin ja niiden käyttöön. Neljännessä alaluvussa (3.4) syvennetään näitä luottamuksenarvoisuuden teemoja ja edellytyksiä ja viidennessä alaluvussa (3.5) tarkastellaan tekoälyn etiikkaa ja eettisiä periaatteita yritysten näkökulmasta. Käsitteitä avataan liitteessä 1.

Työntilaajalle tehtyä selvitystä kuvataan pääluvuissa neljä (4) ja viisi (5), joissa kuvataan, miten tutkimus metodologisesti tehtiin, miten tutkimuksen luotettavuus varmistettiin ja miten varsinainen tiedonkeruu ja aineistonanalyysi toteutettiin ja millaisia tuloksia haastatteluin saatiin. Liite 4 täydentää tutkimustuloksia. Haastattelukysymykset löytyvät liitteistä 2 ja 3. Johtopäätöksissä pääluvussa kuusi (6) vastataan tutkimuskysymyksiin, pohditaan tavoitteiden ja odotettujen tulosten saavuttamista, prosessia ja omaa oppimista sekä ehdotetaan aihioita jatkotutkimuksiksi. Dokumenttianalyysin julkinen versio on liitteessä 5a. Työntilaajan haastatteluiden ja dokumenttianalyysin tarkemmat tulokset sekä jatkotoimenpide-ehdotukset on salattu työn liitteisiin 5b–10. Viimeinen pääluku (7) kokoaa lyhyesti yhteen koko tutkimuksen.

Peittomatriisilla (taulukko 1) kuvataan tietoperustan ja tutkimuksen välistä yhteyttä. Siinä esitetyt haastattelukysymykset olivat joko kaikille tai osalle vastaajista yhteisiä.

Taulukko 1. Peittomatriisi

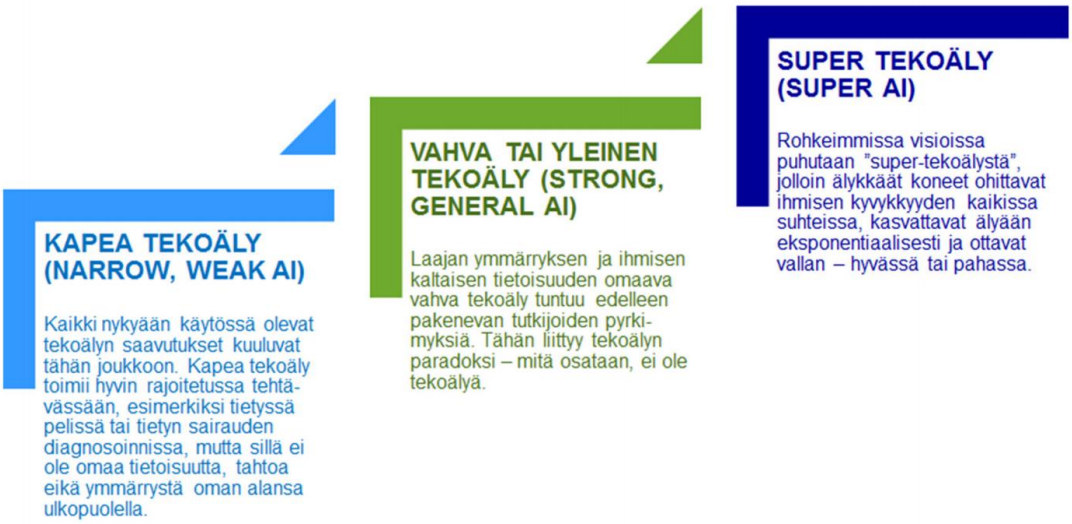
Tutkimuskysymykset	**Tietoperusta**	**Tulokset**	**Tiedonkeruutapa / Haastattelujen kysymykset**
TK1. Mitkä tekijät jo ohjaavat tekoälyn käyttöä?	3.2 Työeläketoimialan ominaispiirteet 3.3.1 Datan keruu ja käyttö 3.4.1 Eettisyys 3.4.2 Laillisuus, sääntely ja valvonta 3.5 Luottamuksenarvoisuuden rakentaminen yrityksissä	5.2 Ulkoisten haastattelujen tulokset 5.3 Työntilaajan haastattelujen tulokset (pääosin salattu) 6 Tulkinta ja johtopäätökset Salatut liitteet	Onko yksityisen toimijuuden ja lakisääteisen tehtävän hoitamisen välinen ero selvä? Asettaako se rajoitteita tekoälyn käytölle? Mitkä tekijät ohjaavat tekoälyn kehittämistä ja hyödyntämistä yrityksessä? / Mitkä tekijät ja tahot ohjaavat tekoälyn hyödyntämistä omalla alallasi? Miten ja kenen edellä mainittuja haasteita tulisi ratkoa? / Minkä tekijöiden tai tahojen pitäisi ohjata datan ja tekoälyn hyödyntämistä? Lisäksi dokumenttianalyysi työntilaajaa ohjaavista periaatteista ja arvoista suhteessa eri tahojen eettisiin periaatteisiin (julkinen liite 5a, salattu liite 5b)
TK2. Mitä tekoälyn eettisillä periaatteilla voidaan ratkoa?	3.4.9 Yhteenveto: Luottamuksenarvoinen tekoäly 3.5 Luottamuksenarvoisuuden rakentaminen yrityksissä	5.2 Ulkoisten haastattelujen tulokset 5.3 Työntilaajan haastattelujen tulokset (pääosin salattu) 6 Tulkinta ja johtopäätökset Salatut liitteet	Miten suhtaudut yritysten omiin tekoälyn eettisiin periaatteisiin? / Tulisiko yrityksen luoda omat eettiset periaatteet tekoälyä varten? Miksi? Millaisena yrityksen tulisi tekoälyn hyödyntäjänä profiloitua ja viestiä? Mitä tekoälyn saralla tapahtuu seuraavan viiden vuoden sisällä? Entä luottamuksenarvoisuuden saralla?
TK3. Miten tekoälyä käytetään luottamuksenarvoisesti?	3.1 Tekoäly käsitteenä ja teknologioina 3.3 Luottamuksenarvoisuuden riskit ja haasteet 3.4 Luottamuksenarvoisuuden edellytykset 3.5 Luottamuksenarvoisuuden rakentaminen yrityksissä	5.2 Ulkoisten haastattelujen tulokset 5.3 Työntilaajan haastattelujen tulokset (pääosin salattu) 6 Tulkinta ja johtopäätökset Salatut liitteet	Mitä on tekoäly? Mistä koostuu luottamuksenarvoinen tekoäly? Mitkä ovat kolme suurinta haastetta tekoälyn käytössä luottamuksenarvoisuuden näkökulmasta? Miten edellä mainittuja haasteita tulisi ratkoa? Kohtaatko datan tai tekoälyn luottamuksenarvoisuuteen liittyviä kysymyksiä omassa työssäsi ja jos kyllä, niin millaisia? / Minkälaisia muita tekoälyn eettisyyteen, laillisuuteen, luotettavuuteen tai vastuullisuuteen liittyviä kysymyksiä on tullut tai voisi tulla vastaan? Mitkä ovat tämän hetken tärkeimmät toimet ja tahot luottamuksenarvoisen tekoälyn edistämisessä?

3 Tietoperusta

3.1 Tekoäly käsitteenä ja teknologioina

Tekoälystä puhutaan paljon, mutta sen määritelmä ei ole yksiselitteinen. Koska tekoälyä voidaan pitää jatkuvasti kehittyvänä joukkona teknologioita, on tärkeää ymmärtää, mitä tällä hetkellä voidaan perustellusti kutsua tekoälyksi ja mitä ei olla vielä kehitetty. Minusta tekoälyteknologiaa ja sen älyllistä tavoitetta kuvaavat parhaiten Euroopan komission (2020, 2) määritelmä, että tekoäly on *”joukko teknologioita, joissa yhdistetään dataa, algoritmeja ja laskentatehoa”* ja Merilehdon (2018, 18) määritelmä, että tekoäly on *”koneen suorittamaa toimintaa, joka ihmisen tekemänä olisi älykästä”.* Merilehto kuitenkin toteaa, että tekoälyä pyritään kehittämään niin, että se ei jää ihmisen älylliselle tasolle vaan kykenee itsenäiseen oppimiseen, päättelemiseen ja ennakoimiseen paremmin kuin ihminen (Merilehto 2018, 18).

Alun perin tekoälyä pyrittiin luomaan ihmisen ajattelua ja päättelykykyä vastaavaksi. Kun tällaisen *yleisen tai vahvan tekoälyn,* saati super tekoälyn, rakentaminen ei vielä ole onnistunut, on keskitytty niin sanottuun *kapeaan tai heikkoon tekoälyyn*, joka suoriutuu kapealaisesta ja sille määritellystä, yksittäisestä tehtävästä (kuva 1). *Vahvaa tekoälyä* pyritään kuitenkin jatkuvasti kehittämään, mutta se edellyttää *sekä* sitä, että tekoäly kykenee sekä *oppimaan itsenäisesti tai ohjaamattomasti* että *yleistämään oppimaansa*. Esimerkiksi neuroverkot oppivat jo itsenäisesti ja syvät neuroverkot puolestaan kykenevät osin jo yleistämään, mutta varsinaisen laajan, vahvan tekoälyn ajatellaan olevan vielä vuosikymmenien päässä. (Merilehto 2018, 18, 23–25.)



Kuva 1 Tekoälyn kehitysaskeleet (Ailisto, Heikkilä, Helaakoski, Neuvonen & Seppälä 2018, 53)

Filosofi Maija-Riitta Ollilan (2019, 8) mukaan heikosta tekoälystä voitaisiin useimmissa yhteyksissä puhua termillä ’koneoppiminen’. On kuitenkin hyvä ymmärtää, että kaikki tekoäly ei ole koneoppimista, eivätkä kaikki koneoppimisen alalajit sisällä tekoälyn ominaisuuksia, jos ajatellaan, että tekoälyn älykkyys ilmenee sen autonomisuutena, adaptiivisuutena, korkeana suorituskykynä ja itseoppivuutena. Käsitteiden selventämiseksi ja tässä työssä käytetyn näkemyksen esittämiseksi oleellisimmat tekoälyn osaamisalueet on syytä avata. Ailiston ym. (2018, 1) mukaan tekoäly voidaan jakaa kymmeneen tieteellisteknologiseen osaamisalueeseen, joista tässä tarkastellaan tarkemmin data-analyysiä, koneoppimista, luonnollisen kielen prosessointia ja havainnointia.

3.1.1 Data-analytiikka ja koneoppiminen

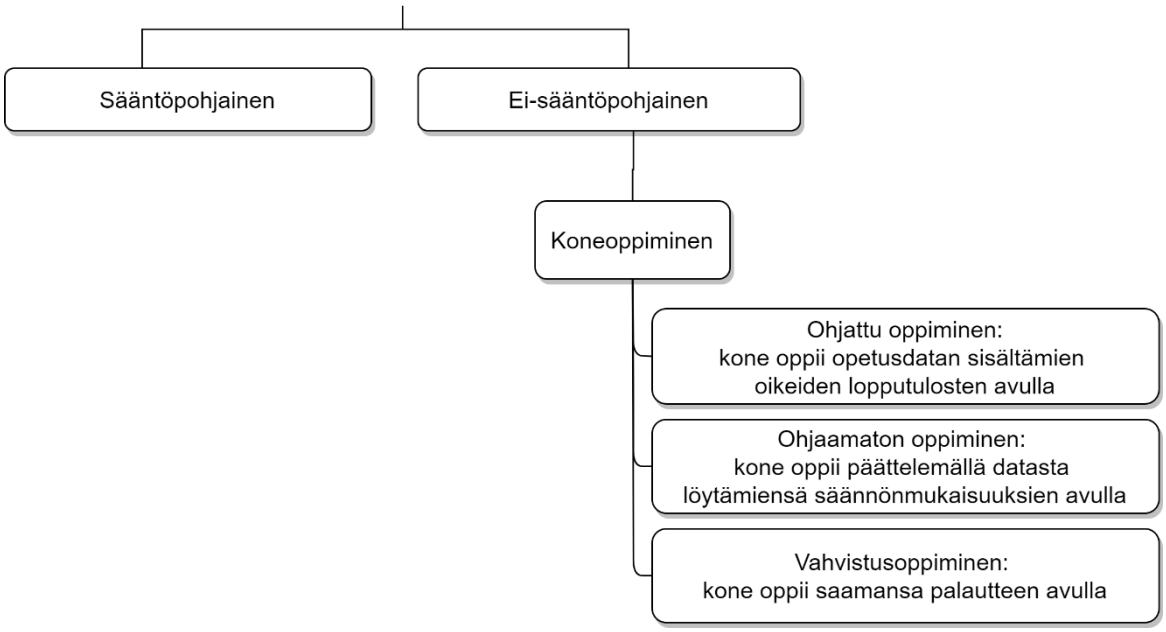
Data-analytiikan osaamispohjan luo datatiede. Data-analytiikalla tarkoitetaan nimensä mukaisesti datan analysoimista ja jalostamista tiedoksi ja yhä edelleen johtopäätöksiksi. Sen piiriin kuuluu koko ketju datan valmistelusta sen visualisointiin, analysointiin ja tulkintaan. Datan hallintaan ja esikäsittelyyn katsotaan kuuluvan datan yhdistely eri datalähteistä, sen muokkaus, suodatus, tarkistus, puhdistus ja annotointi sekä datan hankinta ja tallennus. (Ailisto ym. 2018, 8.)

Ailiston ym. mukaan (2018, 46) *”koneoppiminen (machine learning) on tietokonetekniikan osa-alue, jossa yleensä käytetään tilastotieteen menetelmiä, jotka antavat tietokoneille kyvyn oppia datasta (s.o. parantaa suorituskykyään tietyn tehtävän suorittamisessa) ilman eksplisiittistä ohjelmointia”.* Tässä työssä koneoppimista käsitellään Merilehdon (2018, 28-29) määritelmän mukaisesti tekoälyn osa-alueena, joka käyttää algoritmejä sekä datan kuvaamiseen, luokitteluun, klusterointiin että datasta oppimiseen ja ennustamiseen. Asian konkretisoimiseksi taulukossa 2 on havainnollistettu tyypillisiä koneoppimisen sovelluksia sekä tapaa, jolla koneoppimismallia usein opetetaan: *syöte* kuvastaa dataa, jota mallille syötetään ja *vaste* puolestaan haluttua lopputulosta. Mallia opetetaan jakamalla syötedata kahteen osaan niin, että ensimmäisellä datalla opetetaan mallia ja toisella datalla testataan, että malli toimii halutusti. Mallin opettamisesta puhuttaessa puhutaan käytännössä siis algoritmien sisäisten parametrien säätämisestä mahdollisimman hyvän lopputuloksen (esimerkiksi mahdollisimman tarkan ennusteen tai luokittelun) saavuttamiseksi. (Merilehto 2018, 27–29; Ailisto ym. 2018, 9, 14.)

Taulukko 2. Koneoppimisesimerkit (mukaillen Merilehto 2018, 29)

SYÖTE	VASTE	SOVELLUS
Ääninauhoite	Litteroitu teksti	Puheentunnistus
Historiallinen markkinadata	Tulevat kurssit	Treidausbotit
Valokuva	Kuvateksti	Kuvien merkintä
Luottokorttiosto	Petos vai ei?	Petosten esto
Ostohistoria	Tulevat ostot	Asiakaspito
Kuvia kasvoista	Nimiä	Henkilön tunnistaminen

Koneoppiminen jakautuu kuvan 2 mukaisesti ohjattuun, ohjaamattomaan tai vahvistettuun oppimiseen. Yksinkertaistettuna ohjattua oppimista käytetään, kun data on luokiteltua ja tiedetään tarkasti, mitä koneelta halutaan lopputuloksena. Ohjaamaton oppiminen puolestaan sopii tilanteeseen, jossa valmiita luokkia ei ole ja koneen on löydettävä datan avulla lopputulokseksi esimerkiksi erilaisia klustereita. Vahvistusoppimisella tarkoitetaan koneen saamaa negatiivista tai positiivista palautetta itsenäisestä toiminnastaan kompleksisessa ympäristössä, jossa ideana on palautteen avulla vahvistaa koneen pyrkimystä positiiviseen toimintaan vaihtuvissa tilanteissa. (Ollila 2019, 55–56.)



Kuva 2 Koneoppiminen (mukaillen PwC 2018, 3; Merilehto 2018, 19)

Erilaisia koneoppimisen malleja ovat esimerkiksi *luokittelu, ryhmittely (l. klusterointi), regressio, suosittelu* ja *poikkeaminen etsiminen.* Asiakaskannan ryhmittely on esimerkki ohjaamattomasta oppimisesta ja kohdennettu markkinointiluokittelu tai kannattavuuden ennustaminen regression avulla esimerkkejä puolestaan ohjatusta oppimisesta. Lisäksi mallien opettamisesta puhuttaessa voidaan puhua niin sanotuista online- ja offline-malleista sen mukaan, halutaanko mallin kehittyvän jatkuvasti ja löytävän uusia yhteyksiä uuden datan avulla vai pysyvän opetetun kaltaisena. (Merilehto 2018, 33–34.)

3.1.2 Syväoppiminen, koneaistit ja luonnollisen kielen käsittely

Neuroverkot ja syväoppiminen on yksi koneoppimisen osa-alue. Syväoppimisen uusi tuleminen ulottuu viimeiseen kymmeneen vuoteen ja perustuu laskentatehon kehitykseen, kasvavaan datan määrään ja sen myötä laajaan sovellettavuuteen. Neuroverkot koostuvat kerroksellisista, matemaattisista yksiköistä, neuroneista, jotka ovat yhteydessä toisiinsa ja jotka keskittyvät kukin havainnoimaan yhtä asiaa kerrallaan ja välittämään prosessoimaansa eteenpäin. Jokaisella kerroksella on myös oma tehtävänsä ja mitä enemmän neuroverkossa on kerroksia, sitä syvemmästä verkosta puhutaan, mistä myös juontuu termi *syväoppiminen*. Neuroverkot toimivat sitä paremmin, mitä enemmän niillä on opetusdataa käytettävissään. (Merilehto 2018, 20,45, 47, 48, 51, 55–57, 163.)

Mikäli tekoälyn halutaan tulevaisuudessa olevan tietoisempi ympäristöstään ja sen halutaan reagoivan omien havaintojensa pohjalta, on sen kyettävä havainnoimaan ympäristöään. Ailisto ym. on lähestynyt aihetta koneaistien kautta, joista yksi on konenäkö ja toinen esimerkiksi kyky ymmärtää puhetta. Yhtä lailla koneaistit voivat olla myös sellaisia aisteja, joita ihmisillä ei ole, kuten esimerkiksi tutkat tai paikannustavat. Konenäöllä tarkoitetaan *”menetelmiä, joilla pyritään edistämään kuvassa olevan tiedon automaattista irrottamista (extraction) ja kuvan sisällön ymmärtämistä”*. Kuvalla tässä kontekstissa voidaan tarkoittaa yhtä lailla harmaasävykuvia, värikuvia, 3D-kuvia kuin ultra- tai röntgenkuviakin eli mitä tahansa digitaalisessa muodossa olevia kuvia tai kuvasarjoja. Konenäön osa-alueet koostuvat koko ketjusta: kuvan muodostamisesta kuvan käsittelyn ja analyysin kautta kuvan sisällön ymmärtämiseen. Koneaistien ja objektintunnistuksen sovelluskohteita ovat esimerkiksi asuntojen, autojen, puhelimien ja sosiaalisen median kasvojentunnistussovellukset. (Ailisto ym. 2018, 10–11.)

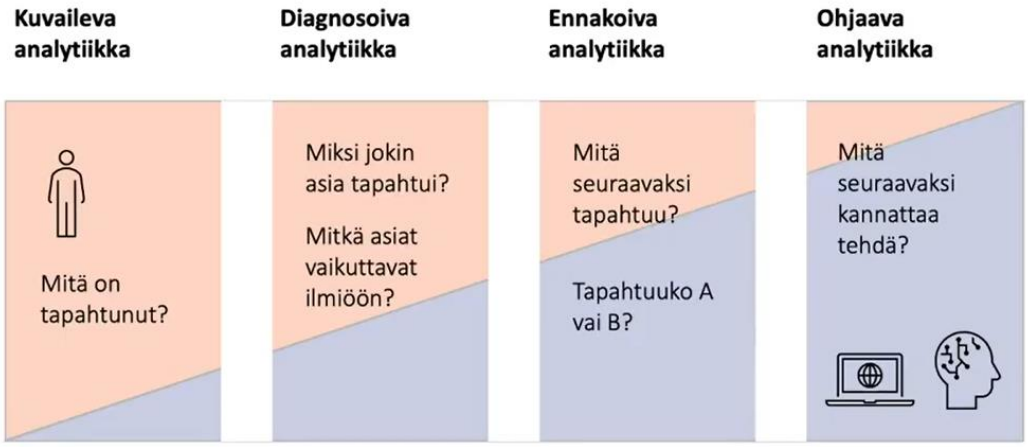
Luonnollisen kielen käsittelyllä (Natural Language Processing, NLP) tarkoitetaan tietokoneen suorittamaa kirjoitetun ja puhutun kielen analysointia ja tuottamista, kuten edellä mainittua kykyä ymmärtää puhetta. Käytännön sovelluksia on jo useita: esimerkiksi puheluita ja saneluita tekstiksi kääntävät sovellukset, käännösohjelmat sekä asiakasrajapinnan ja tiedonhaun virtuaaliassistentit sekä juuri Euroopan komission syyniin joutuneet (kts. lisää

luku 3.3.2) puheentunnistukseen perustuvat digitaaliset assistentit, kuten Applen Siri, Googlen Assistant ja Amazonin Alexa. (Ailisto ym. 2018, 11; Pervilä 2020; Euroopan komissio 2020b.)

3.1.3 Alakohtaiset teknologiasovellutukset etiikan kannalta

Yksi syy, miksi tekoälyn etiikkaa tutkittaessa tulisi ymmärtää tekoäly joukkona teknologioita ja erottaa toisistaan esimerkiksi edellä kuvattuja koneoppimisen erilaisia oppimisparadigmoja, liittyy siihen, että tekoälyltä vaadittavaa "älykkyyttä" rakennetaan erilaisiin tekoälyratkaisuihin eri tavoin. Ihmisen rooli älykkyyden luomisessa eri ratkaisuihin on erilainen, jolloin myös eettiset kysymykset voivat poiketa toisistaan. Ohjatussa oppimisessa koneen oppima älykkyys tulee ihmiseltä; ihminen on koostanut tekoälylle oikeat vastaukset. Ohjaamattomassa oppimisessa ihminen on yhtä lailla valmistellut datan, mutta sillä ei ole antaa koneelle oikeita vastauksia, vaan ihmisen rooliksi ja ihmisen vastuulle jää tekoälyn tekemien päätelmien, esimerkiksi klusterien, tulkitseminen. (Lehtimäki 29.10.2020.)

Jo data-analytiikka muuttaa ihmisten roolia sekä ihmisen ja koneen välistä suhdetta. Kuvailevassa analytiikassa ihmisen rooli on suuri, mutta se pienenee, kun mennään kohti ohjaavaa analytiikkaa (kuva 3). Ihminen ei pärjää enää koneelle ennakoivassa analytiikassa ja ohjaavassa analytiikassa ihmisen rooli on enää pieni, vaikka ihminen on opettanut algoritmin ja valmistellut datan. Algoritmi ei enää välitä tietoa ihmiselle vaan pyrkii päättelemään itse. Datahavainnosta on päästy tilanteen optimointiin. (Lehtimäki 29.10.2020.)



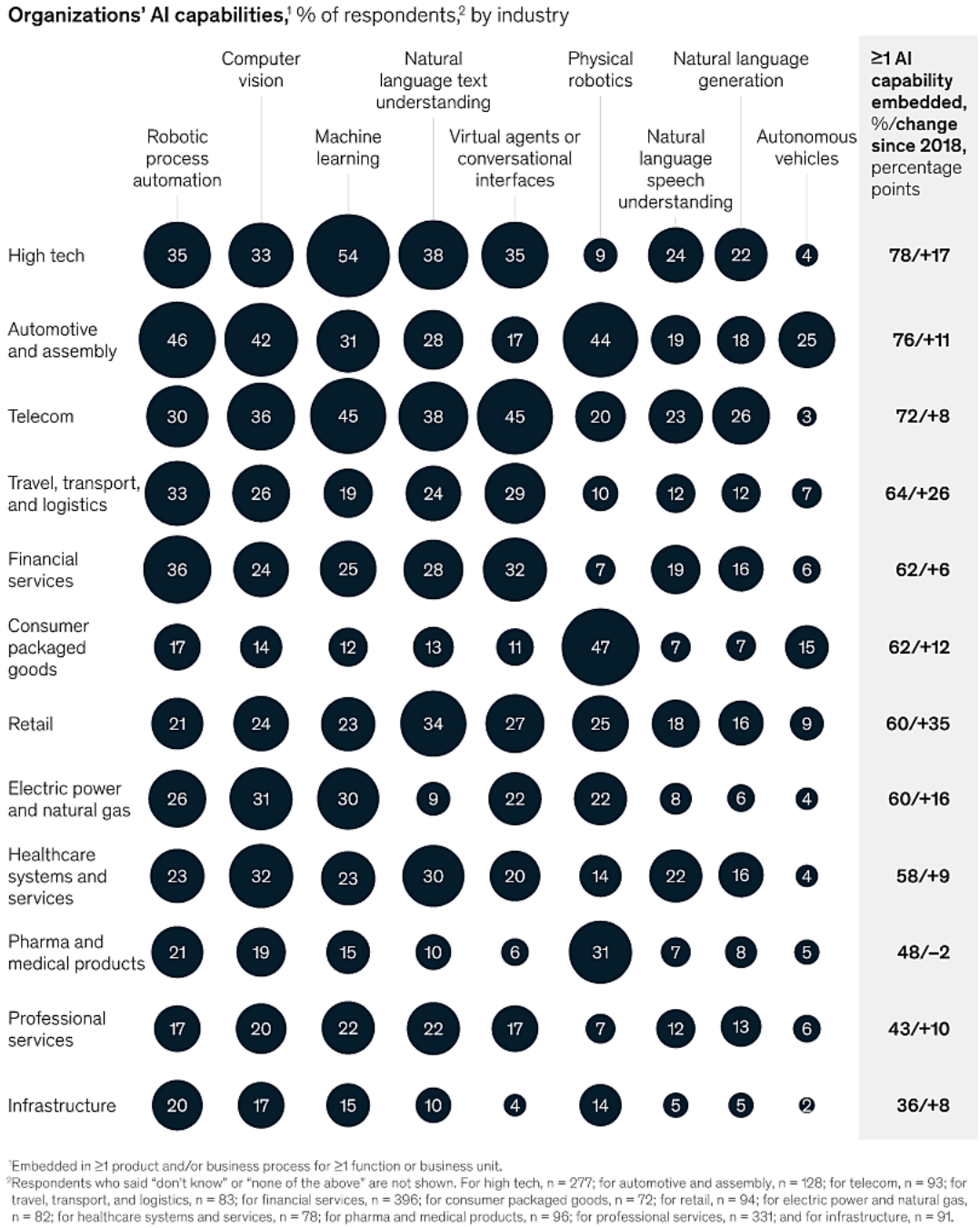
Kuva 3 Ihmisen ja koneen suhde data-analytiikassa (Lehtimäki 29.10.2020)

McKinseyn tutkimuksen mukaan (taulukko 3) eri alat käyttävät yleensä arvoketjunsa kannalta olennaisimpia teknologioita. Siinä missä suuret teknologiajätit ja teleoperaattorit ovat hyödyntäneet tekoälyteknologioista pitkälti koneoppimista, on finanssiala tutkimuksen mukaan keskittynyt tekoälyä sisältävien ohjelmistorobottien ja virtuaaliassistenttien luomiseen.

Terveyspalvelut ovat olleet kiinnostuneita konenäön ja luonnollisen tekstin ymmärtämisen luomista mahdollisuuksista, kun taas lääketeollisuus on ollut kiinnostuneempi tekoälyä hyödyntävistä fyysisistä roboteista. (McKinsey 2019.)

Käytännössä tekoälyä on siis selvästi lähestytty oman alan haasteiden ja mahdollisuuksien kautta, mutta toisaalta erot alojen välillä vaikuttavat melko suurilta. Tekoälysovelluksissa voidaan hyödyntää toki myös useita eri teknologioita samassa yhteydessä (Ailisto ym. 2018, 24).

Taulukko 3. Organisaatioiden AI-kyvykkyydet (McKinsey 2019)



3.2 Työeläketoimialan ominaispiirteet

Mietittäessä tekoälyn luottamuksenarvoisuuden rakentamista työeläketoimialalla on ensin ymmärrettävä, millainen ala on juridiselta ja liiketoiminnalliselta pohjaltaan. Työeläkelainsäädäntö on peräisin vuodelta 1961. Tuolloin työnantajille tuli pakolliseksi maksaa työntekijöidensä palkanmaksun yhteydessä myös työeläkevakuutusmaksu valitsemalleen yksityiselle työeläkeyhtiölle. Vaihtoehtona oli, ja on yhä, perustaa oma työeläkekassa tai työeläkesäätiö, mikä osaltaan luo tehokkuutta ja asiakaspalvelun laatua edistävää kilpailua julkista tehtävää hoitavien yksityisoikeudellisten eläkelaitosten välillä. Laitokset eivät nimittäin voi kilpailla hinnoittelulla ja lakisääteisten etuuksien ominaisuuksilla, mutta ne voivat kilpailla palveluidensa laadun, sijoitustoiminnan tuoton ja hoitokustannusten avulla, eli asiakkaille voidaan maksaa suurempia asiakashyvityksiä, jotka määräytyvät toiminnan tehokkuuden ja sijoitustoiminnan tuoton perusteella. (Sosiaali- ja terveysministeriö 2019, 19; Halila 2005, 1, 9.)

Toinen työeläketoimialaa leimaava tekijä on sidosryhmien ja alaa valvovien tai ohjaavien tahojen moninaisuus. Julkista tehtävää hoitavina ja julkista valtaa käyttävinä työeläkelaitokset ovat eduskunnan oikeusasiamiehen valvontavallan alaisuudessa. Lisäksi alaa sitovat varsinaisen työeläkelainsäädännön lisäksi useat eri lait, Finanssivalvonnan määräykset ja sosiaali- ja terveysministeriön asetukset. Sosiaali- ja terveysministeriö vahvistaa vuosittain muun muassa monimutkaisen vakuutusmaksun laskuperusteen korkoineen. Valmistelussa ovat kuitenkin mukana niin työeläkelaitokset kuin työmarkkinain keskusjärjestötkin. (Sosiaali- ja terveysministeriö s.a.; Eläketurvakeskus s.a.; Halila 2005, 6, 11, 12.)

3.2.1 Työeläketoimialan lainsäädäntö tekoälyn näkökulmasta

Kysymys työeläketoimialan tekoälyn käyttöä ohjaavista laeista kulminoituu kysymykseen työeläkelaitosten yksityis- ja julkisoikeudellisista tehtävistä. Tämä kysymys puolestaan liittyy Halilan mukaan (2005, 2) isompaan kysymykseen siitä, mitä pidetään julkis- ja mitä yksityisoikeudellisena toimintana ja mitä työeläkelaitosten toiminnassa voidaan pitää julkisen hallintotehtävän hoitamisena. Työeläkelaitoksia ei voida pitää viranomaisina, mutta kaikki työeläkelaitokset hoitavat lakisääteistä ja julkista tehtävää osana sosiaaliturvaa ja käyttävät sitä kautta julkista valtaa, mutta ovat silti yksityisoikeudellisia toimijoita.

Toinen juridinen kysymys liittyy teknologioiden, myös tekoälyn, hyödyntämiseen hallintopäätöksissä. Tekoäly kykenee tekemään yhä enemmän sellaisia toimintoja, jotka ihminen on aiemmin tehnyt, mukaan lukien erilaisia yksilöitä koskevia päätöksiä. Tekoälyn, erityi-

sesti koneoppimisen, yksi merkittävimmistä käyttökohteista onkin automaattinen päätöksenteko, jolla päätöksentekoprosessia voidaan tehostaa ja päätösten yhdenvertaisuutta parantaa (Koskinen 2017, 34, 61, 79).

Opinnäytetyön kirjoitushetkellä automaattista päätöksentekoa koskeva laki on valmisteilla. Kiireellinen lainsäädännön arvioiminen liittyy muun muassa oikeuskanslerin ja apulaisoikeusasiamiehen selvityksessä viime vuosina olleisiin, Maahanmuuttoviraston, Kelan ja Veron, päätöksentekomenettelyihin, joista on myös uutisoitu laajalti. Ongelmallista siis on, että jo käytössä oleva automaattinen päätöksenteko ja sen osittainen vakiintuneisuus ei välttämättä ole nykylainsäädännön puitteissa laillista (taulukko 4 seuraavalla sivulla).

Lainsäädännön esiselvitystä varten viranomaiset – mutta myös Eläketurvakeskus ja työeläkeyhtiöt – vastasivat tietopyyntöön automaattisen päätöksenteon asteesta, määristä, kohteista ja edellytyksistä. Nykytilakartoituksen mukaan Suomessa tehdään jo päätöksiä täysin automaattisesti, mutta nämä päätökset eivät edellytä asianomaisen kuulemista ja perustuvat lakiin, joskaan *”yhteenvedossa ei ole arvioitu viranomaisten vastauksista ilmenevien päätöksentekokäytäntöjen suhdetta hallintolakiin tai perustuslakiin”* (Vainio ym. 2020, 15). Automatisoidut päätökset ovat hakijalle usein myönteisiä*:*

> ”**Eläketurvakeskuksella** edellytyksenä automaattiselle päätöksenteolle on asian ratkaiseminen hakijan kannalta myönteisesti ja siten, että päätöksen perusteiksi otettavat tiedot ovat rekisteristä ja suoraan asianosaiselta jo sähköisesti saatujen tietojen perusteella selvillä” (Vainio ym. 2020, 16).

Esiselvityksessä kävi kuitenkin ilmi, etteivät edellytykset automaattiselle päätöksenteolle ole yhtenevät vaan eri toimijat soveltavat niihin omia ehtojaan. Lainsäädännön arviomuistiossa tähdennetään myös, että automaattinen päätöksenteko tulisi rajata tilanteisiin, jotka eivät edellytä ihmisen harkintaa eivätkä asianomaisen kuulemista ja että päättelysäännöt voivat perustua vain lakiin. Lisäksi muistiossa todetaan, että lainsäädäntöä on tarkennettava virkavastuun ja päätöksentekoprosessin läpinäkyvyyden osalta. (Vainio ym. 2020, 15–16, 55, 63–65.)

Lainvalmistelussa on ehditty siihen vaiheeseen, että *Arviomuistio hallinnon automaattiseen päätöksentekoon liittyvistä yleislainsäädännön sääntelytarpeista* on kommentointikierroksella. Siinä arvioidaan, miten hyvin ihmisten perusoikeudet, oikeusturva, julkisuusperiaate, hyvä hallintotapa, hallinnon lainalaisuus ja virkavastuu toteutuvat Euroopan unionin tietosuojasääntelyssä ja Suomen perustuslaissa. Muistiossa ehdotetaan, että automaattiset, hallinnolliset päätökset on rajattava vain sääntöpohjaisiin päätöksiin, jotka eivät edellytä lainkaan harkintaa. (Vainio ym. 2020, 10.)

Taulukko 4. Perustuslakivaliokunnan lausunnot lakiesityksiin, joihin sisältyi automaattiset yksittäispäätökset mahdollistava säännösehdotus (Vainio, Tarkka & Jaatinen 2020, 27)

HE	Säännösehdotus	PeVL
HE 224/2018 vp maahanmuuton henkilötietolaki	Päätös, jonka Maahanmuuttovirasto tekee ulkomaalaisasioiden asiankäsittelyjärjestelmässä, voidaan tehdä menettelyssä, joka perustuu pelkästään automaattiseen käsittelyyn, **ellei asian laatu tai laajuus, yhdenmukainen kohtelu, lapsen etu taikka muu erityinen syy muuta edellytä**. Rekisteröidyllä on oikeus saada selvitys hänelle automaattisessa käsittelyssä tehdystä yksittäispäätöksestä.	PeVL 62/2018 vp: Perustuslakivaliokunnan mielestä sääntelyä on kuitenkin välttämätöntä täsmentää. Säännöksessä on säänneltävä ehdotettua täsmällisemmin, millaisin perustein asioita voidaan valikoida automaattisen päätöksenteon piiriin.
HE 298/2018 vp potilasvakuutuslaki	Potilasvakuutuskeskuksella on tämän lain toimeenpanossa oikeus tehdä luonnollisten henkilöiden suojelusta henkilötietojen käsittelyssä sekä näiden tietojen vapaasta liikkuvuudesta ja direktiivin 95/46/EY kumoamisesta annetun Euroopan parlamentin ja neuvoston asetuksen (EU) N:o 2016/679 (yleinen tietosuoja-asetus), jäljempänä tietosuoja-asetus, 22 artiklan 1 kohdan mukaisia automaattisia päätöksiä, **jos automaattisen päätöksen antaminen on käsiteltävänä olevan asian laatu ja laajuus sekä tämän lain ja hyvän hallinnon vaatimukset huomioon ottaen mahdollista**.	PeVL 70/2018 vp: Laissa ei kuitenkaan säädettäisi yksityiskohtaisesti siitä, missä tilanteissa tai mihin asiaryhmiin liittyviä päätöksiä olisi mahdollista käsitellä täysin automaattisesti. Perustuslakivaliokunnan mielestä sääntelyä on välttämätöntä täsmentää. Säännöksessä on säänneltävä ehdotettua täsmällisemmin, millaisin perustein asioita voidaan valikoida automaattisen päätöksenteon piiriin.
HE 52/2018 vp sosiaaliturva- ja vakuutuslainsäädäntö	Kansaneläkelaitos voi sen toimeenpanemien etuuksien toimeenpanossa tehdä luonnollisten henkilöiden suojelusta henkilötietojen käsittelyssä sekä näiden tietojen vapaasta liikkuvuudesta ja direktiivin 95/46/EY kumoamisesta annetun Euroopan parlamentin ja neuvoston asetuksen (EU) N:o 2016/679 (yleinen tietosuoja-asetus), jäljempänä tietosuoja-asetus, 22 artiklan 1 kohdassa tarkoitettuja automaattisia päätöksiä, **jos automaattisen päätöksen antaminen on käsiteltävänä olevan asian laatu ja laajuus ja hyvän hallinnon vaatimukset huomioon ottaen mahdollista**.	PeVL 78/2018 vp: Laissa ei kuitenkaan säädettäisi yksityiskohtaisesti siitä, missä tilanteissa tai mihin asiaryhmiin liittyviä päätöksiä olisi mahdollista käsitellä täysin automaattisesti. Perustuslakivaliokunnan mielestä sääntelyä on välttämätöntä täsmentää. Automaattista päätöksentekoa koskevassa sääntelyssä on säänneltävä ehdotettua täsmällisemmin, millaisin perustein asioita voidaan valikoida automaattisen päätöksenteon piiriin.
HE 18/2019 vp maahanmuuton henkilötietolaki	Päätös, jonka Maahanmuuttovirasto tekee ulkomaalaisasioiden asiankäsittelyjärjestelmässä, voidaan tehdä menettelyssä, joka perustuu pelkästään automaattiseen käsittelyyn, **jos asian saa ratkaista hallintolain (434/2003) 34 §:n 2 momentin 5 kohdan nojalla asianosaista kuulematta**. Maahanmuuttoviraston on julkistettava automatisoidussa päätöksentekomenettelyssä lopulliseen päätökseen johtanut algoritmi. Lisäksi rekisteröidyllä on oikeus saada erillinen selvitys hänelle automatisoidussa käsittelyssä tehtyyn lopulliseen yksittäispäätökseen käytetystä algoritmista. Maahanmuuttoviraston ylijohtaja vastaa automaattista päätöksentekoa koskevasta menettelystä ja automatisoidusta yksittäispäätöksestä.	PeVL 7/2019 vp: Valiokunnan mielestä ehdotettua sääntelyä on edellä sanotun johdosta rajattava siten, että automatisoitua päätöksentekoa ei voi käyttää sillä perusteella, että kuuleminen asiassa olisi muusta syystä ilmeisen tarpeetonta. Automatisoitu päätöksenteko on käsillä olevassa sääntelykontekstissa rajattava vain sellaiseen päätöksentekoon, jossa hyväksytään vaatimus, joka ei koske toista asianosaista. Vaihtoehtoisesti sääntelyä on olennaisesti täsmennettävä määrittelemällä lakiehdotuksen 21 §:ssä ne hallintolain 34 §:n 2 momentin 5 kohdan jälkimmäisen lauseen alaan kuuluvat tilanteet, joissa päätöksenteko ei edellytetä viranomaisen käyttävän harkintavaltaa.

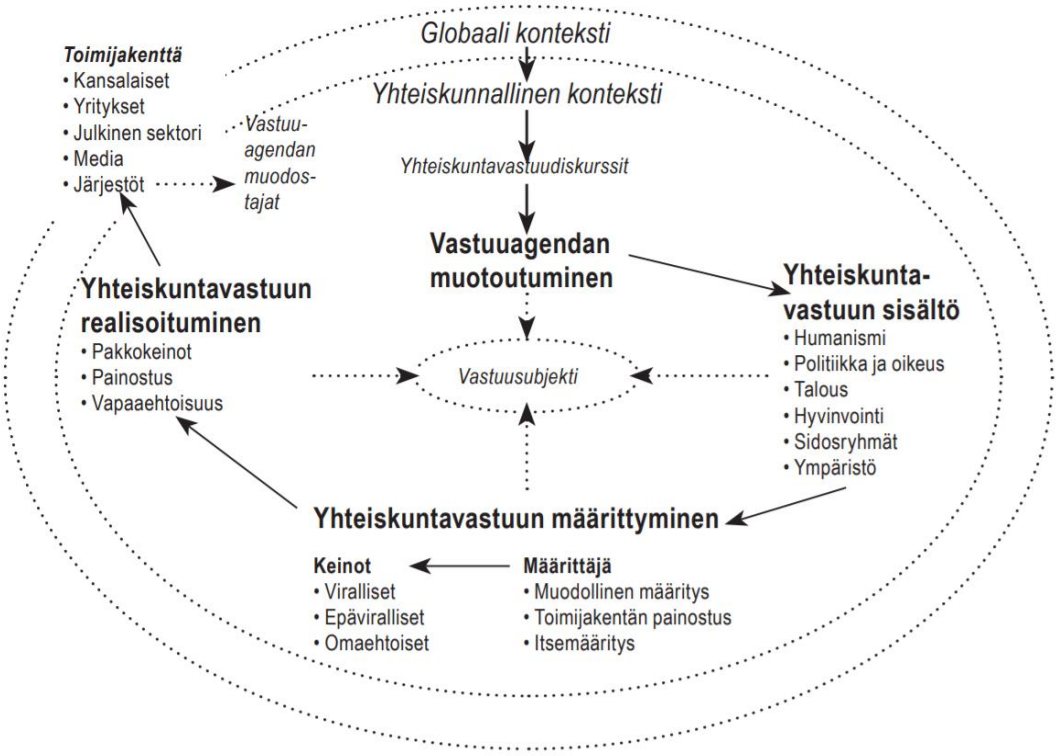
Finanssiala ry on kuitenkin ottanut kantaa kesken olevaan automaattisen päätöksenteon lainvalmisteluun toteamalla, että esiselvityksen näkökulma on liian suppea vakuutus- ja eläkeyhtiöiden näkökulmasta eikä se huomioi riittävästi tekoälyn mahdollisuuksia: *"Niin sanottujen rutiinipäätösten hoitamista vakuutus- ja eläkeyhtiöissä nopeuttaisi, jos hallinnon automaattista päätöksentekoa koskeva laki säädettäisiin mahdollisimman pian ja siinä mahdollistettaisiin myös tekoälyn käyttäminen"* (Wennberg 2020). Tähän liittyen on kuitenkin ymmärrettävä, että työeläkeyhtiöt poikkeavat vahinkovakuutusyhtiöistä vielä siinä, että niiden kaikki päätökset perustuvat lakiin eikä työeläkeyhtiöissä ole yksityisoikeudellisia vakuutustuotteita.

On siis epäselvää, missä määrin automaation, mukaan lukien tekoälyn, käyttöä työeläketoimialalla ollaan rajoittamassa tai ohjaamassa juridisin keinoin. Tällä hetkellä on tulkittava, että kaikki harkintaa edellyttävät hallintopäätökset on tehtävä manuaalisesti – ei siis vielä edes automaattisesti, saati tekoälyä hyödyntämällä autonomisesti. Avusteista automaatiota ja tekoälyä on kuitenkin pidettävä sallittuna esimerkiksi ratkaisukäytäntöä varmistavassa roolissa. Laki toteutuessaan rajaisi silti tekoälyn käyttökohteita merkittävästi työeläketoimialalla ja on alalle siten erityistä, ominaista ja huomioonotettavaa.

3.2.2 Yhteiskuntavastuullisuus työeläketoimialalla

Lakisääteisen ydintehtävänsä lisäksi koko työeläketoimialan yhteisiä yhteiskunnallisia tehtäviä ja haasteita ovat työeläkejärjestelmän kestävyys, suomalaisten työhyvinvoinnin tukeminen sekä eläkemaksujen ja -menojen sopeuttaminen talouteen (Sorsa 2006, 82). Työeläkeyhtiöitä ei siis koske vain yritysten yhteiskuntavastuullisuus, vaan myös laajempi poliittinen ja juridinen vastuu, koska työeläkeyhtiöt on nähtävä yhteiskunnallisesti merkittävinä sosiaaliturvalaitoksina, joiden toiminnassa on vahva paradigma. Järjestelmän sisältö määritellään julkisesti ja järjestelmää hallinnoidaan julkisesti, mutta sitä toimeenpannaan yhteisvastuullisesti ja kilpailurajoittein yksityisissä yrityksissä, joita säädellään yksityisinä vakuutusyhtiöinä. (Sorsa 2006, 76–77, 83.)

Sorsa (2006) on tutkinut, miten yhteiskuntavastuun kautta voidaan analysoida työeläketoimialan sijoitustoiminnan legitimiteettiä ja tuo siinä kontekstissa esiin Anttiroikon (2004, 42) yhteiskuntavastuun prosessit globaalissa kontekstissa (kuva 4). Kauston mukaan (Sorsa 2006, 5) legitimiteetti työeläketoimialalla on ymmärrettävä luottamuksena koko työeläkejärjestelmään, sen laillisuuteen, mutta myös rakenteen antaman turvan uskottavuuteen ja järjestelmän sisältöön.



Kuva 4 Yhteiskuntavastuun prosessi (Sorsa 2006, 79, primäärilähde Anttiroiko 2004, 42)

Vastuuagenda muotoutuu Sorsan mukaan ensinnäkin yhteiskuntavastuun sisällöstä, mutta myös vastuun määrittymisestä. Määrittäjät työeläketoimialalla ovat EU ja valtio sekä työmarkkinajärjestöt, TELA ja työeläkelaitokset, joiden keinot vaihtelevat virallisista linjauksista, epävirallisista vastuista ja suosituksista yritysten omaehtoisiin linjanvetoihin. Yhteiskuntavastuun toteutumisen varmistavat muun muassa velvoittavasti lainsäädäntö, tuomioistuinten päätökset, viranomaisten pakkokeinot ja sopimukset mutta myös viranomaisten suositukset, yhteisölliset toimintaperiaatteet, yritysten omat toimintaperiaatteet, julkinen mielipide, julkisuus, kansalaistoiminta ja kulutuskäyttäytyminen. Sorsan mukaan on ymmärrettävä, että eri toimijoiden välillä on jännitteitä ja yhteiskuntavastuun määrittymisessä esille tulevat valtasuhteet. (Sorsa 2006, 78, 85–86.)

Vastuullisuus edellyttää siis lain ja hyväksyttyjen toimintaperiaatteiden noudattamista, mutta myös *”totuudellisuus ja rehellisyys sekä riittävä avoimuus, tiedonkulku ja läpinäkyvyys – ovat realisoitumisen ehtona”.* Mikäli näin ei toimita, voi yritykselle Sorsan mukaan langeta sanktioita tai korvausvaateita, mutta yhtä tärkeää on ennakoida ja ehkäistä rikkomuksia. Yhteiskuntavastuun ei tulisi olla maineenhallintaa ja vastuusta raportoimista vaan ajattelua ja toimintaa ohjaavaa läpi organisaation. Mikäli näin ei ole, on koko yhteiskuntavastuun prosessi pahimmillaan merkityksetön. (Sorsa 2006, 86–87.)

3.2.3 Tekoäly työeläketoimialalla

Digi- ja väestötietoviraston pääjohtaja Viskari puhui Finanssivalvonnan vuosiseminaarissa 2019 datan ja tekoälyn vastuullisesta hyödyntämisestä erityisesti julkishallinnossa, mutta puheenvuoro sopii sellaisenaan mielestäni myös työeläketoimialalle.

Viskari korosti, että varsinkin julkisen hallinnon puolella tekoälyllä tavoitellaan kahta asiaa: parempia ja sujuvampia palveluita ja toiminnan tehostamista. Väestön ikääntyessä ja syntyvyyden laskiessa julkishallinnon kulut nousevat eikä samoja palveluita pystytä tuottamaan entisin tavoin ja kustannuksin. Tieto ei kuitenkaan liiku riittävän hyvin eri toimijoiden välillä ja vaikka dataa olisi kattavasti, sijaitsee se siiloissa. (Viskari 26.11.2019.)

Teknologian kehittymisen suhteen Viskari peräänkuulutti pitkäjänteisiä toimia ja sopivankokoisten hankkeiden läpivientiä sen sijaan, että *”juostaan hype-asioiden perässä”* tai että toisaalta tehtäisiin vuosia kestäviä hankkeita. Lainsäädännön suhteen tekoälyhankkeiden ajoitus on myös tärkeää, mutta vaikeaa, koska dataan ja tekoälyyn liittyviä sovelluksia tehdään koko ajan ilman lainsäädännöllistä tukea. Hän toivoi, että lainsäädäntö etenee niin, että juridinen kehys on valmis ennen kuin tekoälysovellukset ovat liian pitkällä, jotta toimijoiden ei tarvitsisi jälkikäteen tehdä niihin radikaaleja muutoksia, jos lainsäätäjä onkin reunaehdoista eri mieltä. (Viskari 26.11.2019.)

Samalla Viskari totesi, että kansalaisten luottamus yhteiskuntaan on korkea eikä sitä saa vaarantaa. Hän piti tärkeänä, että tekoälyn käyttö on niin läpinäkyvää, että kansalainen voi myös itse arvioida suhtautumistaan siihen. Huomioon tulisi ottaa myös kansalaisten mahdollinen huoli sähköisen asioinnin kasvusta suhteessa muihin asiointikanaviin ja omaan osaamiseen, jolloin puhutaan jo kansalaisten yhdenvertaisuudesta ja perusoikeuksista. Huomioon tulee ottaa toisaalta myös kansalaisten kasvavat odotukset ja erilaiset tottumukset. (Viskari 26.11.2019.)

Hyvin samoilla linjoilla olivat tekoälyetiikan asiantuntija Haataja ja viestintäoikeuden professori Korpisaari vuoden 2019 Työeläkepäivillä. He toivat esiin yritysten vastuun suojella arvojen, etiikan ja lakien avulla perusarvoja ja ihmisoikeuksia myös uusia teknologioita sisältävässä digitaalisessa maailmassa. Heidän mukaansa yritysten tulee varmistaa, että tekoälytuotteet ja -palvelut ovat teknisesti luotettavia, hyvän hallintotavan mukaisia ja laillisia. Yritysten tulee kantaa vastuu tekoälyratkaisuista läpi niiden elinkaaren riskit ja vaikutukset huomioiden ja tietää, mikä prosessissa on ihmisen tehtävää. (Iivonen 2019, 11.)

3.2.4 Tekoälyn käyttökohteet

Julkisuudessa ei juuri ole esitelty työeläketoimialan tekoälyratkaisuja, mutta alalta löytyy kuitenkin muutama esimerkki. Eläketurvakeskus kokeili vuonna 2018, voisiko koneoppimisen avulla ennustaa työkyvyttömyysriskiä. Kokeilussa tutkittiin neljännesmiljoonan suomalaisen työkyvyttömäksi joutuneen anonymisoituja, sosioekonomisia tietoja ajalta ennen hyväksyttyä työkyvyttömyyseläkepäätöstä ja verrattiin niitä yhtä suureen joukkoon ihmisiä, jotka eivät olleet joutuneet työkyvyttömiksi. Koneoppiva malli pääsi 80%:n todennäköisyyteen. ETK:n mukaan mallia saisi tarkennettua esimerkiksi terveystiedoilla, mutta terveystietojen käyttöön algoritmin opettamisessa tarvittaisiin erikseen lupa. Työkyvyttömyysriskitapauksia voisi kuitenkin lopputuloksen avulla tunnistaa, jolloin työkyvyttömyyseläkkeelle jäämisen riskiä voisi ennaltaehkäistä, millä voisi olla kansanterveydellisiä vaikutuksia. Eläketurvakeskus on tuonut esiin myös, että tekoäly voisi olla avuksi eläkeneuvonnassa tai Suomen sosiaaliturvaan kuulumista todistavan A1-todistuksen hakemisessa, jotka molemmat ovat ETK:n palveluja. Tekoälyä voisi keskuksen mukaan tulevaisuudessa hyödyntää myös viranomaisten välisissä chatboteissa. (Karkiainen 2018; Rissanen 2018; Varis 2018.)

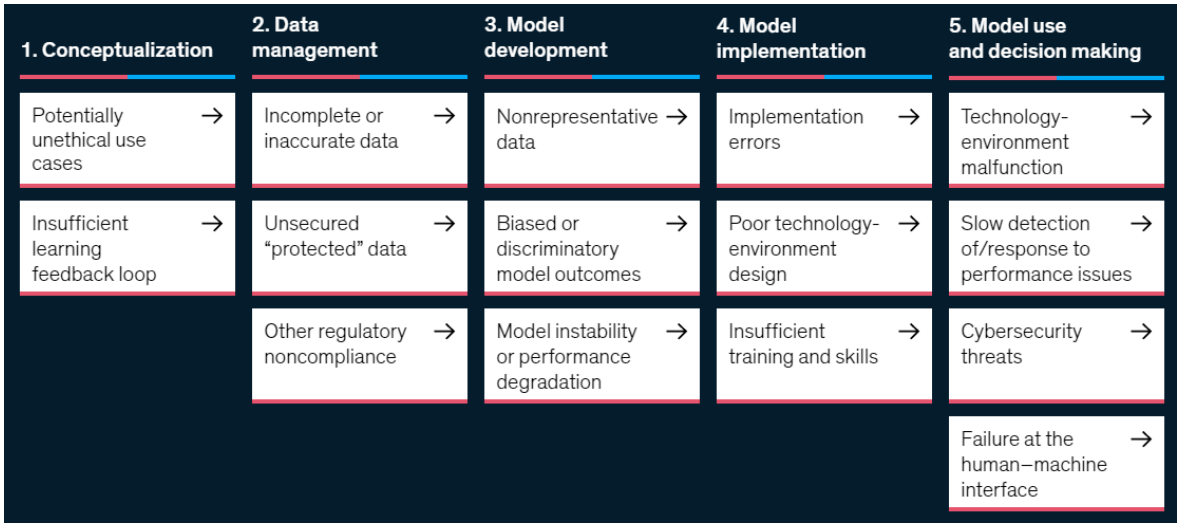
Myös Keskinäinen työeläkeyhtiö Varma on tuonut esiin tutkineensa luonnollisen kielen käsittelyn avulla työkyvyttömyyden syitä ja saaneensa selville, että alle 45-vuotiaista mielenterveyden häiriöt ja niihin liittyvä työkyvyttömyys tai työkyvyttömyyden uhka näkyy erityisesti hoito- ja myyntityössä. Varma toteaa myös, että sijoitustoiminnassa tekoälyä voisi hyödyntää yritysten tilinpäätöstietoihin ja osakekursseihin perustuvassa osakepoiminnassa. Myös muut eläkeyhtiöt ovat raportoineet tekoälyhankkeistaan: Keskinäinen Työeläkevakuutusyhtiö Elo on tuonut esiin 2018 käynnistetyn asiakasdatahankkeensa, jossa dataa pyrittiin keskittämään yhteen paikkaan ja siten myös mahdollistamaan tekoälymallien vienti tuotantoon asti. Hankkeella pyrittiin niin sanottuun asiakas 360 -näkymään eli parempaan asiakaskokemukseen, asiakashallintaan ja myyntiin, mutta myös esittelemään liiketoiminnallisia mahdollisuuksia uusien käyttökohteiden oivalluttamiseksi. Keskinäinen Eläkevakuutusyhtiö Ilmarinen hyödyntää vuoden 2018 yritysvastuuraportin mukaan tekoälyä muun muassa kuntoutushakemustensa rutiinitarkastuksissa ja myös Finanssivalvonta käyttää tekoälyä omassa valvonnassaan sijoituspalveluyritys- ja rahastonotifikaatioiden käsittelyssä, jossa tekoälyä tarvitaan muun muassa dokumenttien luokitteluun ja vapaamuotoisen tekstin tunnistukseen. (Dain Studios s.a.; Ilmarinen 2018, 13; Knowit 2018; Varma 2019; Vatanen 2017.)

3.3 Luottamuksenarvoisuuden riskit ja haasteet

Vaikka tekoäly mahdollistaa paljon, sen kehityksen ja käytön eri vaiheisiin liittyy paljon kysymyksiä useastakin näkökulmasta, joista merkittävimmät liittyvät Euroopan komission

(2020, 10) ja Ailiston ym. mukaan (2018, 40) yksityisyyteen, syrjimättömyyteen, turvallisuuteen, vastuuseen, hallittavuuteen ja läpinäkyvyyteen eli pitkälti ihmisten perusoikeuksien kunnioittamiseen.

Cheatham ym. (2019) ovat koonneet McKinseylle eri tekoälyprosessin vaiheiden merkittävimmät riskit (kuva 5). He jakoivat tekoälykehityksen viiteen eri vaiheeseen: konseptualisointiin, datanhallintaan, mallin kehitykseen, mallin toteutukseen ja mallin käyttöön ja päätöksentekoon. Ensimmäisen vaiheen riskit liittyvät epäeettiseen käyttökohteeseen ja riittämättömään palautesykliin. Datanhallinnan ja mallin kehityksen haasteet liittyvät puutteelliseen, epätarkkaan, suojaamattomaan ja ei-kattavaan dataan sekä vinoutuneisiin tai syrjiviin mallin lopputuloksiin, mallin epävakauteen tai toimintahäiriöihin ja muuhun säännöstenvastaisuuteen. Mallin implementoinnin, käytön ja päätöksenteon osalta riskit liittyvät kehittäjien kykyihin, ympäristöön, jossa mallia on tarkoitus käyttää, käyttöönotossa tehtyihin virheisiin, hitaaseen reagoimiseen, mikäli mallin toiminnassa havaitaan häiriöitä, kyberturvallisuusuhkiin ja ihmisen ja koneen vuorovaikutukseen. (Cheatham ym. 2019.)



Kuva 5 Vaiheet, joissa tekoälyn riskit syntyvät (Cheatham ym. 2019)

Seuraavissa alaluvuissa käytännön haasteita tuodaan esiin erityisesti datanhallinnan ja tekoälyn teknisten ekosysteemien kautta ja luvussa 3.4 poraudutaan syvemmälle löydettyihin teemoihin.

3.3.1 Datan keruu ja käyttö

"On päätettävä millaista aineistoa voi kerätä tai käyttää, mihin tarkoituksiin ja kenen toimesta, missä kulkevat hyväksyttävän ja vältettävän rajat, ja kuka niihin voi vaikuttaa ja millä aikavälillä", toteavat kuluttajatutkijat Lehtiniemi & Ruckenstein (2019) tuoden esiin useam-

mankin kysymyksen liittyen nimenomaan tekoälyn polttoaineen eli datan keruuseen ja käyttöön. Näiden lisäksi datan keräämiseen ja käyttöön liittyvät kysymykset siitä, mistä dataa hankitaan, miten dataa käsitellään ja kuka datan omistaa (ennen ja jälkeen tekoälyn prosessoinnin) sekä viestinnällinen kysymys siitä, missä määrin datan keruuta ja käyttöä on avattava eri sidosryhmille, kuten esimerkiksi kuluttajalle ja kenen tehtävä se on (Merilehto 2018, 164–165, 187).

Vastauksena muun muassa tämänkaltaisiin kysymyksiin vuonna 2018 voimaan astui EU:n yleinen tietosuoja-asetus 679/2016, josta puhutaan yleisesti GDPR:nä. Sen tarkoituksena on turvata yksilöiden oikeus omiin henkilötietoihinsa ja oikeus kontrolloida tietojen käsittelyä ja varmistaa, että henkilötietoja käytetään tarkoituksenmukaisesti ja vain tarveperusteisesti (Koskinen 2017, 11). Asetusta on sekä kiitelty että kritisoitu. Koneoppimisen ja laajemmin tekoälyn näkökulmasta asetukseen liittyy kuitenkin selviä ongelmakohtia. Koskinen (2017, 13) vertaa työssään tietosuojan ja big datan perusperiaatteita toisiinsa mutta toteaa niiden pätevän myös koneoppimiseen: kun suuresta määrästä dataa halutaan selvittää yhteyksiä ja riippuvuuksia kokeilemalla ja koska mahdollisesti käyttöön jääviä koneoppimismalleja opetetaan vain osin anonymisoidun datan avulla ja mallien kehittäminen edellyttää pääsyä dataan, ollaan tietosuojan kanssa helposti ongelmissa. Tietosuojan periaatteet, joiden kanssa koneoppimisen traditiot ovat ristiriidassa, ovat ainakin anonymisointi, tietojen minimointi, säilytyksen rajoittaminen ja käyttötarkoitussidonnaisuus (Koskinen 2017, 13).

Anonymisoinnilla tarkoitetaan henkilötietojen muuttamista eli tilastollistamista tai aggregoimista siten, ettei yksittäinen henkilö ole enää niiden avulla tunnistettavissa. Muutos on tehtävä niin, ettei sitä enää voi peruuttaa. Anonymisoituun tietoon ei sovelleta EU:n tietosuoja-asetusta, koska sen ei katsota olevan enää henkilötietoa. Toinen vastaavanlainen yksilöä suojaava mekanismi on pseudonymisointi eli tietojen muuttaminen sellaiseen muotoon, että ilman lisätietoja henkilöä ei voida aineistosta tunnistaa. Pseudonymisoitu tieto on kuitenkin edelleen henkilötietoa, koska yksilö on tunnistettavissa, kun koodaus tai salaus puretaan tai henkilöön liittyvät tiedot yhdistetään. Anonymisoidun datan problematiikka liittyy siihen, että vaikka anonymisoitua dataa saisikin käyttää, on tutkimuksissa osoitettu, että on haastavaa yhtä aikaa pyrkiä hävittämään datasta kaikki yhteydet, jotka johtaisivat henkilön tunnistamiseen, ja pyrkiä löytämään datasta kaikki ne yhteydet ja säännönmukaisuudet, jotka tekevät datasta merkittävää. Sitäkään ei yksiselitteisesti määritellä, milloin data muuttuu henkilötiedosta anonymisoiduksi dataksi. (Tietosuojavaltuutetun toimisto s.a.; Koskinen 2017, 25–27, 32–33.)

Tietojen minimoinnin haasteet taas liittyvät sekä siihen, miten ja kuinka paljon dataa saa kerätä, että siihen, mikä määritellään tarpeelliseksi datan käsittelyksi. Huomionarvoista nimittäin on myös Koskisen esiin nostama seikka, että asetusta noudattamalla saatetaan vaikuttaa opetusdatan kattavuuteen ja edustavuuteen, jolloin mahdolliset vinoumat siirtyvät koneoppimismalliin. (Koskinen 2017, 14–15.)

Käyttötarkoituksen ja säilytyksen rajoittamisen problematiikka liittyy sekä käyttötarkoitukseen että käyttötarkoituksen rajoittamaan säilytysaikaan. Ensinnäkään henkilötietoja ei saa käyttää muuhun kuin niiden alkuperäisen keräyssyyn edellyttämään käyttöön eikä niitä siten myöskään saa säilyttää pidempään kuin alkuperäinen käyttötarkoitus edellyttää. Haaste on varsin konkreettinen, koska tekoälyn näkökulmasta data on usein sitä arvokkaampaa ja mallit sitä virheettömämpiä, mitä pidemmältä ajanjaksolta ja mitä monipuolisempaa se on. Aina ei kuitenkaan datan keruu- tai käyttöhetkellä tiedetä, mihin dataa voisi hyödyntää myös tulevaisuudessa. Dataa voi kuitenkin käyttää ainakin, jos henkilö itse on antanut siihen luvan, jos data on anonymisoitu tai jos dataa ei käytetä alkuperäisen käyttötarkoituksen kanssa yhteensopimattomalla tavalla. (Koskinen 2017, 13–18.)

Nämäkään eivät kuitenkaan ole aivan yksiselitteisiä asioita, sillä muun muassa sitä voidaan pohtia ja arvioida, onko henkilö tunnistettavissa myös anonymisoidusta datasta varsinkin, jos datoja yhdistellään, missä määrin henkilön on mahdollista ymmärtää, millaisia seurauksia henkilötietojen käyttöluvan antamisella voi olla, mitä tarkoitetaan asetuksessa mainitulla yhteensopimattomuudella ja onko lain kiertämistä määritellä käyttötarkoitus niin laveasti, että se mahdollistaa kerätyn datan käytön myös tulevaisuuden tarpeisiin. Asetuksen mukaan henkilöllä on myös oikeus tulla unohdetuksi eli poistattaa kaikki omat tietonsa, jolloin esille tulee kysymys siitä, pitääkö henkilö poistaa myös opetetusta koneoppimismallista, jos hänen tietojaan on käytetty sen kouluttamiseen ja jos pitää, millaista taloudellista vahinkoa tämä yritykselle voi aiheuttaa, onko se ylipäätään mahdollista ja mihin suuntaan se koulutettua mallia muuttaa. (Koskinen 2017, 13, 21, 48, 51.)

On myös huomattava, että henkilötietojen anonymisointi sinänsä on henkilötietojen käsittelyä ja että jos henkilö pystytään tunnistamaan yhdistelemällä eri tietoja, kyseessä on GDPR:n mukaan henkilötieto. Dataan liittyy myös paljon sellaisia haasteita, jotka eivät liity suoraan henkilötietojen käsittelyyn ja tietosuojaan. Esimerkiksi tasa-arvoon ja oikeudenmukaisuuteen liittyvät kysymykset palaavat usein opetusdataan ja siinä esiintyviin vinoumiin. Vinoumat voivat liittyä niin datan kattavuuteen ja edustavuuteen, datan laatuun kuin esimerkiksi ihmisen tekemiin epäoikeudenmukaisiin päätöksiin ja vääriin tulkintoihin tai valintoihin.

Ollila tuo esiin datan laadukkuuden alttiuden ihmisen erehtyväisyydelle ja valinnoille. Ensinnäkin datan laadukkuus on kontekstisidonnaista; datan on oltava niin oikea-aikaista, oikeellista ja tarkkaa kuin käyttökohde edellyttää. Toisekseen ihminen usein myös ohjaa datan valintaa, varastointia, analysointia ja tulkintaa. Valintoihin voivat vaikuttaa niin yksilön kuin organisaationkin valinnat (vinoumat), kuten tiimien organisoituminen, valittu strategia, johtajien näkemykset tai yksinkertaisesti joidenkin datasettien helpompi saatavuus. Toisaalta itse datakin voi olla etiikan kannalta varsin ongelmallista, koska data kertoo totuuden, ei niitä moraalisia hyveitä, joita ihmiset sanovat noudattavan, jolloin algoritmeille ei voi opettaa moraalia ihmisen käyttäytymisen perusteella. (Ollila 2019, 116, 126–127.)

Rusanen & Lappi huomauttavatkin, että vaikka julkinen keskustelu vinoumien osalta on koskenut pääasiassa dataan liittyviä vinoumia, ne eivät varsinaisesti ole tekoälyn tuottamia: tekoäly ei vain kykene tunnistamaan ja poistamaan virheellisiä säännönmukaisuuksia. Sen sijaan algoritmien, tekoälyarkkitehtuurien tai tutkimuksen aiheuttamat vinoumat saattavat tuottaa vääristymiä: esimerkiksi satunnaismenetelmien vaikutuksia ei voida tarkkaan ennakoida. (Rusanen & Lappi s.a. 3–4.)

Ongelmallista voi olla myös tekoälyjärjestelmien pohjautuminen dataan. Adoben Data Science Lab:n johtajan mukaan datan avulla ei voida selittää kaikkia tapahtumia eikä siten koneoppimismalleiltakaan ole mahdollista saada tarvittavaa selitystä. Hänen mukaansa tilastollinen päättely ei voi tuottaa selityksiä syy-seuraussuhteista, vaan siihen tarvitaan lisää dataa esimerkiksi seurauksista ja toimenpiteistä, ja koska data edustaa jo tapahtunutta tarvittaisiin lisää oppimisdataa, jota ei kuitenkaan välttämättä ole olemassa. Tämä pätee esimerkiksi autonomisten autojen ohjaukseen erilaisissa olosuhteissa. (Pietikäinen & Silvén 2019, 216.)

3.3.2 Sovellukset ja ekosysteemit

”Tekoälyn laskentaympäristöjä voi yksinkertaistaen tarkastella pinona, jossa alimpana on laskentayksikkö eli prosessori, sen yläpuolella käyttöjärjestelmä ja ohjelmointikieli, ylimpänä itse varsinainen tekoälysovellusohjelma. Yleensä laskentaympäristö on ns. pilvessä, ja laskenta suoritetaan palveluntarjoajan alustalla ja se ostetaan palveluna. Eri toimijat muodostavat ekosysteemin, joka yhdessä toteuttaa halutun tekoälylaskennan tai sovelluksen.” (Ailisto ym. 2019, 6.)

Muun muassa Ailisto ym. (2018, 40) on kiinnittänyt huomiota sellaisiin tekoälyn ominaisuuksiin, jotka on otettava huomioon teknologian kehityksessä: toisaalta teknologialla on oltava älykkäitä ominaisuuksia, kuten oppivuutta, autonomisuutta ja suorituskykyä, toisaalta kehityksessä on huomioitava tekoälyn lähdekoodin, menetelmien ja päätösten läpinäkyvyys

sekä varmistettava hallittavuus, (tieto)turvallisuus ja riittävä yksityisyys. Tekoälyn kehitykseen ei kuitenkaan ole olemassa yhteisiä standardeja, menettelytapoja ja ohjeistuksia, joilla varmistuttaisiin esimerkiksi, että tekoälysovelluksissa käytettävä opetusdata on laadukasta, datasta tai käytetystä menetelmästä johtuvia vinoumia ei ole ja muutokset sovelluksen käyttöympäristössä on ennakoitu ja sovellus on verifioitavissa ja validoitavissa. Tästä syystä riski voi kohdistua niin yksityishenkilöihin kuin yrityksiinkin eikä viranomaisillakaan välttämättä ole toimivaltuuksia tai teknisiä valmiuksia puuttua epäkohtiin. Yksityishenkilöiden kannalta epäselvää on esimerkiksi, ovatko tekoälyn tekemät virheelliset päätökset jäljitettävissä, ja kuka tai mikä taho on korvausvelvollinen. Syy-seuraussuhteiden, virheiden ja vahinkojen todistaminenkin voi olla haastavaa ja vaatimusten puute vaikeuttaa myös yritysten kilpailukykyä. (Ailisto ym. 2018, 43; Euroopan komissio 2019, 14.)

Datan lisäksi myös algoritmien kohdalla voidaan puhua vinoumista, joita voidaan tarkastella niin päätöksentekoprosessin kuin tuloksen neutraaliuden kautta. Käytettyjen algoritmien tulisi olla yleisesti hyväksyttyjä ja läpinäkyviä, mutta sekään ei vielä takaa lopputuloksen puolueettomuutta. Tähän on ainakin kolme syytä myöhemmin käsiteltävän (luku 3.4.6) mustan laatikon ongelman lisäksi ja ne kaikki liittyvät ihmisen toimintaan: algoritmit sisältävät usein liikesalaisuuksia, jolloin niiden toimintaa ei haluta avata ulkopuolisille, toisekseen algoritmien kehittäjät saattavat tiedostamattaan siirtää omia asenteitaan osaksi algoritmin toimintaperiaatetta ja kolmannekseen algoritmien käyttäjät saattavat järjestelmällisesti manipuloida algoritmien toimintaa. Näiden osalta kysymys saattaakin olla enemmän ihmisen kognitiivisesta ajattelusta kuin itse algoritmien vinoumista. (Ollila 2019, 126.)

"Algoritmien tavoitteena on mallintaa ihmisen ajattelua ohjelmallisin keinoin. Algoritmisuunnittelun haasteena on luoda eri tilanteet ja olosuhteet hallitseva algoritmi ja samalla antaa autonomiselle robotille kyky ihmismäiseen loogiseen ja analyyttiseen päätöksentekoon." Tästä syystä algoritmeihin liittyy myös useita turvallisuusnäkökulmia. Sen lisäksi, että algoritmin koodaaja voi tehdä virheen, voi algoritmi myös itse toimia virheellisesti tai ennakoimattomasti. Lisäksi algoritmien säännönmukaisuutta voidaan hyödyntää hakkeroinnissa siinä missä dataakin voi manipuloida syöttämällä sinne sellaista dataa, jota ihminen ei kykene huomaamaan, mutta joka muuttaa tekoälyn tuloksia. (Siukonen & Neittaanmäki 2019, 302–303.)

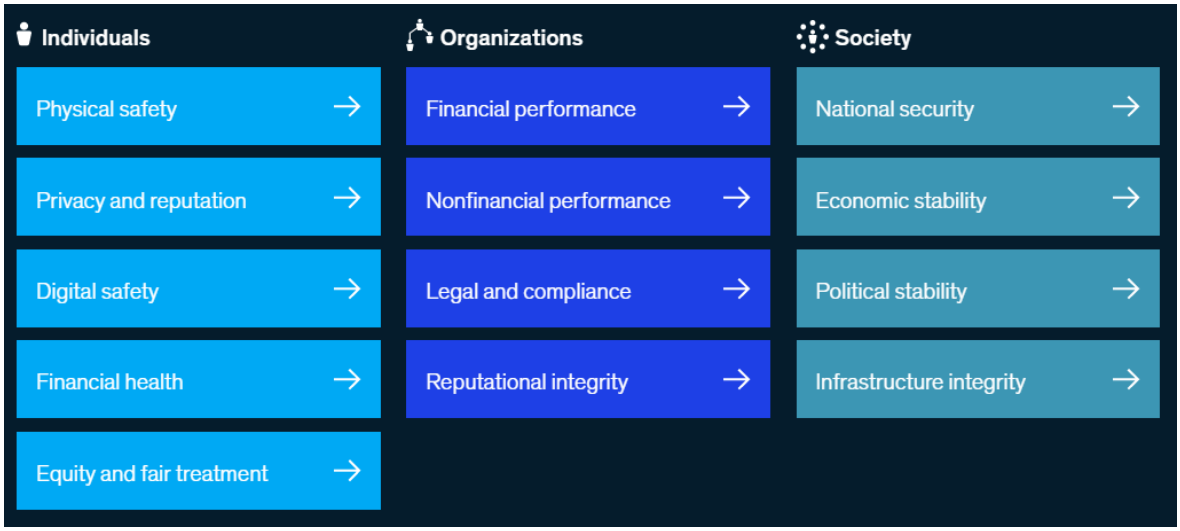
Lisäksi datan määrä ja tarvittava laskentateho edellyttävät usein pilvessä sijaitsevaa, kaupallista tekoälyalustaa, joka tällä hetkellä usein on jonkun teknologiajäteistä. Ailiston ym. mukaan (2018, 19) alustojen heikkoutena pidetään erityisesti yksityisyydensuojaan ja salassapidettävään tietoon liittyviä kysymyksiä, kuten viranomaistietoa ja liikesalaisuuksia.

Saman huolen esiin nostaa myös Ojanen ym. (2019,12), jonka mukaan alustatalouden kehittyminen avaa laajemminkin yksityisyyteen liittyviä kysymyksiä, koska niiden myötä yhteystietoja ja muuta henkilökohtaista tietoa on yhä helpompi kerätä ja hyödyntää yhä tarkempaan ennustamiseen, profilointiin ja ihmisten toiminnan seuraamiseen ja manipulointiin, jopa massavalvontaan.

Teknologiajättien kohdalla on myös mainittava riski monopoliasemasta, joka vähentää kilpailua ja innovaatioita sekä supistaa kuluttajan vapautta valita. Euroopan komissio on nostanut esiin esimerkiksi huolen dataan pohjautuvista IoT-laitteista ja pitää mahdollisena, että suuryritykset käyttävät väärin tekoälyä hyödyntävien älylaitteiden ja ääniassistenttien keräämää dataa kulutustottumuksistamme. Komissio on tästä syystä käynnistänyt tutkimuksen, jossa se pyrkii selvittämään kodin älylaitteiden, ääniassistenttien ja puettavan teknologian (esimerkiksi älyvaatteet ja älykellot) myyjien ja niihin liittyvien palveluntarjoajien datan keräämistä, käyttöä ja datalla ansaitsemista. (Pervilä 2020; Euroopan komissio 2020b.)

3.3.3 Riskit ja vaikutukset eri sidosryhmille

Vaikutustenarvioinnissa keskitytään usein yksilöön ja yksityisyydensuojaan, mutta oleellista on ulottaa vaikutusten ja riskien arviointi myös laajempaan kontekstiin. Ollila (23.1.2020) tuo esiin esimerkiksi henkilökohtaiseen dataan perustuvan profiloinnin yhteiskunnalliset vaikutukset: mitä demokratialle tapahtuu, jos esimerkiksi viestintää voidaan vaalien alla kohdentaa profiloinnin avulla tavalla, jolla voidaan vaikuttaa myös vaalituloksiin, mitä tapahtuu datan käyttöä koskevalle luottamukselle, mikäli esimerkiksi sosiaalisesta mediasta vuotaa jatkossakin tietoa kolmansille osapuolille tai miten jatkuva tietojemme tallentaminen ja valvonta muuttaa käyttäytymistämme. Riskejä kartoitettaessa läsnä on siis aina kysymys, mistä näkökulmasta niitä tarkastellaan ja siksi on syytä jäsennellä vielä riskien merkittävimpiä, pääosin tahattomia, seurauksia eri sidosryhmille (kuva 6).



Kuva 6 Tekoälyn epätoivotut vaikutukset (Cheatham ym. 2019)

Yksilöiden riskit liittyvät niin fyysiseen kuin digitaaliseen turvallisuuteen, yksityisyyteen ja maineeseen, talouteen ja oikeudenmukaiseen kohteluun. Fyysistä turvallisuutta voivat uhata esimerkiksi väärät diagnoosit ja autonomisten kulkuvälineiden toimintahäiriöt. Yksityisyyttä loukkaa tietojen käyttö ilman suostumusta tai tietojen puutteellinen suojaaminen. Epäoikeudenmukainen kohtelu voi ilmentyä vähemmistön syrjintänä päätöksenteossa. (Cheatham ym. 2019.)

Organisaatioissa vaikutukset ulottuvat ainakin talouteen, toimintakykyyn, maineeseen ja vastoin lakeja ja sääntöjä toimimiseen. Talouteen vaikuttaa esimerkiksi kaupantekoalgoritmi, joka ei osaakaan huomioida muuttuneita olosuhteita, tai tekoäly, joka tekee epäedullisia hinnoittelupäätöksiä tai antaa perusteetonta joustoa. Organisaation toimintakykyyn voi vaikuttaa rekrytointialgoritmi, joka johtaa homogeenisempiin henkilöstövalintoihin, tai resursseja virheellisesti arvioiva algoritmi, mikä johtaa puutteelliseen valmistautumiseen esimerkiksi hätätilanteita varten. Maineeseen vaikuttavat esimerkiksi puutteellinen tietosuoja-asetusten huomioiminen tai suojattujen tietojen, kuten terveystietojen, paljastaminen. (Cheatham ym. 2019.)

Yhteiskuntaa uhkaavat ainakin taloudellinen tai poliittinen epävakaus tai vaikutukset kansalliseen turvallisuuteen ja infrastruktuurin eheyteen. Nämä voivat ilmentyä esimerkiksi vaaliprosessien manipulointina virheellisen tiedon avulla, sähkön- tai vedenjakelun tai tietoliikenneyhteyksien häiriöinä tai algoritmien luomana epävakautena valuuttamarkkinoilla. (Cheatham ym. 2019.)

3.4 Luottamuksenarvoisuuden edellytykset

”Teknologiaa ei voida koskaan erottaa materiaalisesta, sosio-kulttuurisesta ja yhteiskunnallisesta käyttökontekstistaan. Eettinen suunnittelu ei siis koske vain ja ainoastaan robotin toiminnallisuutta ja käytettävyyttä, vaan sen istumista laajempaan ekologiseen ympäristöön, osaksi ihmisten toimintaa ja yleisesti hyväksyttyjä yhteiskunnallisia käytäntöjä.” (Ojanen ym. 2019, 20.)

Tekoälyn ja laajemmin teknologian etiikassa oleellista, sen lisäksi, että keskustellaan ja päätetään siitä, mikä on oikein ja väärin, mitä arvoja haluamme edistää ja millaisia normeja noudattaa, on sekä tarkentaa, mitä periaatteet tarkoittavat teknologian suunnittelun ja käytön kannalta, että altistaa eettisyys julkiselle keskustelulle. Tuolloin joudutaan pohtimaan myös sitä, millaisia uhkia teknologiaa koskevilla eettisillä periaatteilla torjutaan tai toisaalta mitä niillä mahdollistetaan. (Koivisto ym. 2019, 10.)

Voidaan esimerkiksi Iso-Britannian kansallisen datatiede- ja tekoälyinstituutin, the Alan Turing Institute:n, tavoin puhua *eettisistä arvoista ja toiminnallisista periaatteista*. Eettiset arvot tukevat ja motivoivat vastuullista tekoälyn suunnittelua ja käyttöä. Tällaisia eettisiä arvoja ovat *”kunnioita yksilöiden ihmisarvoa”* tai *”huolehdi kaikkien hyvinvoinnista”.* Esimerkiksi ihmisarvoa kunnioittavat päätökset varmistavat muun muassa yksilöiden mahdollisuuden ja kyvyn tehdä omaa elämäänsä koskevia päätöksiä tietoon perustuen ja niillä turvataan yksilöiden itsenäisyys, valta ilmaista itseään ja oikeus tulla kuulluksi. (Leslie 2019, 7, 10.)

Toiminnalliset periaatteet puolestaan liittyvät käytännöllisemmin tekoälyn kehittämiseen ja käyttöön ja ne juontavat juurensa edellä mainittuun perustavanlaatuiseen dilemmaan siitä, mitä koneilta odotetaan. Kun ihmiset tekevät älykkyyttä vaativia tehtäviä, pidämme heitä vastuussa toiminnan ja päätöksenteon huolellisuudesta ja luotettavuudesta. Lisäksi vaadimme, että päätökset tehdään oikein perustein ja pidämme päätöksentekijöitä vastuussa siitä, että päätökset ovat kohtuullisia ja oikeudenmukaisia. Kun päätöksenteko on siirretty koneelle, on älykkyyttä edellyttävän asian suorittaminen edellyttänyt useiden kognitiivisten toimintojen siirtämistä algoritmisiin prosesseihin, joita ei voida pitää moraalisessa vastuussa toimintansa seurauksista. Tähän tarpeeseen monet eettiset periaatteet ovat syntyneet: kaventamaan kuilua koneen älykkään päätöksenteon ja sen perustavanlaatuisen moraalisen vastuun puutteen välillä. (Leslie 2019, 12.)

The Alan Turing -instituutin *toiminnalliset periaatteet* ovat oikeudenmukaisuus, läpinäkyvyys, kestävyys ja vastuuvelvollisuus, joista oikeudenmukaisuutta ja kestävyyttä instituutti pitää sellaisina algoritmisten järjestelmien ominaisuuksina, joista tekoälyn suunnittelijat ja toteuttajat ovat suoraan vastuussa. Läpinäkyvyyden ja vastuuvelvollisuuden periaatteet sen

sijaan ovat näiden järjestelmien suunnittelussa, toteutuksessa ja tulosten arvioinnissa kriittisiä, huomioonotettavia aspekteja ja ne tarjoavat menettelytapoja ja keinoja, joiden kautta suunnittelijat ja toteuttajat osoittavat vastuunsa sekä käytännöistään että järjestelmien lopputuloksista ja joiden avulla tekoälyjärjestelmät voidaan perustella. Instituutti myös huomauttaa, että läpinäkyvyys, vastuuvelvollisuus ja oikeudenmukaisuus kuuluvat myös tietosuojaperiaatteisiin ja jos algoritmiseen käsittelyyn liittyy henkilötietoja, niiden noudattaminen ei ole pelkästään eettistä tai hyvien käytäntöjen noudattamista, vaan oikeudellinen vaatimus, joka on vahvistettu yleisessä tietosuoja-asetuksessa. (Leslie 2019, 12–13.)

Seuraavissa luvuissa käsitellään sitä, mitä eurooppalaisessa kontekstissa luottamuksenarvoiselta tekoälyn käytöltä tulee edellyttää ja poraudutaan jo edellistä lukua syvemmälle esiteltyihin teemoihin. Samalla tuodaan esiin teemoja koskevia ajankohtaisia ratkaisuvaihtoehtoja ja eri lausuntojen kautta ilmenneitä kehittämistarpeita, esimerkiksi puutteita lainsäädännössä, jotka voivat edellyttää yrityksiltä laajempaa tai yksityiskohtaisempaa itsesääntelyä.

Teemat on johdettu pitkälti AI HLEG:n (2019, 7) seitsemästä luotettavan tekoälyn vaatimuksesta, jotka ovat

- Ihmisen toimijuus ja ihmisen suorittama valvonta
- Tekninen luotettavuus ja turvallisuus
- Yksityisyyden suoja ja datanhallinta
- Läpinäkyvyys
- Monimuotoisuus, syrjimättömyys ja oikeudenmukaisuus
- Yhteiskunnallinen ja ekologinen hyvinvointi
- Vastuuvelvollisuus

sekä the Alan Turing -instituutin (Leslie 2019, 5–6), vastuullisen tekoälyn tavoitteista, jotka vapaasti käännettyinä ja lyhennettyinä ovat seuraavat:

- Varmista, että tekoälyhankkeessa on huomioitu yksilöihin ja yhteisöihin kohdistuva oikeudenmukaisuus ja syrjimättömyys kiinnittämällä huomiota vinoumiin, jotka voivat vaikuttaa käytetyn tekoälymallin tuottamaan lopputulokseen ja tiedostamalla oikeudenmukaisuuteen liittyvät riskit kaikissa tekoälyn suunnittelu- ja toteutusvaiheissa.
- Varmista, että tekoälyhanke on perusteltavissa sekä suunnittelu- että toteutusprosessin läpinäkyvyyden että sen muodostamien päätösten ja käyttäytymisen läpinäkyvyyden ja tulkittavuuden osalta.
- Varmista, että tekoälyhanke on julkisen luottamuksen arvoinen tekoälyn turvallisuuden, luotettavuuden ja tarkkuuden osalta.
- Varmista, että tekoälyhanke on eettisesti hyväksyttävä ottamalla huomioon sen vaikutukset sidosryhmien ja yhteisöjen hyvinvointiin.

3.4.1 Eettisyys

Varsinaiset syyt, miksi tekoälyn etiikka on viime vuosina noussut ihmisten tietoisuuteen, ovat varmasti moninaisia, mutta yksi on se, että ihmiset tulevat yhä useammin ja yhä useammassa kontekstissa olemaan vuorovaikutuksessa tekoälyn kanssa tilanteissa, joissa ihminen on aiemmin kommunikoinut toisen ihmisen kanssa. Tekoälyasiantuntija Pari Lehtimäen mukaan turvassa ollaan niin kauan kuin tekoälyä käytetään esimerkiksi sujuvoittamaan ihmisten arkea ja tekemään elämästä helpompaa, mutta koska tekoälyä voidaan käyttää myös esimerkiksi ihmisen ajatusten ja käyttäytymisen muokkaamiseen, ihmisten tulee olla varautuneita ja tietoisia siitä, kenen hyödyksi, mihin tarkoitukseen ja millä tavalla tekoälyä käytetään. (Lehtimäki 29.10.2020.)

Tekoälyyn liittyy aina eettisiä ja moraalisia kysymyksiä, koska tekoäly insinööritieteenhaarana ja teknologiana pyrkii *”vaikuttamaan ja muuttamaan maailmaa”.* Kysymykset voivat olla luonteeltaan niin *teknisiä* (tekoälyn virheiltä välttyminen), *taloustieteellisiä* (työttömyys, vaurauden keskittyminen) kuin *yhteiskunnallisiakin* (ihmisen käyttäytymisen muutos, vuorovaikutus) ja liittyä niin *kyberturvallisuuteen* (kyberhyökkäykset) kuin *moraalifilosofiaankin* (koneiden tietoisuus ja oikeudet, singulariteetti). (Ailisto ym. 2018, 21–22.)

Tekoälyn etiikalla ei tarkoiteta sanatarkasti tekoälyn luomaa etiikkaa, vaan tekoälyn kehitykseen ja käyttöön sovellettavaa etiikkaa, jota voidaan tarkastella ainakin kahdesta eri näkökulmasta: toisaalta tekoälyn etiikalla voidaan tarkoittaa tekoälyn käyttötarkoituksen eettisyyttä ja toisaalta taas itse tekoälytoteutusten ja tulosten eettisyyttä (Ollila 2019, 11, 146; Jääskeläinen 2019, 81). Tekoälyyn liitettyjä etiikan ja moraalin osa-alueita ovat esimerkiksi jo mainittu *moraalifilosofia* (esimerkiksi hyve-etiikka ja seurausetiikka*), soveltava etiikka* (esimerkiksi tapauskohtainen päätöksenteko), *teknologian etiikka* (esimerkiksi suorat ja välilliset vaikutukset ihmisiin) ja *sodankäynnin etiikka.* Tekoäly tuo näihin uusia näkökulmia ja haasteita, sillä esimerkiksi teknologian etiikka on perustunut aiemmin teknologiaa suunnittelevan ihmisen toimijuuteen, ei siihen, että eettisen ja moraalisen päätöksen tekisi autonominen kone. (Ailisto ym. 2018, 21–22.)

Ollila (2019, 13, 53) kyseenalaistaakin *”riittääkö tähän asti luotu etiikka ratkomaan tekoälyyn liittyvät ongelmat”,* sillä tekoälyn ominaisuuksista erityisesti autonomisuus (itsenäinen suoriutuminen) ja adaptiivisuus (kokemuksesta oppiminen) asettavat etiikalle haasteita. Myös eetikko Anna Seppänen (29.10.2020) korostaa, että tekoälyn etiikan kysymykset eivät ole teknologian etiikkaa *”as usual”*, koska tekoäly valtaa alaa kaikkialla, myös moraalisen toiminnan ydinalueilla, kuten hoivatyössä ja demokratiassa. Toisekseen tekoälyn etiikka ei ole *”vain etiikkaa”*, vaan tekoälyn osalta eettiset kysymykset ovat vasta muotoutumassa. (Seppänen 29.10.2020).

Toisaalta tekoälyn etiikkaan liittyvä kiinnostava kulma on myös se, miltä osin eettiset kysymykset liittyvät puhtaasti vain tekoälyyn ja miltä osin tekoäly suurine yhteiskunnallisine vaikutuksineen vain ajaa ihmiset miettimään eettisiä ja moraalisia kysymyksiä. Ollila (2019, 9) puhuukin laiskanläksyistä: *”Tekoälyn esiinmarssi vain pakottaa meidät lopultakin tekemään yhteisiä päätöksiä asioista, jotka olemme tähän asti jättäneet kunkin yksilön harkintavaltaan”.*

3.4.2 Laillisuus, sääntely ja valvonta

Niin lainsäädännöllä kuin muulla sääntelyllä voidaan sekä edistää teknologioiden kehitystä ja kaupallistumista mutta myös hidastaa niitä. Sääntely voidaan jakaa *teknologiseen* (esimerkiksi standardointi), *taloudelliseen* (markkinoiden tehostaminen), *sosiaaliseen* (esimerkiksi ympäristön, yhteiskunnan ja ihmisten suojelu) ja *hallinnolliseen* (esimerkiksi yksityisen ja julkisen sektorin käytännöt) sääntelyyn, ja näiden osalta eri tahoilla on jo käynnissä toimenpiteitä. Keskeiset tekoälyyn liittyvät lainsäädännölliset kysymykset koskevat datanhallintaa, yksityisyydensuojaa, kansalaisten yhdenvertaisuutta, vastuukysymyksiä, kilpailua ja määräävää markkina-asemaa, kansallista turvallisuutta sekä kansainvälisiä sopimuksia liittyen sodankäyntiin, joista seuraavassa tarkastellaan tarkemmin vastuukysymyksiä ja perusoikeuksien turvaamista. (Ailisto ym. 2018, 23.)

Euroopan Unionin tuoteturvallisuus- ja tuotevastuulainsäädäntö alakohtaisine sääntöineen ja kansallisine lisälainsäädäntöineen ottaa kantaa ihmisten perusoikeuksiin ja kuluttajan oikeuksiin, esimerkiksi tietosuoja-asetuksin. Lainsäädäntöä sovelletaan jo nyt tekoälyn hyödyntäjiin, mutta Euroopan komission mukaan tätä muutoin mittavaa lainsäädäntöä olisi kuitenkin uudelleentarkasteltava tekoälyn osalta viidestä eri näkökulmasta: soveltaminen ja täytäntöönpano, soveltamisala, muuttuva toiminnallisuus, vastuun jakautuminen ja turvallisuuden käsite. (Euroopan komissio 2020a, 11, 15–16.)

Tekoälyteknologioilla on piirteitä, joita ei ole nimenomaisesti otettu huomioon voimassa olevassa lainsäädännössä, ja jotka vaikeuttavat lainsäädännön toteutumisen valvomista ja siten täytäntöönpanoa ja oikeussuojan toteutumista. Sen enempää lainsäätäjillä kuin kuluttajillakaan ei ole välttämättä keinoja osoittaa ja todistaa säännöstenvastaisuutta, jos esimerkiksi tekoälyä päätöksenteossaan hyödyntävä taho ei avaa tekoälysovelluksensa toimintaa ja sitä, miten päätös on muodostunut. Komission mukaan lainsäädäntöä tulisi siis tarkastella erityisesti sellaisten *”säännösten osalta, jotka koskevat perusoikeuksien suojelua, vastuuseen asettamista tai korvauksen vaatimisen edellytysten täyttymistä”.* (Euroopan komissio 2020a, 11, 13–16.)

Soveltamisalan, muuttuvuuden ja vastuun jakautumisen kysymykset puolestaan liittyvät pitkälti tekoälyn tuotteistamiseen ja palvelullistamiseen. Esimerkiksi jos tekoälyä sisältävä ohjelmisto toimitetaan tiettyyn alaan liittyvän, esim. lääkinnällisen, laitteen mukana, se ei välttämättä kuulu EU:n tuoteturvallisuuslainsäädännön soveltamisalaan, vaan se liittyy lääkinnällisiä laitteita koskevaan asetukseen. Lisäksi tuoteturvallisuuslainsäädännöllä pyritään valvomaan erityisesti markkinoille tulevia tuotteita, jolloin on mahdollista, että tuote on ollut lainmukainen tullessaan markkinoille, mutta muuttunut elinkaarensa aikana (perustuessaan esimerkiksi koneoppimisalgoritmeihin) ja täten aiheuttanut uuden turvallisuusriskin, johon lainsäädäntö ei enää ulotu. (Euroopan komissio 2020a, 15–16.)

Tähän liittyy myös vastuun jakautumisen problematiikka, sillä lainsäädännön mukaan vastuu tuotteesta, myös tekoälyjärjestelmästä, kuuluu tuotteen markkinoille saattajalle, mutta lainsäädäntö ei ota suoraan kantaa, kenelle vastuu kuuluu, jos valmiiseen tuotteeseen lisätään tekoälyä. Lainsäädäntö ei myöskään ulotu palveluihin, jotka perustuvat tekoälyn hyödyntämiseen, vaan pelkästään tuotteisiin. Lainsäädäntö ei myöskään ota riittävästi kantaa uudentyyppisiin ja turvallisuuden käsitettä laajentaviin turvallisuusriskeihin, kuten verkkoyhteyteen, henkiseen turvallisuuteen (kanssakäyminen koneiden kanssa) ja henkilökohtaiseen turvallisuuteen (kuten kodinkoneiden turvallisuuteen) tai kyberturvallisuuteen. (Euroopan komissio 2020a, 15–16.)

Komissio esittääkin, että datan laatuun, algoritmien läpinäkyvyyteen, ohjelmistoihin ja toimitusketjuihin liittyen voisi asettaa erityisiä velvoitteita ja vaatimuksia sekä toteaa, että *"suojatoimenpiteenä saatetaan tarvita myös ihmisen suorittamaa valvontaa tuotesuunnittelusta alkaen tekoälytuotteiden ja -järjestelmien koko elinkaaren ajan".* Komission mukaan tarvittaneen siis sekä tekoälyä koskevaa lainsäädäntöä että mukautuksia olemassa olevaan lainsäädäntöön. (Euroopan komissio 2020a, 17–18.)

3.4.3 Oikeudenmukaisuus ja syrjimättömyys

"Oikeudenmukaisuuden periaate pyrkii tarkastelemaan sitä, miten tietyssä yhteisössä yhteisön jäsenten tulisi toimia ja miten haittojen ja etujen tulisi yhteisössä jakaantua. Oikeudenmukaisuudessa on siten kyse myös siitä, mitä oikeuksia ja velvollisuuksia yhteisön jäsenillä on." (Koivisto ym. 2019, 13.)

Kun ajatellaan tekoälyjärjestelmien oikeudenmukaisuutta, on tärkeää ymmärtää, että järjestelmät eivät koskaan ole itsessään puolueettomia, koska ne ovat ihmisten kehittämiä. Inhimilliset virheet, ennakkoluulot ja virhearvioinnit voivat luoda vinoumia missä tahansa tekoälyhankkeen vaiheessa. Lisäksi datavetoiset teknologiat perustuvat data-aineistoihin, jotka

tallentavat monimutkaisia sosiaalisia ja historiallisia malleja, jotka voivat sisältää puolueellisuuden ja syrjinnän muotoja ja siksi oikeudenmukaisuuden varmistaminen ei ole yksinkertainen tai puhtaasti tekninen kysymys. Tämä ei kuitenkaan tarkoita sitä, etteikö oikeudenmukaisiin, moraalisesti hyväksyttäviin ja hyödyllisiin lopputuloksiin tulisi pyrkiä tekoälymallia luodessa ja etteikö oikeudenmukaisuutta tulisi pitää tärkeänä periaatteena. Oikeudenmukaisuuden periaatteella halutaan varmistaa, että kaikille tekoälyn vaikutuksen alaisille sidosryhmille turvataan tasapuolinen, oikeudenmukainen, moraalisesti hyväksyttävä ja hyödyllinen lopputulos ja syrjimättömyys on sen minimivaade. (Leslie 2019, 13–14.)

Oikeudenmukaisuuden tarkastelu voidaan jakaa kolmeen: datan, suunnittelun ja toteutuksen sekä lopputuloksen oikeudenmukaisuuden tarkasteluun. Dataa kerätessä on kiinnitettävä huomiota datan kontekstisidonnaiseen riittävyyteen ja toisekseen itse datalähteeseen. Huomiota on kiinnitettävä niin datan eheyteen kuin sen puolueettomuuteen eli siihen, ettei malliin siirry datan mukana ihmisten ennakkoluuloja tai puolueellisia päätöksiä. Eheyden lisäksi on ymmärrettävä, että vanhentuneella datalla voi olla vaikutuksia datan yleistettävyyteen. Lisäksi käytetyn data-aineiston on vastattava mallinnettavaa joukkoa siten, ettei mikään ryhmä esiinny datassa yli- tai aliedustettuna. Riski liittyy esimerkiksi lain nojalla suojattuihin ryhmiin. Kehittäjien tulisi myös huolehtia läpi prosessin esimerkiksi data-aineistojen metadatasta ja kontekstuaalisista tiedoista, jotta mahdollisesti myöhemmin syntyviin huolenaiheisiin pystyttäisiin prosessin avulla vastaamaan. Datan käsittelyssä tarvitaan siis niin teknistä osaamista kuin substanssiosaamista, jotta *datan edustavuus, riittävyys, tarkoituksenmukaisuus, ajantasaisuus ja merkityksellisyys* voidaan varmistaa. (Leslie 2019, 15–16.)

Läpi tekoälyn suunnittelun ja toteutuksen ihmiset tekevät valintoja, joilla on merkitystä siihen, millaisia vaikutuksia algoritmeillä on ja kohdallaanko esimerkiksi erityisryhmiä oikeudenmukaisesti. Oikeudenmukainen suunnittelu lähtee siitä, että jo varhaisessa vaiheessa ongelma ja toivottu lopputulos muutetaan tavoitteiksi ja tavoitteet mitattaviksi päämääriksi. Näihin päämääriin ja mahdollisiin vaikutuksiin, joita järjestelmän tuloksilla on asianomaisiin, myöhemmin suhteutetaan eri vaiheiden valinnat ja arvioidaan niiden kohtuullisuutta ja perusteltavuutta. Ihminen vaikuttaa algoritmin toimintaan jo siinä vaiheessa, kun se esikäsittelee eli järjestää, annotoi, merkitsee tai tekee esimerkiksi luokittelua koskevia valintoja. Myös attribuutteja ja ominaisuuksia määritettäessä, parametrejä ja mittareita säätäessä sekä testaus- ja arviointivaiheissa ihminen tekee valintoja, joilla voi olla merkitystä tarkan tai puolueettoman luokituksen tai ennusteen saamiseksi. Kehittäjien on lisäksi tutkittava merkittäviksi luokiteltujen korrelaatioiden moraalinen perusteltavuus ja päättelyt, jotka malli oppimismekanismiensa mukaan itse tuottaa. Vertaisarvioinnilla ja konsultoinnilla on siksi tärkeä rooli sen varmistamisessa, että tehdyt valinnat ovat linjassa oikeudenmukaisuuden vaatimuksen kanssa. (Leslie 2019, 16–17.)

Jos syrjimättömyyttä ei pystytä varmistamaan mallin monimutkaisuuden vuoksi, tulisi varsinkin sosiaalisten tai demografisten tietojen käsittelyssä käyttää selitettävämpiä malleja (kts. 3.4.6 Selitettävyys). Tekoälyjärjestelmässä käytettäviä sääntöjä tai menettelyjä on sovellettava johdonmukaisesti ja yhdenmukaisesti kaikkeen päätöksentekoon, joita kyseinen järjestelmä tekee. Toteuttajien olisi tässä yhteydessä kyettävä myös osoittamaan, että algoritminen tulos on toistettavissa, ja säännöt ja menettelytavat ovat samat kaikille, joiden tietoja algoritmi käsittelee. Tämä yhdenmukaisuus takaa päätöksen kohteena olevien henkilöiden yhdenvertaisen kohtelun. (Leslie 2019, 17–18.)

Lopputuloksen oikeudenmukaisuuteen liittyy kysymys siitä, miten määritellään ja mitataan tekoälyjärjestelmän vaikutusten ja tulosten oikeudenmukaisuutta. Tulosten oikeudenmukaisuuden määrittämisen pitäisi riippua sekä käyttötapauksesta että siitä, onko valitut kriteerit mahdollista sisällyttää tekoälyjärjestelmän rakentamiseen. Leslien mukaan tulos on oikeudenmukainen esimerkiksi, jos jokainen kyseessä olevan data-aineiston ryhmä hyötyy päätöksestä yhtä paljon. Sitä voidaan pitää oikeudenmukaisena myös, jos algoritmisen ennusteen tai luokituksen todellisten positiivisten määrä on sama eri ryhmissä eli jos oikein ennustettujen positiivisten osuus kaikista ennakoiduista positiivisista tapauksista on yhtä suuri kaikissa ryhmissä ja jos tekoäly kohtelee sellaisia henkilöitä samoin, joilla on samanlaiset ominaisuudet. Tämä lähestymistapa perustuu sellaisen samankaltaisuusmittarin luomiseen, joka osoittaa, missä määrin yksilöt ovat samankaltaisia tietyssä kontekstissa. (Leslie 2019, 18–19.)

Oikeudenmukaisuutta voidaan ajatella myös sekä aineellisena että menettelyllisenä ulottuvuutena. Aineellisesti tarkasteltuna on varmistettava hyötyjen ja kustannusten tasapuolinen jakautuminen ja se, että yksilöitä tai ryhmiä ei kohdella sen suhteen puolueellisesti, syrjivästi, leimaavasti tai epäoikeudenmukaisesti. Tähän liittyy myös sen varmistaminen, että tekoälytuotteet ja -palvelut ovat tasavertaisesti kaikkien käytettävissä eikä käyttöä rajoita esimerkiksi käyttäjien ikä, sukupuoli, tausta, ominaisuudet, toimintakyky, digitaidot tai kielitaito, vaan suunnittelussa on kiinnitetty huomiota ”*käytön ja käyttöliittymän tarkoituksenmukaisuuteen, helppouteen ja selkeyteen”.* Aineelliseen oikeudenmukaisuuteen liittyy myös se, ettei käyttäjien valinnanvapautta rajoiteta tai heitä johdeta harhaan. Menettelyllisesti oikeudenmukaisuus tarkoittaa mahdollisuutta muutoksenhakuun, mikä puolestaan edellyttää tekoälyjärjestelmältä selitettävyyttä ja vastuuvelvollisuutta. (Koivisto ym. 2019, 54 ; AI HLEG 2019, 15, 22–24.)

Algoritmisen oikeudenmukaisuuden varmistamiseen löytyy myös työkaluja, joista yksi on AI Fairness 360, joka on avoimen lähdekoodin työkalukokoelma Pythonille ja R:lle. Kirjasto sisältää erilaisia mittareita data-aineistojen ja mallien oikeudenmukaisuuden testaamiseksi,

selitykset näille mittareille ja algoritmit vinoumien minimoimiseksi koneoppimismalleissa läpi niiden elinkaaren. Kokoelmaan kuuluvat myös erilaiset käyttöohjeet ja tutoriaalit sekä verkkoympäristö, joka johdattaa käsitteisiin ja työkalukokoelman ominaisuuksiin. (Bellamy ym. 2018, 1.)

3.4.4 Vastuuvelvollisuus ja auditoitavuus

Kun tekoälyn luottamuksenarvoisuutta tarkastellaan erityisesti julkisen sektorin näkökulmasta, on tarkasteltava vastuuvelvollisuuden roolia koko tekoälyhankkeen elinkaaren aikana ja kiinnitettävä erityistä huomiota uusiin haasteisiin, joita tekoälyjärjestelmien suunnittelu ja täytäntöönpano aiheuttavat. Sen lisäksi, että automaattiset päätökset eivät ole luonnostaan perusteltuja, on ihmisen otettava moraalinen vastuu päätöksestä, vaikka se olisi koneen tuottama. Ihmisen vastuuttaminen ei kuitenkaan ole aivan yksiselitteistä: tekoälyjärjestelmien kehittäminen on pitkällinen prosessi, jonka eri vaiheisiin osallistuu monen erialan osaajia eri rooleissa. (Leslie 2019, 23–24.)

Se, miksi vastuuvelvollisuutta ylipäätään täytyy käsitellä erillisenä teemana, juontaa Ollilan mukaan juurensa siihen, että tekoäly saatetaan nähdä oliona, toimijana, vaikka vain ihmisillä on riittäviä kyvykkyyksiä moraaliseen vastuullisuuteen. Ollila esimerkiksi puhuu ihmislajin taipumuksesta siirtää vastuuta, ja tässä tapauksessa *”olioittaa algoritmejä*”, sen sijaan, että puhuttaisiin ihmisten vastuusta algoritmien kirjoittajina tai teknologioiden luojina. (Ollila 23.1.2020.)

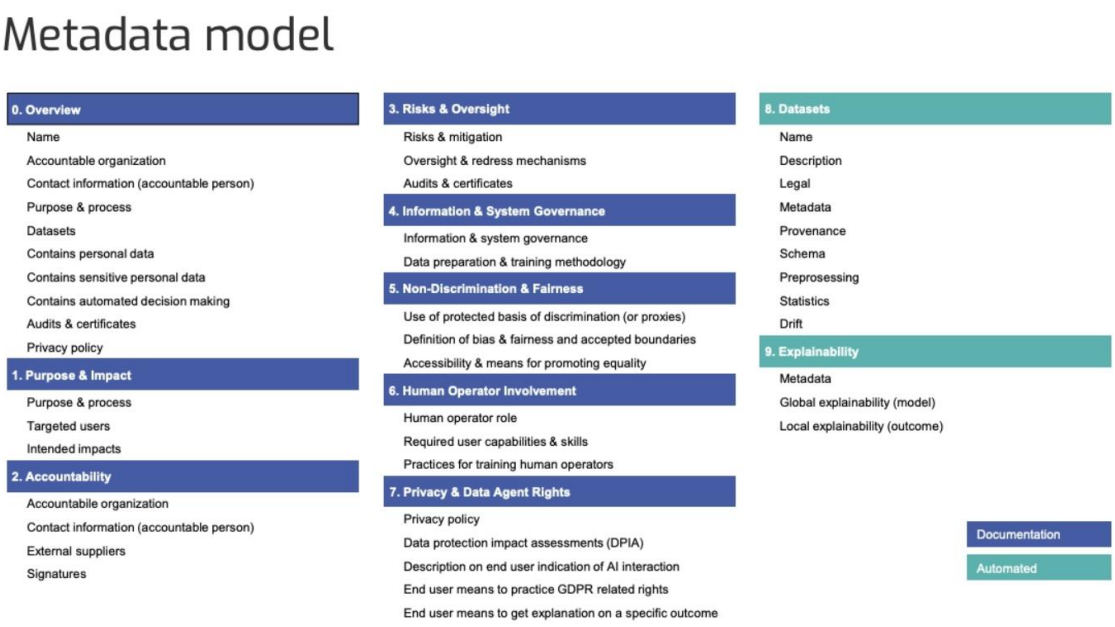
Vastuuvelvollisuus tarkoittaa, että algoritmisesti tuettujen päätösten perusteleminen on vastuutettu jollekin taholle, esimerkiksi tekoälyjärjestelmän luojalle. Käytännössä se kuitenkin edellyttää ehjää ihmisvastuullisten ketjua läpi koko tekoälyhankkeen elinkaaren järjestelmän suunnittelun ensimmäisistä vaiheista algoritmisesti tuotettuihin tuloksiin. Leslien mukaan vastuullisuus edellyttää, että toimivaltaiset ihmiset pystyvät selittämään sekä algoritmisten päätösten sisällön että niiden tuotannon taustalla vaikuttavat prosessit perusteluineen selkeällä, ymmärrettävällä ja johdonmukaisella kielellä. Näiden selitysten ja perusteluiden on perustuttava vilpittömiin, järkeviin ja puolueettomiin syihin, jotka ovat muidenkin kuin teknisten henkilöiden ymmärrettävissä. (Leslie 2019, 24.)

Vaikka vastuuvelvollisen käsitteellä vastataan ennen kaikkea kysymykseen siitä, kuka on vastuussa tuloksesta, auditoitavuuden käsitteellä vastataan kysymykseen siitä, miten tekoälyjärjestelmien suunnittelijoita ja toteuttajia on pidettävä vastuullisina. Tämä vastuuvelvollisuuden näkökohta liittyy sekä järjestelmän suunnitteluun että käyttöön, vastuun osoittamiseen ja tulosten oikeellisuuteen. On varmistettava, että tekoälyhankkeen suunnittelu- ja toteutusprosessin jokainen vaihe on valvottavissa ja tarkastettavissa. Onnistunut tarkastus

edellyttää sen varmistamista, että tekoälyjärjestelmä on kolmannen osapuolen auditoitavissa, että algoritmisten järjestelmien kehittäjistä ja toteuttajista pidetään kirjaa ja että auditoijien saatavilla on tietoa, joka mahdollistaa tekoälyjärjestelmäprosessin asianmukaisuuden ja huolellisuuden seurannan. Tämä edellyttää sellaisten tietojen tallentamista, joiden avulla voidaan seurata datalähteitä ja datan analysointia ja käyttöä keräämisen, esikäsittelyn ja mallintamisen vaiheista koulutukseen, testaukseen ja käyttöönottoon asti. Lisäksi se edellyttää lopputulosten ja positiivisten ja negatiivisten vaikutusten jäljitettävyyttä ja että tiimi antaa mahdollisuuden tutkia ja tarkastella kriittisesti järjestelmän dynaamista toimintaa sen varmistamiseksi, että mallin toimintaa tuottavat menettelyt ja toiminnot ovat turvallisia, eettisiä ja oikeudenmukaisia. Leslie korostaa, että kaikki tekoälyjärjestelmät on luotava tukemaan päästä päähän suunniteltua vastuullisuutta ja auditoitavuutta, mikä edellyttää sekä ihmisen mukana oloa koko suunnittelu- ja toteutusketjun ajan että aktiivista, jatkuvaa monitorointia ja valvontaa. (Leslie 2019, 24–26; AI HLEG 2020, 21.)

Standardointi ja sertifiointi ovat esimerkkejä puolueettomista keinoista, jonka avulla myös suuri yleisö voi varmistaa, että heidän käyttämänsä tekoälyjärjestelmä todella on oikeudenmukainen ja luotettava. Ainakin IEEE:n Ethics Certification Program for Autonomous and Intelligent Systems -hanke on työstänyt kriteereitä ja vaatimuksia tekoälyn läpinäkyvyydelle, syrjimättömyydelle ja vastuille. Myös toimialakohtaisia sertifikaatteja kehitetään muun muassa finanssialalle. AI HLEG:n mukaan *”sertifioinnilla ei kuitenkaan voida koskaan korvata vastuuta. Tästä syystä sitä olisi täydennettävä vastuuvelvollisuusjärjestelmillä, kuten vastuuvapauslausekkeilla sekä tarkistus- ja korjausmekanismeilla.”* Tulevaisuudessa saatetaan nähdä myös nykyisten ISO- ja IEEE P7000 -standardien lisäksi niin sanottuja luotettavan tekoälyn merkkejä, jolla voidaan ilmaista järjestelmän täyttävän määritellyt ja niitä koskevat standardit. (Korhonen 2020; AI HLEG 2019, 26–27.)

Yksi keino ratkoa vastuuvelvollisuuden haastetta käytännössä on hyödyntää kaupallisten toimijoiden ratkaisuja, joista tässä (kuva 7) on yksi esimerkki, jota on käytetty Citizen Trust Through AI Transparency -projektissa. Kelan, Sitran, oikeusministeriön, Helsingin ja Espoon kaupunkien ja Saidot.ai:n yhteishankkeen tavoitteena oli tutkia, mitä kansalaiset haluavat tietää tekoälyn käytöstä, mitä heidän tarvitsee siitä tietää ja millaiseen tietoon he ovat oikeutettuja ja luoda sen pohjalta metamalli, jonka kautta tiedot voisi välittää kansalaisten saatavilla ymmärrettävässä muodossa. Läpinäkyvyyttä parantavaa alustaa voi käyttää myös raportointiin ja valvoa sen avulla vaatimusten täyttymistä, muun muassa vastuuvelvollisen kirjaamista. Ajatuksena on myös, että tulevaisuudessa yritykset voivat hakea alustan avulla kolmansien osapuolten auditointeja ja että alusta voisi tarkistaa sertifikaatit itsenäisesti. (O’Brien 2020; Korhonen 2020; Mattila 2020.)



Kuva 7 Saidot.ai:n esimerkki läpinäkyvyyttä lisäävästä metadatamallista ja tekoälyrekisterialustasta (O'Brien 2020)

3.4.5 Läpinäkyvyys

Läpinäkyvyydellä tarkoitetaan tässä kontekstissa tekoälyjärjestelmän tulkintaa eli kykyä tietää, miten ja miksi malli toimi tietyllä tavalla tietyssä yhteydessä, ja kykyä ymmärtää, mihin päätös tai toiminta perustuu. Käytännössä läpinäkyvyydellä tavoitellaan siis prosessin ja lopputuloksen perusteltavuutta ja sisällön ja lopputuloksen selitettävyyttä. Läpinäkyvyys mahdollistaa myös vastuuvelvollisuuden toteutumisen, parantaa valvontaa, helpottaa riskien ja vaikutusten arviointia, nopeuttaa ongelmanratkaisua ja mahdollistaa asiakkaille ja kansalaisille heidän omien oikeuksiensa turvaamisen. (Leslie 2019, 35–36; Haataja 29.9.2020.)

Läpinäkyvyyden kautta tekoälyjärjestelmästä tulisi saada tietää, mikä sen käyttötarkoitus on, miten järjestelmän onnistumista mitataan, keihin järjestelmällä vaikutetaan, millaisia vaikutuksia järjestelmällä on sekä tiedot liiketoiminnasta ja teknisestä toteutuksesta vastaavista tahoista. Lisäksi tarkempaa tietoa tulisi olla tarvittaessa saatavilla käytetyistä dataseteistä, automaattisen dataprosessoinnin operatiivisesta logiikasta, malliarkkitehtuurista, järjestelmätoimittajista, valvontamekanismeista, ihmisen suorittamasta valvonnasta sekä siihen tarvittavista kyvyistä. Läpinäkyvyyttä tarvitaan myös sen osoittamiseen, miten järjestelmän käytön riskit ja rajoitteet on huomioitu ja miten järjestelmässä on varmistettu oikeudenmukaisuuden toteutuminen ja vinoumien eliminointi. Lisäksi esimerkiksi erillisin doku-

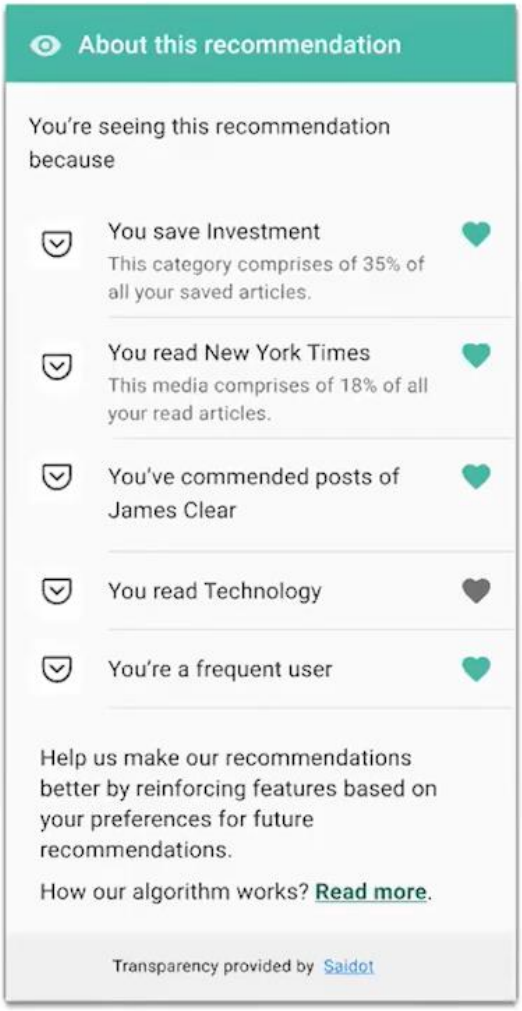
mentein voidaan tuoda esiin, millaiseen viitekehykseen järjestelmä on luotu lakeineen, politiikkoineen ja asetuksineen ja onko järjestelmää auditoitu ja sertifioitu. (Haataja 29.9.2020; Leslie 2019, 39.)

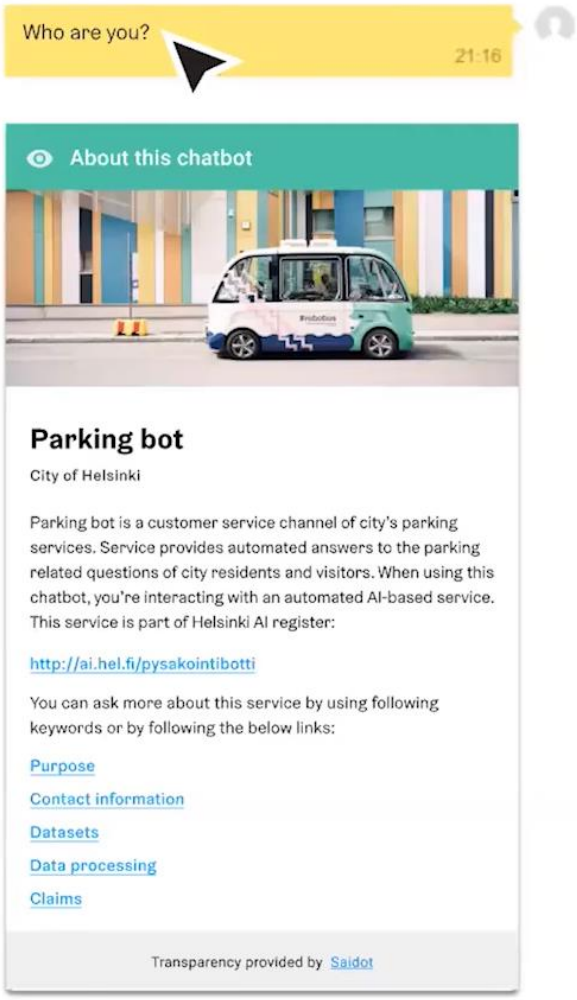
On tärkeää varmistaa, että kaikilla eri tahoilla on tarvitsemansa tiedot arvioidakseen tekoälyjärjestelmän luottamuksenarvoisuutta. Selityksen antamiseksi ja lopputuloksen eettisyyden ja luotettavuuden osoittamiseksi, tulisi mallin toimintaa ja syyseuraussuhteita kyetä avaamaan ymmärrettävällä kielellä. Haataja myös huomauttaa, että läpinäkyvyyden tulisi olla aina kaksisuuntaista, jotta varmistetaan myös kollektiivinen oppiminen. (Haataja 29.9.2020.)

Kela on tuonut julki työpajojensa materiaalin kautta (Leinvuo 2020, 15), mitä läpinäkyvyys yhtenä Kelan kolmesta tekoälyn eettisestä periaatteesta voisi käytännössä tarkoittaa ja millaisia vaikutuksia sillä voisi olla:

- *”Käytetään selkeää ja ymmärrettävää kieltä asiakkaalle: Miksi, missä ja miten tekoälyä on hyödynnetty*
- *Kerrotaan asiakkaalle, mihin algoritmejä käytetään, ja mihin tekijöihin algoritmi perustuu*
- *Mallien dokumentointi, valvonta ja selkeät pelisäännöt < voidaan palata versioon jota päätöksessä on käytetty*
- *Kelassa yhtenevät käytännöt koodin kirjoittamisessa ja jakamisessa*
- *Arvioidaan asiakkaan kokemus läpinäkyvyydestä*
- *Varmistetaan, että tiedonhallinnan periaatteet + eettisyys ovat selkeitä johdolle*
- *Kelan sidosryhmäyhteistyön läpinäkyvyys > tiedon välitys, datan liikkuminen*
- *Open source –tekniikan suosiminen?*
- *Tietosuojavaltuutettu varmistaa että asiakas tietää, mihin tietoihin päätös perustuu*
- *Valtuutetulta mahdollisuus kysyä mitä tietoja tekoäly käyttää itsestä ja mihin tarkoitukseen*
- *Läpinäkyvyyden varmistaminen lisää luottamusta”*

Loppukäyttäjälle läpinäkyvyys voisi puolestaan Haatajan mukaan (29.9.2020) näyttäytyä kuvan 8 mukaisena osallistavana personointina tai mahdollisuutena varmistaa, millaisen tekoälyjärjestelmän kanssa käyttäjä on vuorovaikutuksessa





Kuva 8 Saidot.ai:n esimerkkejä läpinäkyvyydestä käytännössä (Haataja 29.9.2020)

3.4.6 Selitettävyys

Yhteiskunnallisen oikeudenmukaisuuden ja yhdenvertaisuuden sekä objektiivisuuden ja läpinäkyvyyden vaatimukset ovat ehdottomia edellytyksiä, kun tekoälyä aletaan hyödyntää laajemmin. Samalla tiedetään, että ihmisen on vaikea hahmottaa nopeita, monimutkaisia, laskennallisia menetelmiä: esimerkiksi kaikki koneoppimismallit eivät ole luonnostaan läpinäkyviä, intuitiivisia tai helppoja ihmisille ymmärtää. Yksinkertaisiin läpinäkyvyyden haasteisiin voidaan vastata sillä, että perehdytään järjestelmän toimintaan tai viestitään sen toiminnasta paremmin, jos se on mahdollista niin, etteivät esimerkiksi liikesalaisuudet paljastu. Osa haasteista liittyy kuitenkin itse tekoälyjärjestelmiin ja siihen, etteivät edes mallien kehittäjät pysty täysin ymmärtämään ja selittämään mallien toimintaa, jolloin voidaan puhua niin sanotusta mustan laatikon ongelmasta. (Turek s.a.; Rusanen & Koskinen 2018, 48–49, 52–53.)

Mustan laatikon ongelma liitetään usein syviin neuroverkkoihin. Syväoppimisen menestys perustuu siihen, että parhaita hermoverkkoja muutetaan ja mukautetaan entistä parempien luomiseksi, jolloin käytännön tulokset ylittävät ihmisten teoreettisen ymmärryksen. Tämän seurauksena yksityiskohdat koulutetun mallin toiminnasta ovat tyypillisesti tuntemattomia, mustia laatikoita. Mustat laatikot eivät itsessään kerro mitään suunnittelustaan, rakenteestaan tai toteutuksestaan eivätkä perustele antamiaan vastauksia. Mustien laatikoiden osalta tekoälyn toiminta ei ole ymmärrettävää tai läpinäkyvää siis myöskään kehittäjille, koska kehittämiseen ei ole tarjolla riittäviä analyyttisia työkaluja. Tällaisissa tilanteissa, joissa kyse ei ole tiedon tai osaamisen puutteesta, tarvitaan ”*tutkimusta, tutkimuksen menetelmien tarkastelua ja ehkä jopa tekoälysovellusten kehittämiseen liittyvien eettisten ohjeistusten laatimista”*. (Rusanen & Lappi s.a. 2–3; Adadi & Berrada 2018, 52141; Heaven 2020.)

Uusien teknologioiden myötä on jo helpompaa tutkia, mitkä esimerkiksi syötetyn datan tiedoista ovat lopputuloksen kannalta merkityksellisimpiä ja teknologia mahdollistaa kyllä neuroverkkojenkin eri yksiköiden välisen kommunikoinnin tutkimisen mallin sisällä, mutta sellaisenaan tieto ei ole ymmärrettävää. On myös huomioitava, että tällaisia yksiköitä voi suurimmissa neuroverkoissa olla tuhansittain ja yhteyksiä niiden välillä satojatuhansia. (Solita 2019, 21, 23–24.)

Selitettävää tekoälyä (l. explainable AI, XAI) voidaan käsitellä alana, jonka tavoitteena on ymmärtää, miten tekoälyjärjestelmien päätökset tehdään. Se keskittyy ymmärtämään ja selittämään pääasiassa niin sanottuja mustan laatikon päätöksiä ja sitä, millaisia vaiheita ja malleja päätöksentekoon tarvitaan. Kysymykset, joita selitettävä tekoäly pyrkii ratkaisemaan, ovat muun muassa seuraavia: Miksi tekoälyjärjestelmä teki tietyn ennusteen tai päätöksen? Miksi se ei tehnyt jotain muuta? Milloin tekoäly onnistui ja milloin se epäonnistui? Milloin tekoälyyn voi luottaa? Miten tekoäly voi korjata ilmenevät virheet? Toisaalta selitettävän tekoälyn voidaan ajatella olevan myös tapa tai kokoelma menetelmiä ja tekniikoita, joiden avulla selittää tekoälyn, tai laajemmalti automaation, tekemää päätöksentekoa ja sen perusteita. Sen tarkoituksena voidaan ajatella olevan myös sellaisten mallien tuottaminen, joita ihmiset voivat paremmin ymmärtää ja hallita. (Adadi & Berrada 2018, 52140; Schmelzer 2019; Steniche 2019.)

Teknisesti tekoälyn selitettävyys voidaan varmistaa, tai selitettävyyttä voidaan luoda, tekoälykehittämisen kaikissa vaiheissa, jolloin selitettävyys voidaan jakaa kolmeen osaan, jotka ovat mallinnettavissa oleva selitys (pre-modelling explainability), selitettävä mallinnus (explainable modelling) ja mallinnuksen jälkeinen selitettävyys (post-modelling explainability). *Mallinnettavissa oleva selitys* on kokoelma erilaisia menetelmiä, joilla pyritään saamaan pa-

rempi käsitys mallin kehittämiseen käytetystä aineistosta ja se perustuu siihen, että tekoälymallin toimintaa ohjaa suurelta osin data-aineisto, jota käytetään sen kouluttamiseen. (Khaleghi 2019a; Khaleghi 2019b.)

Kun puhutaan *selitettävästä mallinnuksesta*, puhutaan usein mallin valinnan rajoittamisesta niihin malleihin, joita pidetään luonnostaan selitettävinä. Yksi tapa ratkoa selitettävyyden haastetta on käyttää koneoppimisalgoritmeja, jotka ovat luonnostaan selitettäviä ja jotka siten edistävät päätösten jäljitettävyyttä ja läpinäkyvyyttä. Tällaisiksi luonnostaan selitettäviksi malleiksi lasketaan esimerkiksi lineaariset mallit ja päätöksentekopuut. Luonnostaan selitettävän mallin valinta ei kuitenkaan Liptonin ja Khaleghin mukaan takaa selitettävyyttä, koska nämä mallit eivät välttämättä ole läpinäkyviä kaikilla tarkastelun tasoilla: koko tekoälymallin läpinäkyvyyden (l. simuloitavuus), yksittäisten komponenttien, kuten parametrien, läpinäkyvyyden (l. purettavuus) ja koulutusalgoritmin läpinäkyvyyden (algoritminen läpinäkyvyys) osalta. Selitettävyyteen liittyvät algoritmisen läpinäkyvyyden, purettavuuden ja simuloitavuuden lisäksi myös muun muassa mallin koko, laskennan määrä sekä subjektiivisena pidettävä käsite *”kohtuullisessa ajassa”*, jossa ihmisen olisi ylipäätään mahdollista tehdä koneoppimismallin päättelytyö, sekä käyttäjien teknisen lähtötason ymmärtäminen eli tekniikat, joilla tekoäly on selitettävissä myös muille kuin tekoälyyn vihkiytyneille asiantuntijoille. Monimutkaisempien menetelmien, kuten neuroverkkojen, käyttö on usein tehokkaampaa, ne ovat suorituskyvyltään parempia ja antavat tarkempia lopputuloksia, minkä vuoksi myös niiden selitettävyyttä pyritään parantamaan. Ihannetapauksessahan mustan laatikon ongelmaa ei edes synny, jos kehitetään vain malleja, jotka ovat selitettävissä. (Schmelzer 2019; Khaleghi 2019b; Arrieta ym. 2019, 13; Lipton 2017, 4–5, 7.)

Suurin osa tekoälyn selitettävyyteen liittyvästä tutkimuskirjallisuudesta keskittyy jo kehitettyjen mallien selittämisen menetelmiin. *Mallinnuksen jälkeinen selitettävyys* voidaan jakaa neljään tarkastelukulmaan:

- Mikä tuottaa halutun selityksen?
  Kaikki tekijät, joilla on vaikutusta AI-mallin kehitykseen, voivat olla selittäviä tekijöitä eli selityksen tuottajia. Näitä tekijöitä ovat esimerkiksi koulutusdata, parametrien asetukset, optimointialgoritmien valinta ja malliarkkitehtuurin valinta.
- Mitä mallissa selitetään?
  Tekoälyasiantuntijat voivat vaatia mekaanisen selityksen jokaisesta mallin sisällä olevasta komponentista validoidakseen mallin toimintaa, kun taas tekoälyyn vihkiytymättömille voi riittää selitys siitä, miten malli käyttää sille annettuja tietoja ennusteiden tekemiseen sen varmistamiseksi, että malli on puolueeton ja säännöstenmukainen. Selitykset voivat myös vaihdella laajuuden ja monimutkaisuutensa puolesta.
- Miten selitys välitetään käyttäjälle?

Kolmannekseen on eri tapoja välittää selitys käyttäjälle. Selityksiä voidaan luoda myös esimerkiksi mukautettuina visualisointeina, sanallisia selityksinä ja vastakohtaisina selityksinä eli miten erilaiseen lopputulemaan olisi päästy.

- Miten laskennallinen prosessi toimii selityksen saamiseksi?
  Selityksiä voidaan arvioida erilaisilla laajoilla menetelmillä, jotka vaihtelevat pääasiassa mallin sovellettavuuden ja taustalla olevan mekanismin suhteen.

(Khaleghi 2019c.)

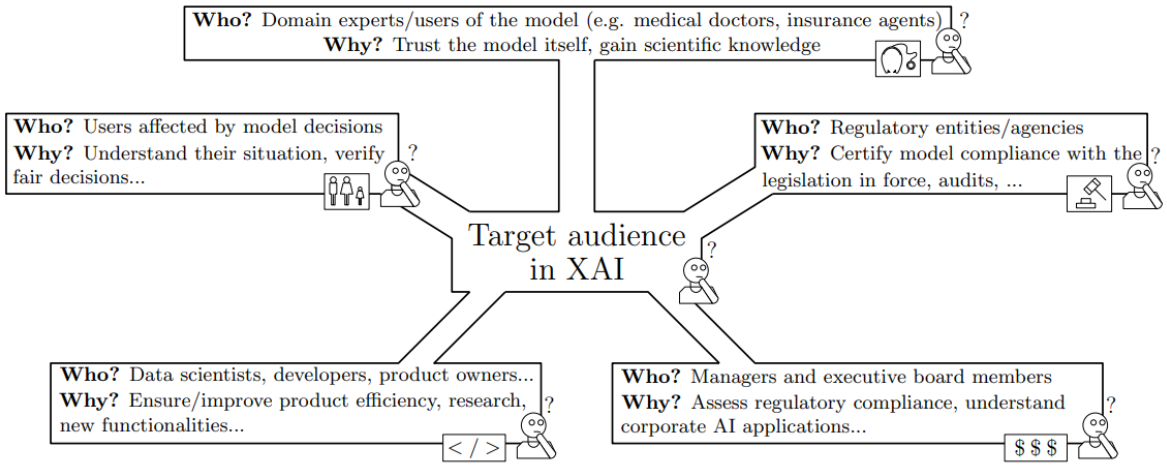
Selitettävyyden ja teknologisten mahdollisuuksien välille pyritään luomaan tasapainoa, mutta perustellusti voidaan myös kysyä, onko tekoälymallin edes mahdollista olla riittävän yksinkertainen, että se voi tarjota ratkaisun läpinäkyvyyden haasteeseen, mutta samalla riittävän tehokas suoriutuakseen monimutkaisesta tehtävästään (Khaleghi 2019b).

Voidaan myös pohtia, onko mahdollinen valinta tehtävä tarkkuuden ja selitettävyyden vai tarkkuuden ja tulkittavuuden välillä. Selitettävyys (explainability) liittyy läheisesti tulkittavuuden (interpretability) käsitteeseen: tulkittavissa olevat järjestelmät ovat selitettäviä, jos ihminen ymmärtää niiden toiminnan. Selitettävyydellä tarkoitetaan enemmän kyvykkyyttä ymmärtää koneoppimisalgoritmien toimintaa kuin sitä, onko malli tulkittavissa eli voiko loppukäyttäjä tutkia, miten syötetty data on matemaattisesti yhteydessä mallin antamaan lopputulemaan eli ovatko sen syy- ja seuraussuhteet havaittavissa ja pystytäänkö ennustamaan, mitä tapahtuu, jos syötettävää dataa tai algoritmin parametreja muutetaan. Tulkittavuus tarkoittaa siis kykyä tulkita algoritmin sisäistä rakennetta, ymmärtää algoritmiä, tietämättä tarkasti, miksi se toimii kuten toimii. Tämä voi auttaa esimerkiksi datatieteilijöitä ja liiketoiminta-analyytikoita keskittymään tarkemmin organisaationsa avainkysymyksiin ja tarpeisiin tekniikan sijaan. Molempia termejä kuitenkin käytetään, myös toistensa synonyymeinä, vaikka näin ei tulisi tehdä. Kolmas termi on vielä ymmärrettävyys: selittäminen riippuu pääasiassa siitä, mitä selitetään (alkuperäinen tekoälymalli) ja miten selitys tehdään (menetelmä), kun taas ymmärtäminen riippuu näiden lisäksi siitä, kenelle selitetään, ja ollakseen selitettävä, on mallin oltava ihmisen ymmärrettävissä, mikä edellyttää ihmiseltä ainakin koneoppimisen sekä ihmisen ja koneen vuorovaikutuksen asiantuntemusta. Selityksen voidaan myös ajatella syntyvän ihmisen ja koneen vuorovaikutuksessa. (Adadi & Berrada 2018, 52140–52141, 52152–52153, 52155; Gall 2018.)

Toisaalta voidaan kysyä myös, onko selitettävyyttä ylipäätään pidettävä edellytyksenä. Tutkimusten mukaan tarve selittää tekoälyjärjestelmiä voi johtua ainakin neljästä, osin päällekkäisestä syystä. Tarve liittyy perustelemiseen, kontrolloimiseen, jatkuvaan parantamiseen ja tutkimukseen. Ensinnäkin selitettävyyden avulla pyritään varmistamaan, ettei tekoälypohjaisia päätöksiä ole tehty virheellisesti. Tällöin tulee tarkastella lopputuloksen sijaan mallin sisäistä toimintaa ja päätöksentekoprosessin taustalla olevaa logiikkaa, jotta myös odotta-

mattomat päätökset voidaan perustella. Samalla varmistetaan, että on olemassa auditoitava ja todistettava tapa puolustaa algoritmin tekemien päätösten oikeudenmukaisuutta ja eettisyyttä. Oikeus selitykseen sisältyy myös GDPR:n vaatimuksiin. Toisekseen järjestelmän ymmärtäminen lisää näkyvyyttä ja mahdollistaa puutteiden, virheiden ja haavoittuvuuksien tunnistamisen ja nopeamman korjaamisen kriittisissä tilanteissa. Kolmannekseen ymmärrettävää ja selittävää mallia on helpompi parantaa, sillä jos käyttäjät tietävät, miksi järjestelmä päätyy määrättyihin lopputulemiin, he tietävät myös, miten siitä voidaan tehdä entistä älykkäämpi. Selitettävyys siis mahdollistaa jatkuvan iteroinnin koneen ja ihmisen välille. Neljännekseen selitysten avulla ihmisellä on mahdollisuus oppia; on toivottavaa, että mikäli kone kykenee tekemään ihmistä paremman tai tarkemman päätöksen, kone kykenisi myös jakamaan tietonsa. (Adadi & Berrada 2018, 52142–52143.)

Arrietan ym. (2019, 7) mukaan selitettävyyttä voi lähestyä myös eri kohderyhmien kautta (kuva 9) ja vaikka tarpeet selitykselle vaihtelevat, ne liittyvät pitkälti joko mallin ymmärtämiseen tai sääntelyn noudattamiseen.



Kuva 9 Selitettävän koneoppimisen eri kohderyhmät (Arrieta ym. 2019, 7)

Kolmas kysymys on, koskeeko selitettävyyden haaste kaikkia käyttötapauksia. Jos mustia laatikoita käytetään peleissä tai elokuvien valitsemiseksi, ei selitettävyydellä ole Heavenin mielestä niin suurta merkitystä kuin jos niitä käytetään lainvalvonnassa, lääketieteellisessä diagnosoinnissa ja autonomisissa autoissa. Tätä käsitystä tukee Bunt & ym.:n tutkimus, jonka mukaan myös läpinäkymättömät tekoälyjärjestelmät voidaan kokea positiivisesti, vaikka merkityksellinen tai helposti saatava selitys puuttuisikin, jos käyttökohde on riskitön. Näissä tilanteissa selitettävyyden kustannukset voisivat tutkimukseen osallistuneiden mielestä olla suurempia kuin niistä saatava lisähyöty. (Adadi & Berrada 2018, 52153–52154; Heaven 2020.)

On olemassa myös tutkimuksia, joiden mukaan selitettävyys ei edes takaa luotettavuutta. Läpinäkyvämpien neuroverkkojen luominen voi johtaa jopa siihen, että luotamme niihin liikaa, mikä vaikeuttaa mallin virheiden havaitsemista ja korjaamista. Heavenin mukaan ihmisten tulisikin uskaltaa olla eri mieltä automatisoidun päätöksen kanssa selityksestä huolimatta. On mahdollista, että jos koneet eivät kykene lunastamaan ihmisten odotuksia kaikissa päätöksentekotilanteissa, ihmiset alkavat vastustaa itse tekniikkaa. Näin on käynyt kasvojentunnistusohjelmien kanssa, joiden käytön lopettamista määritellyissä käyttökohteissa ovat vaatineet jo teknologiayritysten omat työntekijätkin vastoin eettisiä ohjeistuksiaan. (Heaven 2020.)

Selitettävään tekoälyyn liittyy päätöksenteon osalta Robbinsin mukaan myös paradoksi: selitettävä tekoäly on merkityksellistä vain, jos tiedetään, mitkä näkökohdat ovat hyväksyttäviä käsiteltävässä päätöksessä, ja jos ne jo tiedetään, ei ole syytä käyttää tekoälyalgoritmia niiden selvittämiseksi, koska päätöksenteon automatisointi riittää. Robbins myös toteaa, ettei ihmisiltäkään vaadita selitystä jokaiseen päätökseen, eikä edes kykyä selittää, ellei päätös erityisesti sitä vaadi. Tällöinkään annettu selitys ei välttämättä koske kuin lopputulosta. Siksi Robbinsin mielestä olisikin syytä ennemminkin pohtia, mitkä käyttökohteet selitystä edellyttävät. Hän kohdistaa kritiikin kuitenkin ennen kaikkea selitettävyyteen periaatteena kuin itse selitettävyyteen määritellyissä, perustelluissa kohteissa. Selitettävyyden vaade tulisi suunnata yksittäiseen toimintaan tai päätökseen, kontekstiin ja päätöksen negatiivisiin vaikutuksiin, ei koko prosessiin. (Robbins 2019, 495–497.)

3.4.7 Tekninen luotettavuus ja turvallisuus

Teknisesti kestävä ja luotettava tekoäly liittyy oleellisesti vahinkojen välttämisen periaatteeseen ja on turvallinen, tarkka, luotettava ja kestävä (robusti). Tämä edellyttää huolellisen suunnittelun ja riskien minimoimisen lisäksi sen varmistamista, että järjestelmä toimii tarkasti, luotettavasti ja suunnittelijoiden odotusten mukaisesti myös erilaisissa odottamattomissa tilanteissa ja että se kykenee minimoimaan tahattomat ja kohtuuttomat vahingot. Tällaisten turvallisuustavoitteiden mukaisen tekoälyjärjestelmän rakentaminen edellyttää testausta, validointia ja uudelleenarviointia sekä asianmukaisten valvontamekanismien sisällyttämistä tekoälyn toimintaan. Tekoälyn turvaamisen epäonnistuessa tekoäly voi päätyä haitallisiin lopputuloksiin ja yleinen luottamus tekoälyä kohtaan heikkenee. (Leslie 2019, 30; AI HLEG 2019, 20.)

Tekoälyjärjestelmän turvallisuudella tarkoitetaan esimerkiksi useiden eri toiminnallisuuksien suojaamista, kun järjestelmään kohdistuu hyökkäys. Suojatulla järjestelmällä voidaan säilyttää sen muodostavien tietojen eheys. Turvallisuuden varmistamiseen kuuluu myös tekoälyarkkitehtuurin suojaaminen sen osien luvattomalta muuttamiselta tai vahingoittumiselta.

Turvallinen järjestelmä on jatkuvasti toimiva ja valtuutettujen käyttäjien saatavilla ja se pitää luottamukselliset ja yksityiset tiedot turvassa kaikissa olosuhteissa. (Leslie 2019, 30.)

> ”Hyökkäykset voivat kohdistua tietoihin (tietojen korruptointi), malliin (mallivuoto) tai perustana olevaan infrastruktuuriin, sekä ohjelmistoihin että laitteistoihin. Jos tekoälyjärjestelmään tehdään hyökkäys, esimerkiksi kontradiktorinen hyökkäys, tiedot sekä järjestelmän käyttäytyminen voivat muuttua niin, että järjestelmä tekee eri päätöksiä tai lakkaa kokonaan toimimasta. Järjestelmät ja tiedot voivat korruptoitua myös vihamielisten tahallisten hyökkäysten seurauksena tai altistuttuaan odottamattomille tilanteille. Riittämättömät turvallisuusprosessit voivat johtaa myös virheellisiin päätöksiin tai jopa fyysisiin vahinkoihin.” (AI HLEG 2019, 20.)

Sen lisäksi, että ulkoiset uhat voivat olla monenlaisia, tekoälyllä on myös ominaisuuksia, jotka tekevät sen erityisen alttiiksi tietynkaltaisille turvallisuutta uhkaaville tilanteille. Yksi tällainen ominaisuus on tekoälyn kyvyttömyys reagoida odottamattomiin tilanteisiin; tekoälyltä puuttuu ”terve järki”. Tekoäly voi täten tehdä odottamattomia ja vakavia virheitä, koska sillä ei ole valmiuksia kontekstualisoida ongelmia, jotka se on ohjelmoitu ratkaisemaan, eikä määrittää uusia merkityksiä. Lisäksi nämä virheet voivat jäädä selittämättömiksi, kun otetaan huomioon niiden matemaattisten rakenteiden moniulotteisuus ja laskennallinen monimutkaisuus. (Leslie 2019, 30.)

Koneoppimismalleihin kohdistuvat hyökkäykset muokkaavat usein haitallisesti tekoälylle syötettyjä tietoja – usein huomaamattomilla tavoilla – aiheuttaakseen virheellisiä luokitteluja tai ennustuksia. Vaikka huomio kiinnittyy usein syväoppimiseen kohdistuviin hyökkäyksiin, hyökkäykset voivat kohdistua myös yksinkertaisempia koneoppimistekniikoita hyödyntäviin sovelluksiin kuten roskapostin suodatukseen ja haittaohjelmien havaitsemiseen. (Leslie 2019, 32.)

Mikäli tekoäly sallittaisiin hallinnollisessa päätöksenteossa, koskisi sitä myös viranomaisten tietojärjestelmien tietoturvallisuuden ja vaatimuksenmukaisuuden arviointia koskeva sääntely (Vainio ym. 2020, 52):

> "Julkisen hallinnon tiedonhallinnasta annetussa laissa (906/2019) säädetään muun muassa tietojärjestelmien tietoturvallisuuden varmistamisesta. Lain perusteella viranomaisen tulee esimerkiksi seurata toimintaympäristönsä tietoturvallisuuden tilaa, testata tietojärjestelmien vikasietoisuutta ja ottaa käyttöön tiettyjä tietoturvallisuustoimenpiteitä. Näiden toimenpiteiden toteutuminen on varmistettava myös, mikäli yksiköissä tehdään tietojärjestelmiä koskevia hankintoja. Laissa viranomaisten tietojärjestelmien ja tietoliikennejärjestelyjen tietoturvallisuuden arvioinnista (1406/2011) säädetään viranomaisten tietojärjestelmien ja tietoliikennejärjestelyjen tietoturvallisuuden arvioinnista. Arviointi voidaan toteuttaa mainitun lain mukaisella menettelyllä tai käyttäen tietoturvallisuuden arviointilaitoksista annetun lain (1405/2011) mukaista arviointilaitosta.”

Se, mihin laki ei tällä hetkellä riittävästi ota kantaa, on, kenen vastuulla on päätöksentekosääntöjen hyväksyminen, järjestelmän testaus, hyväksyntä, arviointi ja valvonta (Vainio ym. 2020, 52).

Tämä on ongelmallista esimerkiksi siksi, että esimerkiksi AI HLEG on todennut, että perinteinen testaus ei riitä tekoälyjärjestelmien testaamiseen, koska ne ovat kontekstisidonnaisia ja saattavat reagoida yllättävällä tavalla odottamattomiin tilanteisiin tai todelliseen dataan. Siksi mallia tulee seurata tarkasti sekä koulutuksen että käyttöönoton aikana ja siksi on todennettava, että se toimii elinkaarensa kaikissa vaiheissa, etenkin käyttöönoton jälkeen. Testauksessa voidaan hyödyntää myös niin sanottua hyökkäystestausta, jossa järjestelmästä pyritään tarkoituksenmukaisesti löytämään ne virheet ja heikkoudet, joilla järjestelmä saadaan murrettua. Joka tapauksessa *”testauksen ja validoinnin olisi katettava tekoälyjärjestelmän kaikki osatekijät, kuten tiedot, esikoulutetut mallit, ympäristöt ja koko järjestelmän toiminnan. Sen suunnittelevan ja toteuttavan ryhmän olisi koostuttava mahdollisimman erilaisista ihmisistä.”* Lisäksi tekoälyjärjestelmiin on sisällytettävä suojatoimia, joilla voidaan estää tai lieventää väärinkäyttöä. Järjestelmät voidaan esimerkiksi asettaa pyytämään ihmiseltä tarvittaessa vahvistus tai siirtymään tilastollisista menettelyistä säästöpohjaisiksi. (AI HLEG 2019, 20, 26.)

3.4.8 Riskienhallinta ja vaikutusten arviointi

*”Lisäksi olisi otettava käyttöön prosesseja, joilla selvitetään ja arvioidaan tekoälyjärjestelmien käyttöön liittyviä mahdollisia riskejä eri soveltamisaloilla. Tarvittavien turvallisuustoimenpiteiden taso riippuu tekoälyjärjestelmän aiheuttaman riskin suuruudesta, joka puolestaan riippuu järjestelmän kyvyistä. Jos voidaan ennakoida, että kehitysprosessi tai itse järjestelmä aiheuttaa erityisen suuria riskejä, on erittäin tärkeää, että turvallisuustoimenpiteitä laaditaan ja testataan ennakoivasti.” (*AI HLEG 2019, 20.)

Riskienhallinnalla tarkoitetaan sekä kykyä tunnistaa ja raportoida tekoälyjärjestelmän lopputuloksiin vaikuttavista toimista että vastata lopputuloksista aiheutuviin seurauksiin. Riskien tunnistaminen, minimointi, tekoälyjärjestelmien kielteisten vaikutusten arviointi ja dokumentointi on erityisen tärkeää niiden ihmisten kannalta, joihin tekoälyjärjestelmät ensisijaisesti vaikuttavat. Edellisissä kappaleissa esiteltyjen tekoälyn edellytysten välille voi syntyä jännitteitä, jotka edellyttävät kompromissien tekoa. Näitä tehtäessä on punnittava erityisen tarkkaan eri vaihtoehtojen turvallisuuteen ja eettisyyteen liittyviä riskejä. (AI HLEG 2020, 21.)

Anttisen & Lohilahden tutkimuksessa EK:n alaisilta yrityksiltä kysyttiin, miten he ovat minimoineet tekoälyn riskejä. Keinoiksi mainittiin muun muassa riskienhallintaprosessien tiukkojen vaatimusten noudattaminen, tietojenkäsittelyn lokitus, seuranta ja puuttumisen periaate. Kehitysprosesseissa yrityksillä oli tarkistuspisteitä riskien tunnistamiseksi ja välttämiseksi sekä validointimenetelmiä. Yrityksissä myös työskentelee henkilöitä, joiden päätteh-

tävä on riskienhallinta ja oikein toimimisen valvonta. Keinoiksi mainittiin myös mahdollisimman luotettavien palveluntarjoajien käyttö sekä tekoälyratkaisujen ja IT-kumppaneiden keskittäminen. Tekoälyjärjestelmää myös huollettiin ja päivitettiin ja niitä varten oli olemassa erilaisia varajärjestelmiä. Tärkeäksi koettiin tekoälyn testaaminen ensin ammattilaisen työparina ja työntekijöiden kouluttaminen. Yritykset halusivat myös kyetä ymmärtämään ja arvioimaan algoritmiensä logiikkaa ja pitäytyä niissä, jotka ovat selitettäviä. (Anttinen & Lohilahti 2019, 33–34.)

Yksi kulma lähestyä vaikutusten ja riskien arviointia on Kanadan julkisen sektorin malli, joka arvioi päätöksenteon vaikutuksia ja haittoja ja asettaa niille eritasoisia hallinta-, raportointi-, auditointi- ja valvontavaatimuksia. Kanadan malli on neljätasoinen: tason I päätökset eivät juuri vaikuta yksilöiden tai yhteisöjen oikeuksiin, terveyteen, hyvinvointiin, taloudellisiin etuihin tai ekosysteemin kestävyyteen. Tasolla II vaikutukset ovat todennäköisesti lyhytkestoisia. Tasolla III vaikutukset ovat jatkuvia ja vaikeita peruuttaa ja tasolla IV täysin peruuttamattomia. (Kääriäinen ym. 2018, 19; Government of Canada 2019.)

Opinnäytetyön kirjoitushetkellä Kanadalla on julkinen beta-versio tasoluokituksen pohjalta luodusta AIA-työkalusta (Algorithmic Impact Assessment), jonka avulla on mahdollista arvioida ja tarkastella tekoälyjärjestelmän vaikutustasoa, tasonmukaisia vaatimuksia ja keinoja vähentää riskejä. Viitekehys on luotu hahmottamaan paremmin automaattisiin päätöksentekojärjestelmiin liittyviä riskejä ja auttamaan niiden minimoimisessa. (Government of Canada 2019.)

AIA:n (Government of Canada 2020) kysymykset liittyvät projektin tietoihin, tavoiteltaviin hyötyihin, riskiprofiiliin, teknologiaan ja käyttötarkoitukseen, algoritmin ominaisuuksiin, päätöksenteon kohteeseen, vaikutusten arviointiin, datalähteisiin, konsultointitarpeeseen, datan laatuun, menetelmälliseen oikeudenmukaisuuteen ja yksityisyyteen. Lopuksi työkalu pisteyttää 13-sivuisen kysymyslistan vastaukset ja näyttää yksityiskohtaisesti, mikä oli minkäkin vastauksen painokerroin ja vaikutus suosituksiin. Vastausten perusteella työkalu suosittelee tasokohtaisesti esimerkiksi, minkä tahon tulisi suorittaa vertaisarviointi, millaisia selityksiä järjestelmän tulee tarjota, mitä testauksessa, koulutuksessa ja valvonnassa on huomioitava, kenen hyväksyntää järjestelmä edellyttää ja miten ihmisen tulee olla mukana päätöksentekoprosessissa.

Kanadan mallista poiketen Euroopan komissio on lähestynyt riskejä lähinnä erottamalla ”suuririskiset” kohteet muista tekoälyjärjestelmistä, vaikka toteaakin samalla, että *”riskiperusteinen lähestymistapa on tärkeä sen varmistamiseksi, että sääntelytoimet ovat oikeasuhteisia”.* Suurella riskillä komissio viittaa sellaisiin kohteisiin, jotka voisivat merkittävästi

uhata joko turvallisuutta, kuluttajien oikeuksia tai ihmisten perusoikeuksia. Komissio luokittelee suuririskisiksi kohteiksi sellaiset tekoälyjärjestelmät, jotka liittyvät sekä suuririskisenä pidettävään alaan että suuririskisenä pidettävään käyttötarkoitukseen. Näiden lisäksi, ja näistä säännöistä poiketen, suuririskisenä voidaan pitää myös rekrytointiprosesseja ja biometrisia etätunnistusmenetelmiä (esimerkiksi kasvojentunnistusta väkijoukosta vertaamalla tunnisteita tietokantadataan). (Euroopan komissio 2020a, 19, 23.)

3.4.9 Yhteenveto: Luottamuksenarvoinen tekoäly

On ymmärrettävä, että luottamuksenarvoinen tekoäly on uskomattoman kompleksinen ja alati muuttuva kenttä, joka koostuu toisiaan täydentävistä ja toisaalta keskenään ristiriidassa olevista vaatimuksista, niin juridisista, eettisistä kuin sosiaalisista normeista, turvallisuuteen liittyvistä kysymyksistä, teknisistä ohjeista ja parhaista käytännöistä, kattavasta vaikutusten ja riskien arvioinnista, valvonnasta, eri näkökulmia edustavista tahoista ja yleisestä yhteiskunnallisen hyväksyttävyyden tasosta, jota vielä ollaan määrittelemässä. Mittelstadtin (2019, 12) mukaan se myös heijastaa kaikkia niitä haasteita, mitä yhteiskunnassa tapahtuu: "*AI ethics is effectively a microcosm of the political and ethical challenges faced in society.*"

Parhaiten tämän kaiken on mielestäni kiteyttänyt 84:n eri tahon tekoälyn eettisiin periaatteisiin tutustuneet Jobin, Ienca & Vayena (2019, 8–9), jotka ovat koonneet yhteen (taulukko 5), mistä tekoälyn etiikassa on kyse ja mitä siinä on huomioitava, keitä aihe koskee ja miten eettistä tekoälyä rakennetaan:

Taulukko 5. Eettisen tekoälyn ulottuvuudet (Jobin ym. 2019, 8–9)

Question addressed	Thematic family	Themes
What?	**Ethical Principles & Values**	**Ethical Principles** I. Beneficence II. Non-maleficence III. Trust IV. Transparency & Explainability V. Freedom and autonomy (incl. consent) VI. Privacy VII. Justice, Fairness & Equity VIII. Responsibility & Accountability IX. Dignity X. Sustainability XI. Solidarity
	Technical and methodological aspects	**Specific functionalities** I. Feedback & feedback-loop II. Decision-making

		I. Impact II. Goals/Purposes/Intentions III. Public opinion IV. Risk evaluation & mitigation (duplicate in Risks) V. Monitoring/Precaution VI. Future of work
Who?	**Design & development**	I. Industry II. AI researchers III. Designers IV. Developers
	Users	I. End users II. Organisations III. Public sector actors IV. Military V. Communities
	Specific stakeholders	I. Ethical and/or auditing committees II. Government III. Policy makers IV. Researchers & scientists V. Vulnerable groups & minorities
How?	**Social engagement**	I. Knowledge commons II. Education & training III. Public deliberation & democratic processes IV. Stakeholder involvement & partnerships
	Soft policy	I. Standards II. Certification III. Best practices IV. Whistleblowing
	Economic incentives	V. Business model & strategy VI. Funding & investments VII. Taxes/taxation
	Regulation & audits	VIII. Laws & regulation (general) IX. Data protection regulation X. IP law XI. Human rights treaties XII. Other rights & laws XIII. Audits & auditing
		Data & datasets I. Data origin/input II. Data use III. Metadata IV. Algorithms
		Methodological challenges I. Methodology II. Metris & measurements III. Tests, testing IV. Ambiguity & uncertainty V. Accuracy VI. Reliability VII. Evidence and validation VIII. Black-box (opacity) IX. Data security X. Quality (of data/system/etc.)
	Impact	**Benefits** I. AI strengths, advantages II. Knowledge III. Innovation IV. Enhancement
		Risks I. Risks II. Malfunction III. Misuse & dual-use IV. Deception V. Discrimination (duplicate in Justice&Fairness) VI. Surveillance VII. Manipulation VIII. Arms race
		Impact assessment

3.5 Luottamuksenarvoisuuden rakentaminen yrityksissä

> ”Ethics is not an outcome but it's about thinking what is right and what is wrong and acting accordingly. AI creation becomes ethical when AI systems are designed and built in a mindful way, making sure that essential questions are raised and that people who should be included are included. Ways of working should focus on collaboration through asking and answering questions about the system being built. Documentation should be something that happens mostly during the process – increasing the understanding of everyone involved and leading to better, more mindful design and implementation decisions. When everyone is actively thinking about consequences, the work becomes consequence-driven.” (Saidot 2019.)

Lainsäädäntö ja korkean tason eettisten periaatteiden ja vaatimusten luominen on siis vasta alku, eivätkä eettisen tekoälyn avaimet ole vain lainsäätäjien käsissä vaan niiden ihmisten, jotka tekoälyjärjestelmiä luovat (Saidot 2019).

Perinteisesti erilaisia ammatillisia normeja ovat luoneet eri ammattijärjestöt, mutta nykyään erilaisia koodistoja ovat alkaneet tuottaa myös ei-perinteiset toimijat, kuten akateemiset yhteisöt, kansalaisjärjestöt ja yritykset. Lisäksi on alettu nähdä itse ammatinharjoittajien vaatimia normeja, kun työntekijät ovat nousseet vaatimaan muutoksia yrityksen toimintaan. Myös osakkeenomistajat ovat alkaneet vaatia yrityksiä noudattamaan periaatteitaan. Institutionaalisen tai hallinnollisen rakenteen muutoksesta huolimatta periaatteet ovat kuitenkin pysyneet samankaltaisina: niillä pyritään vaikuttamaan ammatinharjoittajien toimintaan ja heidän luomien tuotteiden ja palveluiden yhteiskunnallisiin vaikutuksiin. (Gasser & Schmitt 2019, 5, 14, 23.)

Tässä luvussa keskitytään tarkastelemaan yritysten yhteiskuntavastuuta, yritysten, kansalaisten ja kuluttajien suhtautumista tekoälyn eettisyyteen, eettisen tekoälyn arvoa, eri tahojen luomia tekoälyn eettisiä periaatteita ja toimia niiden varmistamiseksi.

3.5.1 Yritysvastuullisuus

Ojasen ym. (2019, 25) mukaan tekoälyn etiikka on osa yritysten yhteiskuntavastuuta. Yritysten yhteiskuntavastuu puolestaan on osa yritysvastuuta, josta puhuttaessa puhutaan usein yrityksen compliancesta. Aiemmin compliancella käsitettiin kapea-alaisemmin lakien, sääntöjen ja määräysten *vaatimustenmukaisuutta* (jäljempänä käytetään sanaa compliance), mutta nykyään compliancen käsitteeseen liittyy vahvemmin myös eettisyys, joka tuo vaatimustenmukaisuuteen tiukemmin myös yritysten arvot ja yrityskulttuurin. *”Enää ei kuitenkaan riitä, että toiminnassa otetaan huomioon ainoastaan lakipykälien asettamat vaatimukset, vaan toiminnan tulee mukailla myös ulkopuolisten tahojen asettamia moraalisia ja eettisiä vaatimuksia.”* Näiden lisäksi compliance koostuu yritysten itse laatimista, aika- ja paikkasidonnaisista ohjeista ja säännöistä, joita edellytetään sekä omalta henkilöstöltä että

yrityksen kumppaneilta. Compliancen voidaan siis katsoa olevan niin lakien, eettisten ohjeiden kuin yrityksen arvojenkin noudattamista ja näiden mukaisten päätösten tekemistä arjessa. (Ratsula 2016, 12-13.)

Compliance-rikkomusten seuraukset voivat olla moninaisia, mutta Ratsula (2016, 15) on listannut niistä seuraavia: vahingot julkisuuskuvalle, yrityksen arvolle tai osakekurssille, asiakkaiden, henkilöstön tai sijoittajien menettäminen, taloudelliset vaikutukset sakkojen, tulonmenetysten, vahingonkorvausvelvoitteiden, rahoituksen vähenemisen tai liiketoimintakieltojen muodossa, sekä yritysjohdon juridisen vastuun realisoituminen ja kriisinhallinta liiketoiminnan kustannuksella. Compliancen tulee perustua kunkin yrityksen omiin arvoihin ja riskiarviointeihin, mutta usein riskit liittyvät kuitenkin *”kilpailuoikeudellisiin kysymyksiin, lahjontaan ja korruptioon, eturistiriitatilanteisiin, alihankintaketjun hallintaan sekä työympäristöön ja ihmisoikeuksiin liittyviin kysymyksiin”* (Ratsula 2016, 19).

Complianceen kuuluu osana niin sanonut code of conduct -ohjeet. Eettisillä liiketapaperiaatteilla (l. code of conduct, liiketoimintaperiaatteet) – joiden mahdollisiksi täydentäjiksi tekoälyn eettiset periaatteet itse miellän – on ainakin neljä tehtävää. Ensinnäkin ne konkretisoivat henkilöstölle, miten yrityksessä toimitaan, mitä sallitaan ja miten yrityksen arvojen tulee arjessa näkyä. Toisekseen eettisen toiminnan seuraaminen, mittaaminen ja arvioiminen ja niiden perusteella sanktioiminen edellyttää, että periaatteista on viestitty ja ne on kirjattu. Periaatteiden avulla viestitään asiakkaille, yhteistyökumppaneille ja muille sidosryhmille, millaista toimintaa yrityksessä edellytetään. Lisäksi periaatteiden avulla voidaan edistää eettistä keskustelua ja etiikan tuomista liiketoiminnan arkeen ja kulttuuriin. Käytännön varmistaminen edellyttää, että periaatteista on edelleenjohdettu syventäviä ohjeita ja että lupauksista ollaan käytännössä vastuussa. (Ratsula 2016, 18, 48–49, 55–56.)

Anttisen ja Lohilahden tutkimuksen mukaan (2019, 37, 48) yritysten laajempien eettisten koodistojen, liiketapaperiaatteiden, sekä tutkimustyön eettisten periaatteiden, katsottiin ohjaavan myös tekoälyn käyttöä silloin, kun yritys ei ollut luonut tekoälylle erikseen omia eettisiä ohjeita.

3.5.2 Tekoälyetiikan normatiivinen ydin

Eettisiä periaatteita voidaan tarkastella eri tasoilla ja eri näkökulmista. Leikas (Koivisto ym. 2019, 13–14) on listannut jo 12 vuotta sitten seuraavia teknologian etiikan periaatteita käyttäjän, kehittäjän ja yhteisön näkökulmista. Koiviston ym. (2019, 13–14) kattava listaus osoittaa myös sen, että minkäänlaisessa eettisessä tyhjiössä myöskään tekoälyn periaatteita ei luoda:

> ”Teknologian käyttäjän näkökulmasta:
>
> ihmisarvo (dignity), loukkaamattomuus (integrity), oikeuksien kunnioittaminen (respect for rights), itsemääräämisoikeus (autonomy), tietoinen suostumus (informed consent), oikeus kieltäytyä (right to decline), luottamus (trust), pätevyys (competence), yhdenvertaiset mahdollisuudet kaikille, tasa-arvoisuus (democracy), identiteetti (identity), käyttäjän osallistuminen (participation), käyttäjän suojaaminen (protection of user), valvonta (surveillance), turvallisuus (safety), saavutettavuus (access), vahingoittaminen (do no harm), valinnanvapaus (choice), vapaaehtoisuus (voluntariness) ja yksityisyyden suoja (privacy).
>
> Yhteisön näkökulmasta:
>
> tasapuolinen hyödyn jakautuminen (equal benefit), kulttuurinen moninaisuus, yhteistyö (cooperation), yhteiset sopimukset (conventions), syrjintä (freedom from bias) sekä teknologian sosiaalinen ja yhteiskunnallinen vaikutus (social impact of technology and role in the society).
>
> Teknologian kehittäjän näkökulmasta:
>
> luotettavuus (reliability), valvonta (surveillance), turvallisuus (security), yksimielisyys (agreement), pätevyys (competence), vastuullisuus (accountability), tekijänoikeuksien kunnioittaminen (respect for intellectual property rights) ja ymmärrys (comprehension).”

Viime vuosina ainakin 84 eri kansainvälistä tahoa on julkaissut tekoälyn eettiset periaatteensa, joiden tarkoituksena on ollut keskittää julkista keskustelua yhteisiin huolenaiheisiin sekä lisätä tietoisuutta eettisistä kysymyksistä ja haasteista (Mittelstadt 2019, 1).

Kuva 10 Tutkitut eettiset periaatteet aikajanalla (Fjeld ym. 2020, 18–19)

Berkman Klein Center tutki kymmenien eri toimijoiden eettisiä periaatteita (kuva 10) ja huomasi, että periaatteissa toistui kahdeksan niin selvää teemaa, että niistä voidaan puhua tekoälyn periaatteiden normatiivisena ytimenä (Fjeld ym. 2020, 4–5).

Fjeld:n ym. (2020, 4–5, 21, 29, 37, 41, 47, 53, 56, 60) löytämät teemat alateemoineen olivat:

- Yksityisyys
  (Privacy: Consent, Control over the Use of Data, Ability to Restrict Pro cessing, Right to Rectification, Right to Erasure, Privacy by Design, Recom mends Data Protection Laws)
- Vastuu(velvolli)suus
  (Accountability: Verifiability and Replicability, Impact Assessments, Environmental Responsibility, Evaluation and Auditing Requirement, Creation of a Monitoring Body, Ability to Appeal, Remedy for Automated Decision, Liability and Legal Responsibility, Recommends Adoption of New Regulations)
- Turvallisuus
  (Safety and security: Security by Design, Predictability)
- Läpinäkyvyys ja selitettävyys
  (Transparency and explainability: Open Source Data and Algorithms, Open Government Procurement, Right to Information, Notification When AI Makes a Decision about an Individual, Notification when Interacting with AI, Regular Reporting)
- Oikeudenmukaisuus ja syrjimättömyys
  (Fairness and Non-discrimination: Non-discrimination and the Prevention of Bias, Representative and High Quality Data, Equality, Inclusiveness in Impact, Inclusiveness in Design)
- Ihminen kontrolloi
  (Human control of technology: Human Review of Automated Decision, Ability to Opt out)
- Ammatillinen vastuu
  (Professional Responsibility: Accuracy, Responsible Design, Consideration of Long Term Effects, Multi-stakeholder Collaboration, Scientific Integrity)
- Inhimillisten arvojen edistäminen
  (Promotion of Human Values: Human Values and Human Flourishing, Access to Technology, Leveraged to Benefit Society)

Samansuuntaiseen tutkimustulokseen päätyivät myös Jobin ym. (taulukko 6) ja Stanfordin yliopiston HAI-instituutti (Human-Centered Artificial Intelligence Institute) omissa raporteissaan: kolme useimmin tekoälyn periaatteissa esiintyvää teemaa olivat oikeudenmukaisuus, selitettävyys ja läpinäkyvyys (Jobin ym. 2019, 7; Perrault ym. 2019, 7, 149).

Oman analyysini perusteella myös esimerkiksi AI HLEG:n periaatteet korreloivat hyvin löydetyn tekoälyn normatiivisen ytimen kanssa (kts. lisää liite 5a/b).

Taulukko 6. Tekoälyohjeissa esiintyneet eettiset periaatteet (Jobin ym. 2019, 7)

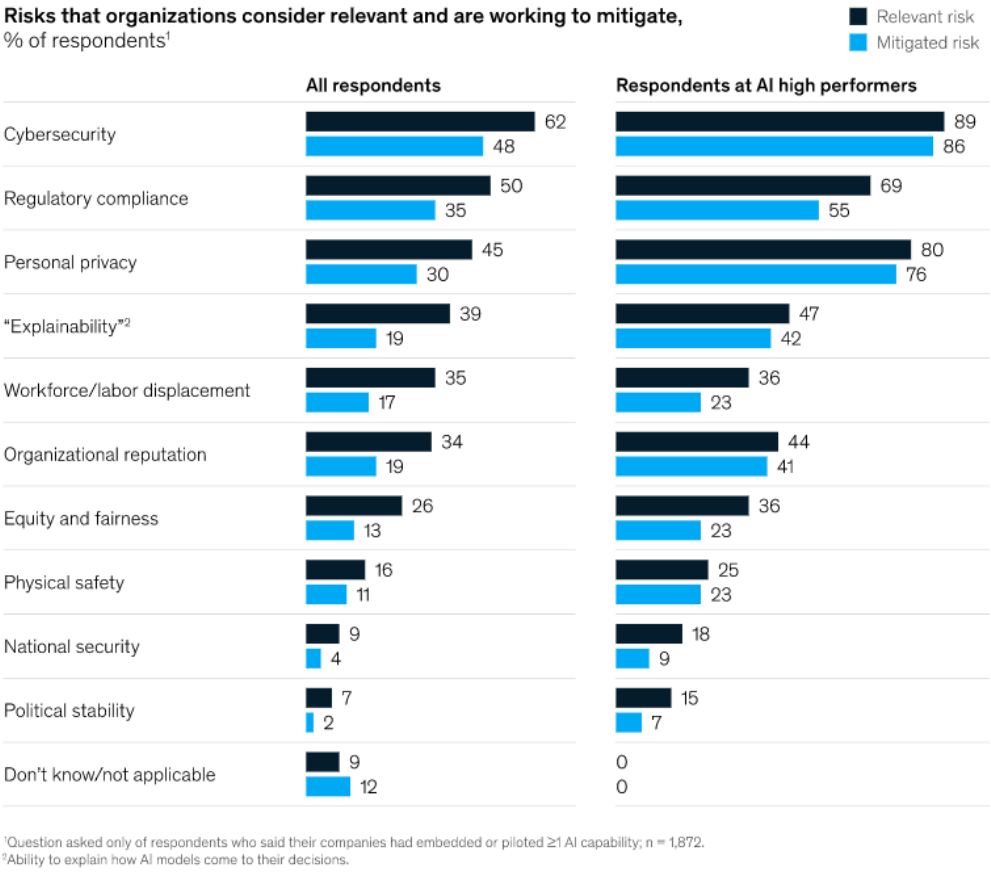
Ethical principle	**Number of documents**	**Included codes**
Transparency	73/84	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice & fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-)bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	60/84	Responsibility, accountability, liability, acting with integrity
Privacy	47/84	Privacy, personal or private information
Beneficence	41/84	Benefits, beneficence, well-being, peace, social good, common good
Freedom & autonomy	34/84	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28/84	Trust
Sustainability	14/84	Sustainability, environment (nature), energy, resources (energy)
Dignity	13/84	Dignity
Solidarity	6/84	Solidarity, social security, cohesion

Berkman Klein Center:n tutkimuksen mukaan eri sektoreiden intressit poikkesivat kuitenkin hieman toisistaan: kuvan 11 perusteella kansalaisyhteiskunnalle tärkeintä oli yksityisyyden, läpinäkyvyyden ja selitettävyyden toteutuminen, yksityistä sektoria kiinnostivat esimerkiksi ammatillinen vastuu, turvallisuus ja vastuullisuus ja hallintoa muun muassa oikeudenmukaisuuden toteutuminen ja se, että ihminen on ohjaksissa (Fjeld ym. 2020, 8–9).

Kuva 11 Tekoälyn normatiivinen ydin (Fjeld ym. 2020, 8–9)

3.5.3 Eettisen tekoälyn arvo ja yritysten toimet

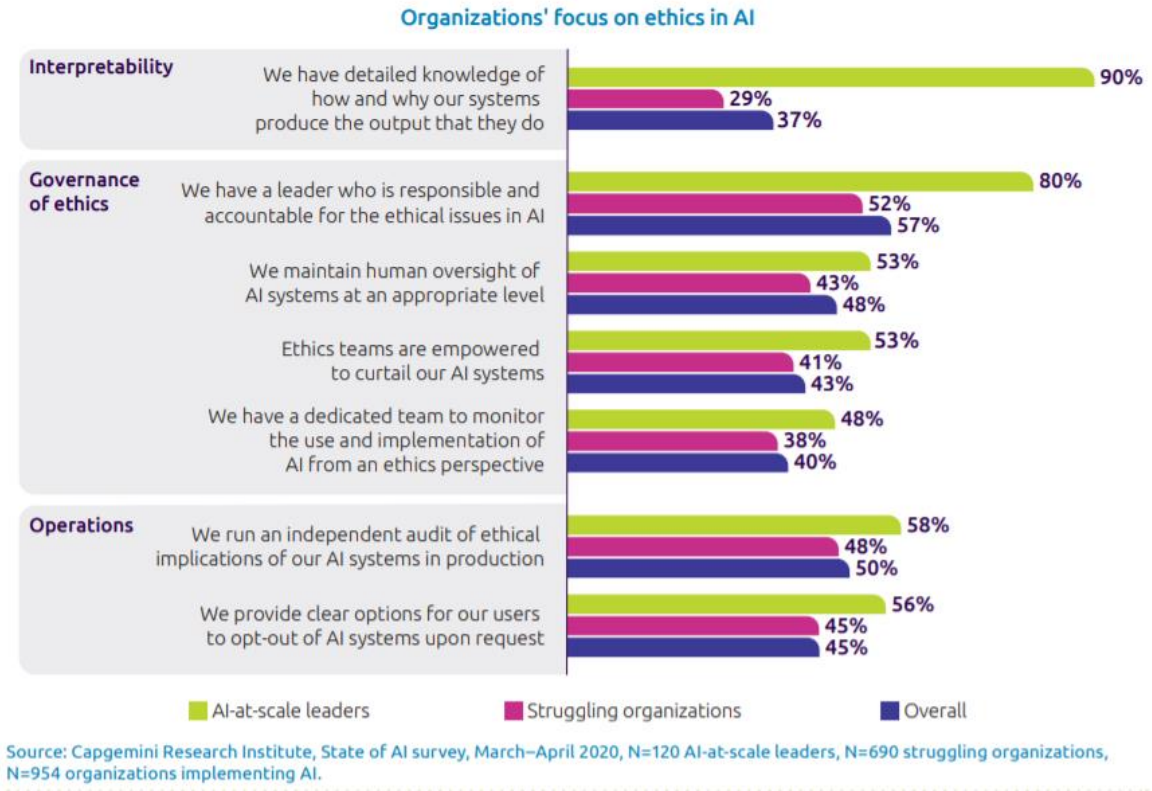
Jaetut periaatteet eivät kuitenkaan välttämättä korreloi yritysten toimien tai riskiarvion kanssa ainakaan globaalisti, sillä esimerkiksi oikeudenmukaisuuteen liittyviä tekoälyriskejä pyrkii McKinseyn (2019) tutkimuksen mukaan taklaamaan vain 13% tekoälyä hyödyntävistä yrityksistä ja selitettävyyttäkin vain 19%. Tämä selittynee sillä, että yritykset itse pitävät merkittävimpinä tekoälyn riskeinä tietoturvaa, compliancea sekä yksityisyydensuojaa. McKinseyn tutkimuksesta (2019) ei käy ilmi, mikä selittäisi sen, että organisaation tietoturvariskiä pyritään vähentämään melko hyvin suhteessa riskin merkittävyyteen, mutta näin ei ole kuitenkaan esimerkiksi ihmistyövoiman korvaamisen, selitettävyyden ja oikeudenmukaisuuden osalta (kuva 12).



Kuva 12 Organisaatioiden merkittävimmät riskit ja niiden mitigointi (McKinsey 2019)

Yksi syy voi olla se, että johtajat ovat tutkimusten mukaan ylipäätään enemmän huolissaan datan laadusta, osaamisen puutteesta, pääoman tuottoasteesta ja hyödystä kuin esimerkiksi tekoälyn etiikasta (Cognilytica Research 2020, 8). Samansuuntaiseen lopputulemaan päätyi tutkimuksissaan myös O'Reilly, jonka mukaan tekoälyn suurimmat haasteet yrityksissä eivät liittyneet complianceen vaan tekoälytarpeen tunnistamiseen (22%), sopivien käyttökohteiden löytämiseen (20%), osaamisen puutteeseen (18%) ja datan puutteeseen tai datan laatuun liittyviin seikkoihin (16%) (Magoulas & Swoyer 2020).

Tutkimukset osoittavat myös, että eri yritysten panostukset tekoälyn etiikkaan ovat osin hyvin eritasoisia (kuva 13). Esimerkiksi vain 37% kaikista Capgeminin kattavaan tutkimukseen vastanneista yrityksistä tietää, miten ja miksi yrityksen tekoälyratkaisut päätyvät lopputuloksiinsa. (Capgemini Research Institute 2020, 21).



Kuva 13 Organisaatioiden panostus tekoälyn etiikkaan (Capgemini Research Institute 2020, 21)

Syyt jo tapahtuneisiin eettisiin rikkeisiin ovat myös moninaisia. Johtajien mielestä suurin syy on paine tekoälyratkaisujen nopeaan implementointiin (34%). Erot suurimpien syiden välillä ovat kuitenkin vain muutamia prosenttiyksiköitä ja muut syyt liittyivät eettisten kysymysten huomioimiseen tekoälyn kehityksessä, tekoälyyn dedikoitujen resurssien (ihmiset, teknologia, rahoitus) puutteeseen, kehitystiimin homogeenisyyteen (rotu, sukupuoli ym.) ja eettisten periaatteiden puuttumiseen tai niistä poikkeamiseen (28%). (Capgemini Research Institute 2019, 11.)

Yritysten nykyiset toimet näyttävät kuitenkin olevan ristiriidassa kuluttajien odotusten kanssa. *Why addressing ethical questions in AI will benefit organizations* -tutkimuksen mukaan yli 70% kuluttajista haluaa tekoälyratkaisujen olevan läpinäkyvämpiä ja he haluavat tietää tulevansa kohdelluksi oikeudenmukaisesti, mutta vain 51% johtajista uskoo läpinäkyvyyden tärkeyteen, vain 37% yrityksistä on kiinnittänyt merkittävästi huomiota eettisiin kysymyksiin implementoidessaan tekoälyratkaisujaan ja vain 44% yrityksistä on valmistautunut vähentämään tekoälyyn liittyviä eettisiä riskejä. (Capgemini Research Institute 2019, 2, 12, 23.)

Myös Suomessa kansalaiset ovat eniten huolissaan inhimillisyydestä ja yksilöllisestä harkinnasta, läpinäkyvyydestä ja prosessin selkeydestä tulevaisuuden viranomaispäätöksissä,

joskaan täyttä yksimielisyyttä eri ikäryhmien välillä ei ole. Kansalaiset myös suhtautuisivat tutkimusten mukaan myönteisemmin tekoälyn käytön lisäämiseen, mikäli sen käyttöä ohjaisivat selkeät, velvoittavat säännöt, ja mikäli heillä olisi mahdollisuus saada tietoa tekoälyn toimintalogiikasta ja tiedoista, joita siinä on hyödynnetty. *”Niin ikään 76 % vastaajista koki, että heidän myönteinen suhtautuminen tekoälyn käyttöön kasvaa, mikäli myös tekoälyn suorittamista toimista ja päätöksistä vastaa aina ihminen. Vastauksissa on selkeästi havaittavissa läpinäkyvyyden ja avoimuuden merkitys.”* (Koivisto ym. 2019, 43–45.)

Kyseessä ei ole vain compliance-asia, vaan tekoälyn eettisyydellä saavutettaisiin tutkimusten mukaan myös merkittäviä liiketoiminnallisia hyötyjä, kuten asiakasuskollisuutta (Capgemini 2020). Kuluttajat ovat nimittäin uskollisempia sellaisia yrityksiä kohtaan, joiden tekoälyn käytön he mieltävät eettiseksi: he luottavat yritykseen enemmän (62%), ovat lojaalimpia, jakavat positiivisia kokemuksiaan (61%), ostavat enemmän ja antavat parempaa palautetta. Toisaalta he myös edellyttävät eettisyyttä: jos eettisiä rikkomuksia käy ilmi, he myös valittavat ja vaativat selityksiä (kuva 14). (Capgemini Research Institute 2019, 5, 7–8.)

Kuva 14 Kuluttajien reaktiot kohdatessaan epäeettisyyttä (Capgemini Research Institute 2019, 8)

3.5.4 Yritysten eettiset periaatteet

Kuluttajatutkijoiden Lehtiniemen ja Ruckensteinin (2019) mukaan *”eettiset viitekehykset ovat erityisen tärkeitä silloin, kun sääntely tai yhteiskunnalliset oikeudenmukaisuuden normit eivät auta jäsentämään toiminnan reunaehtoja".*

Anttinen & Lohilahti tutkivat vuonna 2019, millaisia tekoälyn eettisiä periaatteita Elinkeinoelämän keskusliiton alaiset yritykset olivat luoneet. Tutkimuksen mukaan (taulukko 7) kaikkien yritysten eettiset periaatteet sijoittuvat AI HLEG:n luotettavan tekoälyn vaatimuksiin, tosin erilaisin painotuksin (Anttinen & Lohilahti 2019, 29, 41–42).

Taulukko 7. Synteesi yritysten julkaistuista periaatteista suhteessa AI HLEG:n periaatteisiin (Anttinen & Lohilahti 2019, 41)

Organisaatio	Ihmisen toimijuus ja valvonta	Tekninen vakaus ja turvallisuus	Yksityisyyden suoja ja datan hallinta	Läpinäkyvyys	Monimuotoisuus, syrjimättömyys, oikeudenmukaisuus	Yhteiskunnallinen ja ekologinen hyvinvointi	Vastuuvelvollisuus	Jatkuva arviointi
A	✓	✓	✓	✓	✓	✓	✓	✓
B	✓		✓	✓	✓		✓	
C			✓	✓		✓		✓
D				✓	✓	✓	✓	✓

Tutkimuksessa selvitettiin, mihin tarpeeseen tekoälyn eettisiä periaatteita on luotu. Tuloksissa mainittiin, että tekoälyn eettisyys rinnastuu usein GDPR:ään ja korostuu henkilötietojen käsittelyssä. Yritysten mukaan suurta yleisöä tekoälyn eettisyys kiinnostaa yllättävän vähän, kun taas yrityksiä aihepiiri kiinnostaa, mutta tietoa ei ole riittävästi. Datankäsittelyä, ohjelmistoja, laitteistoa ja tutkimusta koettiin säänneltävän jo voimakkaasti, mutta yritysten itsesääntelyä koettiin silti tarvittavan, koska lainsäädäntö ei pysy teknologisen kehityksen perässä. Toisaalta yksi tutkimukseen osallistunut oli maininnut, että yritykset myös haluavat pitää kiinni mahdollisuudesta itsesääntelyyn. (Anttinen & Lohilahti 2019, 32–33.)

Yritykset ottivat kantaa myös periaatteiden julkaisuun. Toisaalta kaikilla yrityksillä ei ollut aikomusta tiedottaa tekoälynsä eettisyydestä tai he kokivat asian hankalaksi. Toiset taas ajattelivat periaatteiden julkisuuden edistävän tekoälyn etiikkaa yrityksessä. Periaatteet oli voitu sisällyttää myös käytännön työkaluihin. Yritykset itse kertoivat saavuttavansa tekoälyn eettisillä periaatteilla niin imagoetua kuin yhteistä visiota, avoimuutta ja tukea käytännön työhön. Periaatteiden koettiin lisäävän asiakkaan luottoa niin yritykseen kuin tekoälyynkin

ja niiden avulla viestitään yhteiskuntavastuun ja yksilön tarpeiden tärkeydestä ja eettisten kysymysten tiedostamisesta. Eettisyys nähtiin myös kilpailutekijänä. (Anttinen & Lohilahti 2019 35–37, 39, 43.)

Ollila kuitenkin kyseenalaistaa, mitä yritykset tekoälyn etiikalla tavoittelevat, kun yleensä tekoälyä käytetään yrityksen toiminnan tehostamiseen. Myös Anttisen ja Lohilahden tutkimuksen mukaan *”tekoäly toimii organisaatioissa eri prosessien apuna ja sen avulla organisaatiot pyrkivät tehokkuuteen ja tarjoamaan parhaan mahdollisen asiakaskokemuksen”,* vaikka periaatteiden suhteen *”haastatteluissa todettiin, että yleisesti ottaen yrityksen toiminnan lähtökohtana ja missiona pitäisi olla eettisyys ja eettisesti oikein toimiminen*”. Yhtäältä halutaan maksimoida omaa taloudellista hyötyä ja kilpailukykyä, toisaalta jaetaan huoli ihmiskunnasta ja biosfääristä. Ollilan mukaan ongelmallista on, että puuttuvan lainsäädännön vuoksi yritykset pääsevät määrittelemään ja luomaan mielikuvia hyvästä teknologiasta ja valvomaan itseään; toimimaan itse itsensä portinvartijoina. Anttisen ja Lohilahdenkin työssä todetaan, että *”varsinaista tekoälyn etiikan raportointivelvollisuutta haastateltujen mukaan ei ole ollut, koska aihe on uusi ja kansainvälisiä ohjeita aiheen tiimoilta vasta rakennetaan”*. (Anttinen & Lohilahti 2019, 36, 44–45; Ollila 2019, 23, 97, 108, 110.)

Mikäli tekoälyetiikan koodiston tarkoituksena on kasvattaa luottamusta tekoälyyn ja sen turvallisuuteen ylipäätään, voisi Ollilan mukaan olla parempi lähestyä tekoälyn etiikan itsesääntelyä ammattietiikan kautta samaan tapaan kuin toimii esimerkiksi Julkisen sanan neuvosto, diplomi-insinöörien Kunniasääntö tai lääkäreiden Hippokrateen vala, sillä asiakas ei välttämättä ole niin kiinnostunut kunkin yrityksen omista arvoista ja päämääristä kuin häntä koskevan toiminnan eettisyydestä ja toimijoiden moraalista ylipäätään. Myös digi- ja väestötietoviraston nykyinen pääjohtaja Viskari on muistuttanut, ettei kansalaista välttämättä edes kiinnosta, tuottaako tarvittavaa palvelua yksityinen yritys vai viranomainen vai tuottavatko ne palvelua yhdessä eri rooleissa, mikä edellyttäisi erilaisten toimijoiden tiiviimpää yhteistyötä. (Ollila 2019, 113–114 & Viskari 26.11.2019.)

Ollilan ehdotusta kuitenkin kritisoidaan muun muassa siksi, ettei yhtenäistä tekoälykehityksen ammattikuntaa katsota olevan. Tekoälyn kehityksessä ei ole ammattietiikan puitteita kuten identiteettiä, yhtenäistä ammatillista historiaa tai ammatillista kulttuuria. Globaalisti tekoälyjärjestelmiä saatetaan myös rakentaa tiimeissä, joissa kehittäjillä voi olla hyvin monenlaisia taustoja moraalivelvoitteineen, kannustinjärjestelmineen, kulttuureineen ja historioineen, jotka voivat olla ristiriidassa keskenään. Tekoälyn luonteen vuoksi ongelmia havaittaessa vastuuta ei välttämättä voida jäljittää yksittäiseen henkilöön vaan ennemminkin verkostoon, joka vaikutti tekoälyjärjestelmän kehitykseen sen eri vaiheissa. Tällöin on myös

vaikea tietää, miten yksittäisten kehittäjien valinnat vaikuttavat, jolloin myös "hyvän tekoälyn" tai "hyvän tekoälyn kehittäjän" vaatimuksiakaan ei voida luoda. Mittelstadtin mukaan tekoälyn eettistä keskustelua ei myöskään tulisi ohjata yksittäisiin kehittäjiin ja suunnitteluperiaatteiden ja arvojen noudattamattomuuteen, vaan tarkastella sen sijaan tekoäly-yritysten, organisaatioiden ja liiketoimintamallien eettisyyttä. (Mittelstadt 2019, 1–2, 5, 10.)

Tekoälyetiikan asiantuntija Haataja, Tekoälyaika-ohjelman eettisiin kysymyksiin keskittyneen ryhmän vetäjä, kritisoi uusien eettisten periaatelistausten luontia ylipäätään, koska *”niistä on kansainvälisesti periaatetasolla yhteisymmärrys”* mutta kannustaa yrityksiä organisaatiolaajuiseen, tekoälyeettiseen keskusteluun (Korhonen 2020). PwC:n AI Lead (Perrault ym. 2019, 149) on samaa mieltä periaatetason konsensuksesta ja peräänkuuluttaa niistä johdettuja toimia:

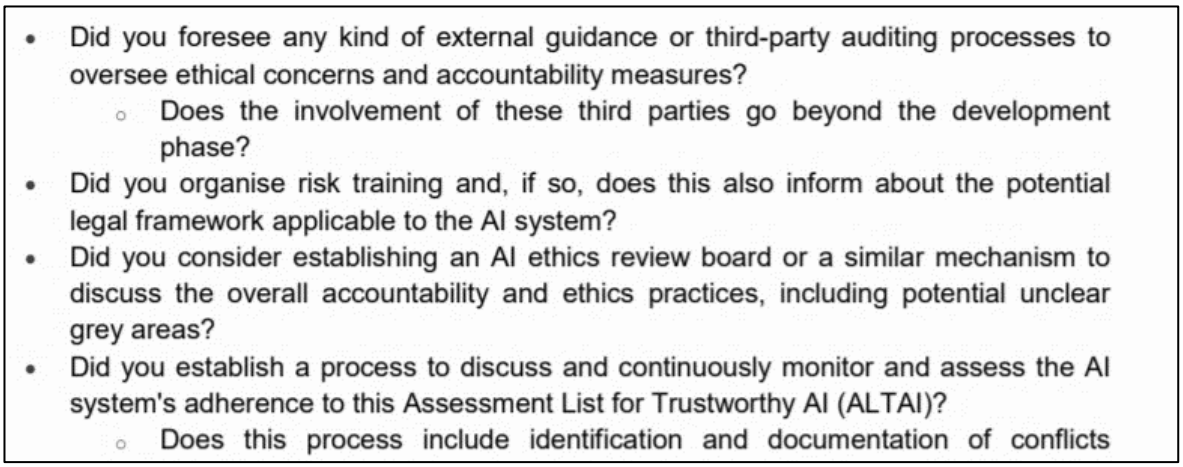
> “While there is a broad consensus emerging on the core set of principles associated with ethics and AI, the contextualization of these principles for specific industry sectors and functional areas is still in its infancy. We need to translate these principles into specific policies, procedures, and checklists to make it really useful and actionable for enterprise adoption.”

3.5.5 Luottamuksenarvoisuuden varmistaminen

”Arvot eli toiminnan päämäärät ovat sellaisinaan kovin abstrakteja, joten niistä on johdeltava periaatteita, joita noudattamalla arvoja voidaan soveltaa. Koska periaatteetkin ovat varsin korkealentoisia, niistä on johdettava sääntöjä, jotka konkretisoivat haluttua periaatetta. Kun sääntöjä edelleen havainnollistetaan, saamme menettelytapaohjeet.” (Ollila 2019, 106.)

Tutkimuksissa on osoitettu, että vaikka periaatetasolla eettisistä kysymyksistä voidaan olla samaa mieltä, eri tahot ovat erimielisiä siitä, miten periaatteita tulkitaan, miksi niitä pidetään tärkeinä, mitä asioita, toimialueita tai toimijoita ne koskevat ja miten ne tulisi panna täytäntöön. Epäselvää on myös, miten periaatteet tulisi priorisoida, miten eettisten periaatteiden välisiä ristiriitoja tulisi ratkoa, kenen eettistä tekoälyä tulisi valvoa ja miten eri tahot voivat noudattaa periaatteita. Lisäksi periaatteiden erilaiset tulkinnat tulevat ilmi vasta, kun periaatteita tai konsepteja testataan kontekstissaan, mikä on tärkeä ymmärtää. Jobin:n ym. mukaan nämä tulokset kielivät siitä, että periaatteiden luonnissa ja käytännön toteutuksen välillä on kuilu. (Jobin ym. 2019, 13–14; Mittelstadt 2019, 5–7.)

Tätä kuilua kaventamaan useat eri tahot ovat luoneet esimerkiksi erilaisia tarkistuslistauksia, varsinkin tekoälyn kehittäjille, joiden avulla varmistaa (mitata), että kehityksessä on huomioitu riittävällä tasolla arvoista, periaatteista ja vaatimuksista johdetut käytännön toimet. Eurooppalaisessa kontekstissa tunnetuin näistä lienee tällä hetkellä ALTAI (the Assessment List for Trustworthy Artificial Intelligence), AI HLEG:n luoma arviointiluettelo itsearviointia varten (kuva 15).


- Did you foresee any kind of external guidance or third-party auditing processes to oversee ethical concerns and accountability measures?
    - Does the involvement of these third parties go beyond the development phase?
- Did you organise risk training and, if so, does this also inform about the potential legal framework applicable to the AI system?
- Did you consider establishing an AI ethics review board or a similar mechanism to discuss the overall accountability and ethics practices, including potential unclear grey areas?
- Did you establish a process to discuss and continuously monitor and assess the AI system's adherence to this Assessment List for Trustworthy AI (ALTAI)?
    - Does this process include identification and documentation of conflicts

Kuva 15 Esimerkki ALTAI:n sisällöstä (AI HLEG 2020, 22)

ALTAI:n tarkoituksena on auttaa organisaatioita ymmärtämään, mitä luotettava tekoälyn käyttö tarkoittaa, miltä AI HLEG:n tekoälylle asettamat vaatimukset näyttävät käytännössä, tuoda esiin tekoälyyn liittyviä riskejä ja keinoja minimoida niitä mutta osoittaa samalla, miten maksimoida tekoälyn hyödyntämistä. Sen halutaan lisäävän eri sidosryhmien osallistumista ja tietoisuutta tekoälyn potentiaalisista vaikutuksista niin yhteiskuntaan, ympäristöön, kansalaisiin, kuluttajiin kuin työntekijöihinkin. AI HLEG korostaa, että arviointilista edellyttää aktiivista sitoutumista esiin nousevien kysymysten ratkaisemiseksi ja se kehottaa käyttämään arviointilistaa joustavasti: siitä voi ottaa tarkasteluun kehityksessä olevien järjestelmien kannalta merkityksellisiä osioita ja siihen voi lisätä omia kontekstisidonnaisia osia, jotta se palvelisi juuri sillä kyseessä olevalla sektorilla parhaiten. Arviointiluettelo ei myöskään poista tarvetta juuri esimerkiksi periaatteita tukevien käytänteiden ja työkulttuurin luomiseksi ja ylläpitämiseksi. (AI HLEG 2020, 3–4.)

Vastaavia tarkistuslistoja on kuitenkin luotu aiemminkin. Esimerkiksi Iso-Britannian hallinnolla on kattava dataetiikan viitekehys lukuisine itsearviointikysymyksineen ja numeerisine itsearviointiasteikkoineen, joiden avulla arvioidaan läpinäkyvyyden, oikeudenmukaisuuden ja vastuuvelvollisuuden toteutumista, kun tarkastellaan käyttötapausta ja sen hyötyjä julkiselle sektorille, datan laatua ja sen rajoituksia, lain ja ohjeistuksen noudattamista, monialaista moniosaajatiimiä ja projektia kokonaisuutena (Government Digital Service 2020).

Useat tahot kritisoivat myös tarkistuslistoja. Googlen mielestä tarkistuslistoilla ei pystytä yhteismitallisesti varmistamaan eettisyyttä kohteissa, joiden data, käytetty teknologia, käyttötapaus ja soveltamiskohde eivät koskaan ole samoja (Frey 2020). Loukides, Mason & Patil ovat kritisoineet edellä mainittua Iso-Britannian dataviitekehystä sen ylätasoisuudesta, kysymysten määrästä ja laadusta johtuen ja luoneet oman, dataprojekteihin osallistuville

suunnatun tarkistuslistan, johon muun muassa filosofi Ollila (23.1.2020) on puheissaan viitannut. Heidän listassaan on vain reilu tusina me-muotoisia, yksinkertaisia kyllä-ei-kysymyksiä, kuten olemmeko *me* listanneet, millä tavoin teknologiaa voisi käyttää väärin tai sitä vastaan hyökätä, tai että olemmeko *me* testanneet ja varmistaneet, että käyttämämme opetusdata on edustavaa ja oikeudenmukaista. (Loukides ym. 2018, 14–15.)

Mittelstadtin mukaan ongelmallista on kuitenkin juuri se, että periaatteita noudatetaan yleensä kirjaimellisesti ja erilaisina tarkistuslistoina sen sijaan, että ne olisivat osa kriittistä ja reflektiivistä käytäntöä ja toimisivat kulttuurin tavoin. Mittelstadtin mukaan monet ajattelevat, että ylätason periaatteet vain käännetään suunnitteluvaatimuksiksi, teknisiksi korjauksiksi ja ammattikoodistoksi. Todellisuudessa se on suuri metodologinen haaste, johon periaatteet eivät juuri tarjoa käytännön suosituksia. Esimerkiksi lääketiede on koko historiansa ajan kehittänyt tehokkaita tapoja, normeja ja hyvien käytäntöjen vaatimuksia ylätason sitoumusten ja periaatteiden toteuttamiseksi käytännössä. On ammatillisia yhdistyksiä, hallituksia, eettisiä arviointikomiteoita, akkreditointia, lisensointijärjestelmiä ja muita mekanismeja, jotka auttavat määrittämään etiikkaa vaikeiden esimerkkitapausten arvioinnilla ja rankaisemalla huolimattomasta tai virheellisestä toiminnasta. Tekoälyn kehityksestä sen sijaan puuttuvat lähes kokonaan oikeudellinen ja ammatillinen vastuuvelvollisuus ja ulkoiset pakotteet ja seuraamukset, joiksi Mittelstadtin mukaan eivät riitä maineriski ja epäeettisyyden julkitulo. (Mittelstadt 2019, 1–4, 6–8.)

Gasser & Schmitt kuitenkin puolustavat periaatteita ja pitävät niitä tärkeänä osana tekoälyn hallintamallia, niiden hallintovaikutusten epäselvyydestä huolimatta. Myös tekoälyn kehitykseen voidaan luoda, ja on luotukin, lääketieteestäkin tuttuja rakenteita, joilla vahvistetaan periaatteiden vaikuttavuutta. Edellytyksenä kuitenkin on, että periaatteita luodessa luodaan myös nämä rakenteet; normit pitää integroida ja toteuttaa. He myös muistuttavat, että ammatillisista normeista voi tulla oikeudellisesti merkittäviä monin eri tavoin: lainsäätäjät delegoivat normien säätämisen ammattiyhdistyksille, osapuolet sisällyttävät normien noudattamisen sopimukseen, tuomioistuin tulkitsee abstrakteja standardeja normien valossa tai sääntelyviranomainen voi hyväksyä asetusta täydentävät code of conduct:t osaksi regulaatiota ja valvontaa. (Gasser & Schmitt 2019, 18, 23–24.)

Yritystasolla tulee päättää, tarvitaanko yrityksessä lainsäädännön lisäksi muuta ohjeistusta ja mihin eettiset standardit asetetaan. Yksilötasolla tarkastellaan sekä tietoisuutta että toimintaa, esimerkiksi tiedetäänkö, mitkä säännöt ja normit ohjaavat toimintaa, puututaanko epäkohtiin ja valitaanko oikea vaihtoehto, vaikka se ei olisi vaihtoehdoista helpoin tai vaikka asian ratkaisemiseksi koettaisiin painetta. Yrityskulttuuri rakentuu juuri arkisista päätöksistä

ja siksi eettisten periaatteiden ja arvojen tulisi aina ohjata päätöksentekoa siten, että ensinnäkin kyetään tunnistamaan ne tilanteet, joissa on tehtävä eettinen valinta. Kun tilanne on tunnistettu, on pysähdyttävä miettimään, mitä tietoa asiasta on päätöksentueksi saatavilla, mihin tai keihin päätös tulee vaikuttamaan ja millaisia vaihtoehtoja on olemassa. Tämän jälkeen on punnittava, onko päätös tai toiminta lain, yrityksen arvojen ja vaatimusten mukainen sekä eettisesti hyväksyttävä: mitä haittaa tai hyötyä päätöksestä on, kehen päätös vaikuttaa ja miltä päätös vaikuttaisi ulkopuolisen silmin. Eettisten sääntöjen tulisi aina kirjallisesti saatavilla, ajantasaisia ja relevantteja, mutta virallisten sääntöjen lisäksi tulisi olla tietoisia epävirallisista ohjeista ja kirjoittamattomista säännöistä ja niiden sävystä. (Ratsula 2016, 20–21, 24–25, 29–30.)

Periaatteet tulee siis viedä käytännön tekemiseen: esimerkiksi tekoälyn eettiset periaatteet osaksi liiketoiminnan eettisiä periaatteita ja asiakaslähtöisyyttä, eettiset ongelmat osaksi riskienhallintaa, eettisyys osaksi kulttuuria ja datanhallinta osaksi tietosuojaa ja tietoturvaa (Vuori 2019, 33).

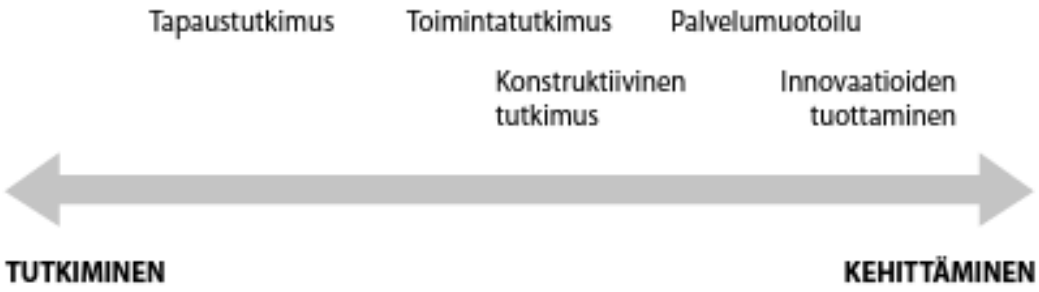
4 Menetelmät

“Laadullista tutkimusta voi luonnehtia prosessiksi. Kun laadullisessa tutkimuksessa aineistonkeruun väline on inhimillinen eli tutkija itse, voi aineistoon liittyvien näkökulmien ja tulkintojen katsova kehittyvän tutkijan tietoisuudessa vähitellen tutkimusprosessin edistyessä.” (Kiviniemi 2018, 72.)

Opinnäytetyön tutkimusstrategiana käytetään laadullisiin tutkimusmenetelmiin kuuluvaa tapaustutkimusta, sillä tehtävänä on selvittää, millaisia haasteita luottamuksenarvoisen tekoälyn käyttöön tai käsitteistöön liittyy, millä tavoin niitä on pyritty ratkomaan ja toisaalta millainen on työntilaajan konteksti luottamuksenarvoisen tekoälyn hyödyntämisen kannalta: millaiset tekijät nykyisellään tekoälyn hyödyntämistä ohjaavat ja onko periaatteellinen lähestymistapa oikea luottamuksenarvoisen tekoälyn varmistamiseksi.

Tapaustutkimukselle on tyypillistä hyödyntää useita tiedonkeruumenetelmiä, jotta tutkimuskohteesta saataisiin mahdollisimman paljon syvällistä ja yksityiskohtaista tietoa ja tutkimusongelmasta kattava kokonaiskäsitys (Ojasalo, Moilanen & Ritalahti 2015, 37, 52). Tapaustutkimus ei pyri tilastolliseen yleistettävyyteen, vaan korostaa ennemmin paikallista ja ajallista kontekstia ja kunnioittaa ilmiön moninaisuutta ja lähtee liikkeelle itse tutkittavasta tapauksesta, ei niinkään mistään yleisestä teoriasta tai mallista (Ojasalo ym. 2015, 52–54).

Opinnäytetyössä ei pyritä erottamaan tutkimista ja kehittämistä, koska ne nivoutuvat työssä tiukasti yhteen. Tapaustutkimuksesta ei ole erotettavissa erikseen konkreettisia tuotoksia, vaan tapaustutkimuksessa tuotetaan kehitysideoita tai kehitysehdotuksia tutkimuksen johtopäätöksenä, mitä kuvaa tapaustutkimuksen asettuminen tutkimuksen ja kehittämisen jatkumolla (kuva 16) enemmän tutkimuksen puolelle (Ojasalo ym. 2015, 36–37, 53).



Kuva 16 Lähestymistavat jatkumolla (Ojasalo ym. 2015, 36)

4.1 Aineistonkeruu- ja aineistonanalyysimenetelmät

Ojasalon ym. mukaan (2015, 40) menetelmien valinnan tulisi pohjautua siihen, millaista tietoa tutkimuskysymyksiin vastaaminen edellyttää ja miten kerättyä tietoa aiotaan tutkimuskysymyksiin vastaamiseksi käyttää. Koska tutkimusongelman ratkaisemiseksi on oleellista tutustua sekä luottamuksenarvoiseen tekoälyyn liittyviin näkökulmiin että yrityksen toimintaympäristöön, ensisijainen aineistonkeruumenetelmäni on haastattelut ja toissijaisesti tutustun yrityksen olemassa oleviin dokumentteihin.

Puolistrukturoidut haastattelut toteutetaan sekä yksilö-, että ryhmäteemahaastatteluina. Teemahaastattelut sopivat tapaustutkimukseen, koska tapaustutkimukset koskevat usein ihmisen toimintaa ja teemahaastatteluissa asiantuntijat saavat itse kuvata niin tekemistä kuin ilmiötäkin (Ojasalo ym. 2015, 55). Puolistrukturoitua haastattelua käytetään puolestaan siksi, että Ojasalon ym. (2015, 41) mukaan se sopii tilanteisiin, joissa tutkimuskohde ei ole entuudestaan täysin tuttu ja haastateltavia ei haluta siksikään ohjailla liikaa. Työntilaajan sisäisissä haastatteluissa käytetään ryhmähaastatteluja yksilöhaastattelujen sijaan, jotta saadaan selville, nouseeko keskusteluun uusia teemoja eri rooleissa tai eri toiminnossa työskentelevien keskinäisessä vuorovaikutuksessa. Ryhmäteemahaastatteluin saadaan myös useamman vastaajan ääni kuuluviin.

Ulkoisten haastattelujen tarkoituksena on varmentaa, kyseenalaistaa ja rikastaa tietoperustasta, sisäisistä haastatteluista ja olemassa olevista dokumenteista esiin nousevia teemoja. Työntilaajan sisäisten teemahaastatteluiden tarkoituksena on tutkia tarkempaan työntilaajan kontekstiin liittyen:

- tekoälyyn ja sen luottamuksenarvoisuuteen liittyviä asenteita ja käsitteitä työntilaajayrityksessä, jotta yhteisen ymmärryksen muodostaminen olisi mahdollista sekä opinnäytetyön aikana että sen jälkeen (käsitteistö, nykytila)
- tekijöitä, jotka nykyisellään tekoälyn käyttöä ohjaavat (nykytila)
- potentiaalisia haasteita, joihin tekoälyn käytön periaatteiden luomisella pyrittäisiin vastaamaan ja joiden avulla luoda jatkokehitysehdotuksia ja avainkysymyksiä periaatteiden validointiin (nykytila)
- mitä työntilaaja haluaa periaatteilla saavuttaa ja millaisena toimijana se haluaa tulevaisuudessa näyttäytyä (tavoitetila).

Haastattelut pelkistetään koodaamalla ja analyysimenetelmänä käytän aineistolähtöistä teemoittelua, vaikka teemahaastattelun runko osin toimiikin osana teemakortiston runkoa, koska haluan selvittää ennen kaikkea, mitä ulkopuoliset haastateltavat tuovat keskusteluun tai millaisia haasteita työntilaajan edustajat nostavat esiin.

Sekundäärisenä aineistonkeruumenetelmänä käytän dokumenttianalyysiä tutustumalla tarkemmin työntilaajan olemassa oleviin eettisiin liiketoimintaperiaatteisiin (Code of Conduct

ja Supplier Code of Conduct), arvoihin, strategiaan, brändikirjaan, työtä ohjaaviin konseptointeihin ja periaatteisiin ja aiemmin tehtyyn selvitykseen työntilaajan palveluiden oikeudellisesta luonteesta selvittääkseni työntilaajan nykytilaa ja tekoälyn käyttöä ohjaavia tekijöitä.

Yrityksen olemassa olevan ohjeistuksen suhteen analyysi tehdään teorialähtöisenä sisältöanalyysinä peilaten yrityksen olemassa olevan ohjeistuksen teemoja tekoälyn etiikan normatiiviseen ytimeen ja muihin työssä käsiteltyihin eettisiin periaatteisiin (liite 5a/b).

Dokumenttianalyysin tarkoituksena on selkiyttää ja järjestellä dokumentin sisältämää tietoa ja tehdä analyysejä ja päätelmiä tiivistetyn, kirjallisen aineiston pohjalta. Käytännössä sen vaiheet ovat aineistonkeruu ja saattaminen tarvittavaan muotoon eli esimerkiksi digitaaliseen muotoon, aineiston pelkistäminen, aineiston ilmiöiden tunnistaminen ja tulkitseminen sekä jatkuva prosessin tarkastelu. Sen etuna on löytää uusia näkökulmia kehitettävään asiaan, jos sitä käytetään muiden tiedonkeruumenetelmien rinnalla, mutta sen edellytyksenä on aina tarkastella kriittisesti sitä, mihin tarkoitukseen dokumentti on alun perin tehty ja tarkastella sitä kontekstissaan. Dokumenttianalyysin vahvuus on kuitenkin juuri sen ilmentämä asiayhteys. (Ojasalo ym. 2015, 43, 136–138.)

Teorialähtöinen sisältöanalyysi tarkoittaa, että ensin muodostetaan analyysirunko, jolla testataan esimerkiksi aiempaa käsitejärjestelmää (Ojasalo ym. 2015, 140). Tässä tapauksessa runko muodostetaan dokumenttianalyysiin valituista periaatteista, joista muodostetaan runko, jota vasten yrityksen aineistoa peilataan. Aineistosta voidaan poimia sekä rungon sisään että ulkopuolelle jääviä asioita ja muodostaa tarvittaessa niistä luokkia eli tässä tapauksessa uusia periaatteita (Ojasalo ym. 2015, 140–141).

Tarkoituksenani on tarkastella eri dokumenttien (mukaan lukien litteroitujen haastattelujen) välisiä yhteyksiä ja esiin nousevia puutteita, ristiriitoja, uusia teemoja tai käytännön ilmentymiä ja verrata näitä toisiinsa ja tietoperustaan, joka on irrottamaton osa tutkimuskysymyksiin vastaamista.

4.2 Luotettavuuden arviointi ja palautteen keruu

Opinnäytetyössä pyritään läpi prosessin aineistojen väliseen vuoropuheluun sekä aineistojen yhdistelemiseen. Tietoperustassa näkökulmia haetaan eri sidosryhmien näkökulmista, eri tahojen toiminnasta ja eri asiantuntijoiden lausunnoista. Toimintaympäristöä ja kontekstia selvennetään sekä työntilaajan dokumentaatiosta että litteroiduin haastatteluin läpi organisaation. Ulkopuolisten asiantuntijoiden haastatteluissa kiinnitetään huomiota eri ammattiryhmien edustukseen ja asiantuntemukseen ja aineiston saturaatiopisteeseen pyritään niin tietoperustan kuin tutkimusaineiston kanssa. Työssä kuvataan prosessia mahdollisimman tarkasti ja rehellisesti.

Opinnäytetyössä jätetään tilaa myös työn aikana esiin tuleville ilmiöille, teemoille tai aineistoille. Siinä missä kvantitatiivinen tutkimus yleensä lähtee liikkeelle teoreettisesta viitekehyksestä, käsitemäärittelystä, ennalta määritetystä hypoteesista ja tutkittavasta aineistosta, tarkentuvat nämä kvalitatiivisessa tutkimuksessa tutkimuksen edetessä. Näin itse aineistoista voi nousta merkityksellisiä teemoja käsiteltäviksi ja analyyttinen kehys voi muotoutua ennemminkin kvalitatiivisen tutkimuksen lopputulokseksi kuin toimia sen lähtökohtana. Tämä ei kuitenkaan tarkoita sitä, ettei tutkijan tarvitsisi tehdä valintoja aineistohavaintojen oleellisuudesta läpi tutkimusprosessin, mutta se mahdollistaa uudet oivallukset ja epätyypillisten ja poikkeavienkin näkökulmien esiin tuomisen, jotka kvalitatiivisessa tutkimuksessa ovat usein sekä kiinnostavia että merkityksellisiä. (Kurunmäki 2007, 86–89.)

Työssä on pyritään jatkuvaan systemaattiseen, analyyttiseen ja kriittiseen tarkasteluun, jonka avulla rakennetaan Ojasalon ym. (2015, 21–22) mukaan tutkimuksellisuutta kehittämistöihin ja osoitetaan tutkimuksentekijän kykyä muuntaa teoriapainotteista tietoa käytäntöön. Teorian ja käytännön vuoropuhelu on Ojasalon ym. (2015, 21) mukaan tärkeää, samoin kuin sen osoittaminen, mitä tapaustutkimuksessa on löydetty suhteessa tietoperustaan.

Kokonaisuuden hahmottamiseksi opinnäytetyössä on käytetty myös peittomatriisia, joka on esitelty opinnäytetyön kappaleessa 2.4 Rakenne. Peittomatriisin avulla hahmotetaan sitä, missä osioissa työtä tavoitteeseen liittyvää tietoperustaa ja tuloksia käsitellään ja millä esimerkiksi haastattelun kysymyksellä tai muulla tutkimusmenetelmällä on pyritty saamaan vastauksia kuhunkin tutkimuskysymykseen. Se kuvaa sekä työn rakennetta että yhteyttä tietoperustan ja oman tutkimuksen välillä ja varmistaa sisällön validiteetin. (Saaranen 2014, 4.)

Kehittämistyön ehdotuksia pyritään arvioimaan lyhyellä kyselylomakkeella. Työ luottamuksenarvoisen tekoälyn osalta kuitenkin jatkuu opinnäytetyöprosessin ulkopuolella ja tuon kollektiivisen työn yhteydessä osallistujia voidaan pyytää arvioimaan opinnäytetyötä ja ehdotuksia alustuksena esimerkiksi työpajatyöskentelyyn. Työtä ohjanneelta ohjausryhmältä pyydetään kirjallista palautetta opinnäytetyöprosessin lopussa.

5 Haastattelujen toteutus ja tulokset

Haastattelut ja dokumenttianalyysi toteutettiin pääosin menetelmäsuunnitelman mukaisesti. Dokumenttianalyysin osalta suunnitelmasta ei poikettu eikä dokumenttianalyysin tuloksia julkisessa versiossa käsitellä, joten sen tarkastelu on salattu opinnäytetyön liitteisiin. Myös työntilaajan haastatteluiden tulokset ovat salaisia, joten tässä luvussa käsitellään kaikkien haastattelujen toteutusta ja ulkopuolisten asiantuntijoiden haastattelujen tuloksia.

5.1 Haastattelujen toteutus

Opinnäytetyötä varten haastateltiin viittä ulkopuolista asiantuntijaa sekä 18:aa työntilaajayrityksen edustajaa. Kaikki haastateltavat valittiin harkinnanvaraisen eliittiotannan ja suosittelujen avulla. Ulkopuoliset haastatellut edustivat oman alansa huippua: tekoälyn etiikkaa, sovelluskehitystä, kyberturvallisuutta, työeläketoimialaa, vakuutusmatematiikkaa, datankäsittelyä ja informaatiomuotoilua. Haastateltavia tai heidän työnantajiaan ei nimetä työssä, koska kaksi ulkopuolisista haastateltavista koki, että he haluavat osallistua tutkimukseen omasta erityisalastaan, omista kiinnostuksenkohteistaan ja omasta roolistaan käsin, eivät niinkään organisaatioidensa edustajina.

Ulkopuolisten haastateltavien teemahaastattelut tehtiin pääosin ennen työntilaajayrityksen edustajien haastatteluja. Yhtä lukuun ottamatta kaikki ulkopuolisten haastattelut olivat yksilöhaastatteluja ja kestivät kukin 1-1,5 tuntia. Puolistrukturoitujen haastattelujen kaikille yhteiset kysymykset liittyivät tekoälyn haasteisiin ja ajankohtaisiin toimiin ratkoa niitä ja niillä varmennettiin, että opinnäytetyössä on huomioitu riittävästi eri näkökulmia ja hyödynnetty oikeita lähteitä. Osa kysymyksistä räätälöitiin vastaajan ammattialan mukaan. Haastatteluiden teemat ja kysymykset löytyvät liitteestä 2.

Työntilaajan haastateltavia oli kaikkiaan kolmetoista eri puolilta organisaatiota ja mukana oli eri rooleissa toimivia johtajia, päälliköitä, matemaatikoita, asiantuntijoita ja konsultteja. Yhtä haastattelua lukuun ottamatta haastattelut tehtiin kahden tai kolmen vastaajan ryhmähaastatteluina, joiden kesto oli kunkin 1-1,5 tuntia. Haastateltavat työskentelivät työkykyriskienhallinnassa, vakuutuspalveluissa, eläkepalveluissa, tietohallinnossa, vakuutustekniikassa, hankinnoissa sekä lakiasioiden tai datan parissa.

Työntilaajayrityksessä tehtyjen haastattelujen teemat koskivat aiheeseen liittyviä käsitteitä, yrityksen nykytilaa, tavoitetilaa ja tarvetta luottamuksenarvoisen tekoälyn käytön eettiselle ohjeistukselle. Haastatteluissa selvitettiin, millaisia luottamuksenarvoista toimintaa ohjaavia tekijöitä yrityksestä löytyy, ovatko ohjeet riittävällä tasolla, jääkö ohjeistuksen ulkopuolelle

kysymyksiä ja miten tuolloin epätietoisuus ilmenee. Lisäksi selvitettiin, miten ylipäätään käsitteet tekoäly ja luottamuksenarvoisuus ymmärretään ja millaisia kysymyksiä, riskejä ja haasteita niihin liitetään. Lisäksi pohdittiin, millaisena toimijana vastaajat yrityksen näkevät ja millaista kehitystä luottamuksenarvoisuuden saralla he toivoisivat näkevänsä. Myös viestinnällistä kulmaa mietittiin. Haastatteluissa käytiin läpi myös tekoälyn mahdollisuuksia, lainsäädäntöä ja erilaisia tekoälyn etikkaan liittyviä ajankohtaisia hankkeita. Haastatteluiden kysymykset ja teemat löytyvät liitteestä 3.

Ensin haastattelut taltioitiin ja litteroitiin, jonka jälkeen poimin ja merkitsin niistä tutkimuskysymysten kannalta oleellisimmat kohdat, jotka koodasin ja teemoittelin. Merkitsin litterointeihin myös, mikäli jokin kohta oli selvästi tullut vastaan toisessa yhteydessä käyden näin jatkuvaa keskustelua aineistojen kanssa ja välillä. Osa merkityistä kohdista noudatteli teemahaastattelun runkoa, jolloin koodasin ne suoraan ensin pääteema-alueittain: tekoälyn ja luottamuksenarvoisen tekoälyn käsitteet, suurimmat haasteet, uhat ja riskit, ratkaisuvaihtoehdot, mahdollisuudet, vaikutukset, ohjaavat tekijät ja eri toimijoiden roolit, työeläketoimialan ominaispiirteet, eettiset periaatteet, ajankohtaiset hankkeet ja tulevaisuuden näkymät. Haastatteluaineistosta nousi myös uusia teemoja, joiden sijoittelu ei ollut yhtä selvää, sillä ne eivät yhtä selkeästi liittyneet aiempiin teemoihin tai niitä ei oltu käsitelty tietoperustassa, kuten esimerkiksi ulkopuolisten haastateltavien esille tuomat design thinking, ihmisen kognitiiviset kyvyt ja vuorovaikutus koneiden kanssa, asiantuntemuksen ja tietoisuuden kasvu, datan- ja tekoälynlukutaito sekä datan ja tekoälyn hyödyntämisen demokratisoituminen ja hyvin vahvasti aikaulottuvuus eri konteksteissa. Jatkoin teemoittelua, kunnes sain luotua oleelliset, pelkistetyt teemat, jotka kuvaavat aineistoa, ja kunnes kaikki esiin tulleet tutkimuskysymysten kannalta oleelliset ilmiöt kuuluivat johonkin teema-alueeseen. Teemoittelin ja kokosin erikseen vielä haastatteluissa ilmenneiden päähaasteiden kuvaukset ja ratkaisuehdotukset taulukkoon, joka ulkopuolisten haastatteluiden osalta löytyy liitteestä 4.

Vaikka teemoittelussa yleensä etsitään usealle haastateltavalle yhteisiä asioita (Ojasalo ym. 2015, 110) analysoin haastatteluaineistoja myös etsimällä erityisesti löydetyistä teemoista poikkeavia ilmiöitä tai erilaisia näkökulmia samaan ilmiöön, koska ulkopuolisten haastattelujen tarkoituksena oli ennen kaikkea täydentää tietoperustaa ja näkökulmia tekoälyn luottamuksenarvoisuuteen ja koska työntilaajan haastattelujen tarkoituksena oli löytää muun muassa ristiriitoja ja merkityksellisiä eettisiä kysymyksiä. Merkityksiä voidaan luoda Ojasalon ym. (2015, 143–144) mukaan muun muassa tunnistamalla toistuvia rakenteita, rakenteesta poikkeavia muuttujia ja niiden välisiä linkkejä, tekemällä vertailuja ja kontrasteja ja näkemällä uskottavia selityksiä.

Kaikki haastattelut käsiteltiin samalla tavalla mutta pidin ulkopuolisten haastateltavien ja työntilaajan edustajien haastattelut erillään toisistaan, koska ulkoisten haastattelujen päätavoitteena oli varmistaa, että tietoperustassa on huomioitu riittävästi eri näkökulmia, ja työntilaajan haastattelujen tarkoituksena oli selvittää nimenomaan työntilaajan käsitteistöä ja kontekstia.

5.2 Ulkoisten haastattelujen tulokset

5.2.1 Tekoälyn käsite

Ensimmäisissä ulkopuolisten haastatteluissa haastateltavat saivat itse vastata, miten he tekoälyn ja luottamuksenarvoisuuden ymmärtävät. Muissa käsitemäärittelyn alustuksena toimi opinnäytetyöhön kirjatut määrittelyt, joista keskusteltiin. Tekoälyä pidettiin jo itsessään filosofisena terminä, mutta teknisesti yksi vastaajista katsoi sillä tarkoitettavan *”ohjelmistoja, prosesseja niiden tukena, joilla pystyy tekemään kehittynyttä, nykyaikaista data-analytiikkaa, ja sitä kautta rakentamaan sellaisia palveluita, ratkaisuja, vaikka mobiikäyttöappseja, jotka pystyy asioita ymmärtämään ja ne pystyy ennustamaan, että käyttäjä kokee, että niissä on älykkyyttä"* ja jotka verrattuna tavallisin sovelluksiin hyödyntävät dataa paremmin ja pystyvät toimimaan itsenäisesti. Vastaaja viittasi määrittelyssään myös Alan Turingin testiin niin, että jos ihminen ei kognitiivisten kykyjensä rajoissa kykene erottamaan konetta ihmisestä, voidaan puhua tekoälystä. Toinen vastaaja ei kokenut, että mekaanisiin, etukäteen määriteltyihin päättelysääntöihin pohjautuvat ratkaisut olisivat tekoälyä, vaikka ne olisivat monimutkaisiakin ja vaikka käyttäjästä palvelut voisivat *”vaikuttaa älykkäältä”.* Myös muut edellyttivät tekoälyltä *”jonkin sortin älykkyyttä” ja ”monimutkaisuuden astetta”.*

Tekoäly ymmärrettiin kattoterminä, laajana konseptina, jonka alle myös laajalti käytetty koneoppiminen kuuluu, vaikka arkisessa keskustelussa saatetaan koneoppiminen nähdä synonyyminä tekoälylle. Koneoppimiseen viitataan, kun yrityksille myydään tekoälyä tai kun puhutaan yrityksissä käytettävästä tekoälystä tämän ajanhetken termein: *”Ja nyt tänä päivänä sitten kaikki, jotka tekee jotain koneoppimisen kanssa, niin tietää, että tekoäly on se termi, millä heidän duunista puhutaan.”* Muun tekoälyn nähtiin jääneen *”paitsioon”*.

Tekoälyn määritelmän katsottiin myös riippuvan termin käyttäjästä. Esimerkiksi ohjelmistorobotiikasta puhutaan toisinaan tekoälynä. Arveltiin, että mikäli puhuja tuntee tekoälyyn liittyvää terminologiaa syvemmin, liittää hän todennäköisesti tekoälyyn esimerkiksi adaptiivisten ja luovien piirteiden vaatimuksen, mikä rajaa yksinkertaiset toteutukset tekoälyn määritelmän ulkopuolelle. Yksi vastaajista viittasi tekoälyn *”viralliseen määritelmään”* spesifioimatta sen sisältöä tai määrittelijätahoa tarkemmin, mutta totesi myös, ettei tekoälyn tarkka määritelmä ole tekoälyn etiikan kannalta edes relevanttia.

Tekoäly nähtiin pääosin työkaluna, yhtenä teknologiana muiden joukossa, mutta toisaalta muistutettiin myös, että tekoäly voi olla sovelluksissa osana kokonaisuutta, kuten esimerkiksi erilaisissa alustoissa. Vastaaja toi esiin myös ihmisen ajattelukyvyn rajallisuuden siinä, että vaikka tekoälyä sisältävät sovellukset ovat vain ohjelmistoja, ne saattavat toimia myös manipulatiivisesti, ihmismielen tiedostamattoman puolen kanssa ja *”se ongelma syntyy siitä, että me ihmisinä ajatellaan, että kaikki mitä me tehdään, niin me ymmärretään se, ne on rationaalisia päätöksiä, me tehdään tiedostettuja päätöksiä”*, mitä vastaaja piti ongelmallisena oletuksena.

5.2.2 Luottamuksenarvoisuuden käsite

Luottamuksenarvoinen tekoäly ei terminä ollut tuttu, vaikka vastaajat olivatkin pääosin tutustuneet esimerkiksi AI HLEG:n luotettavan tekoälyn vaatimuksiin (*7 key requirements for Trustworthy AI*). Yksi vastaajista nosti kuitenkin suoraan keskusteluun kyseiset periaatteet. Muissa pelkkä termi herätti ajatuksen siitä, että luottamuksenarvoisuus liittyy esimerkiksi datan asianmukaiseen ja turvallisuussäädeltyyn käsittelyyn, EU:n tietosuoja-asetukseen ja yksityisyydensuojaan, läpinäkyvyyteen, luottamukseen algoritmeja kohtaan, mustien laatikoiden tekemiin päätöksiin ja oikeudenmukaisuuden vaatimukseen, mutta myös siihen, että tekoälyyn liittyy aina epävarmuuksia ja epäluuloja sekä kansalaisten että asiantuntijoiden puolelta.

Luottamuksenarvoisuuden käsitteen yhteydessä esiin tuotiin myös Asimovin robotiikan perussäännöt: robotti ei saa vahingoittaa ihmistä, sen tulee noudattaa ihmisen määräyksiä ja suojella omaa olemassaoloaan mutta sen toiminta ei koskaan saa olla ristiriidassa edellisten sääntöjen kanssa. Vastaaja katsoi, että AI HLEG:n periaatteissa säännöt näkyvät ihmisen toimijuuden ja ihmisen suorittaman valvonnan vaatimuksena, mitä vastaaja piti yleisinhimillisenä näkökulmana luottamuksenarvoisuuteen. Vastaaja näki myös AI HLEG:n periaatteissa useita tasoja: osa periaatteista on tärkeitä yksilölle mutta yhteisölle saattavat olla tärkeitä esimerkiksi vaikutusten arviointi sekä demokratian, oikeudenmukaisuuden ja syrjimättömyyden vaaliminen.

Yksi vastaajista korosti erikseen sitä, että luottamuksenarvoisuus on erilaista eri toimijoiden näkökulmasta: loppukäyttäjän mielessä luottamusta voi herättää odotusten kohtaaminen ja positiivinen kokemus tilanteesta, kun taas jollekin toiselle taholle luottamuksen rikkoutumisella voi olla suurempia ja laajempia vaikutuksia kuin vain kokemuksellinen pettymys.

5.2.3 Läpinäkyvyys ja selitettävyys

Yksi eniten mainituista haasteista oli läpinäkyvyys. Vastauksissa korostettiin, että läpinäkyvyys ei tarkoita pelkästään tuloksen läpinäkyvyyttä, vaan läpinäkyvä prosessi ohjaa läpinäkyvään tekemiseen, mikä mahdollistaa sen, että tuloksista voidaan myös avoimesti viestiä. Yksi vastaajista korosti kuitenkin, että yritysten tulee olla läpinäkyviä *”sillä tasolla, kun se on tarpeellista ja soveltuvaa sen vastaanottajapään näkökulmasta”* ja tarkensi, että julkisen sektorin läpinäkyvyyden vaade on erilainen kuin yksityisen ja lisäksi auditoijien ja viranomaisten suuntaan saatetaan tarvita eritasoista läpinäkyvyyttä kuin asiakkaiden ja kuluttajien. Vastaaja myös korosti, että läpinäkyvyyden tason suhteen yritysten tulee tehdä omia päätöksiään.

Vinoumat tulevat esiin vain, jos tekoälyjärjestelmä ja prosessit sen taustalla ovat läpinäkyviä, mikä puolestaan on haastavaa, jos dataa, algoritmeja, järjestelmiä ja liiketoimintamalleja ei voida avata kilpailuedun ja arvontuoton vaarantumatta. Yhden vastaajan mukaan tällöin riskinä kuitenkin on, ettei myöskään esimerkiksi kuluttaja kykene arvioimaan, mistä hänen käyttämänsä tekoälyjärjestelmän riskit ja ristiriidat muodostuvat.

Yksi vastaajista uskoi läpinäkyvyyden ja selitettävyyden vaateen lisääntyvän ja *”tulevan osaksi kaikkia palveluita”,* jolloin kuluttajalla olisi mahdollisuus tietää esimerkiksi, kuka tekoälyä sisältävästä palvelusta vastaa, mitä dataa siinä hyödynnetään, miten se toimii ja miten palveluun voi esittää korjauksia. Näin kuluttaja voisi myös oppia arvioimaan palvelun luotettavuutta sen perusteella, ovatko nämä kaikki yleisesti jaetut tiedot hänen saatavillaan. Tämä olisi myös yksi indikaatio tekoälylukutaidon kehittymisestä. Tulevaisuusskenaariona oli myös osallistava personointi, jolloin kuluttaja voisi itse vaikuttaa siihen, mitä hänelle suositellaan, antaa palautetta ja kehittää personointia.

5.2.4 Datan käsittely, GDPR ja yksityisyydensuoja

Yksi vastaaja toi esiin, että mitä enemmän dataa on saatavilla ja mitä suurempi rooli datan käytöllä on ihmisten elämässä, sitä merkittävämmäksi muodostuu kysymys datan laadusta, lähteestä, relevanttiudesta sekä myös datanlukutaidosta, jotta ymmärtää, mitä data kertoo ja miten siihen tulee suhtautua. Hän korosti, ettei datan lähdekään välttämättä kerro datan laadukkuudesta, sillä on mahdollista, ettei siihen päästä käsiksi, datan keruumetodi on ollut väärä, tavat kerätä dataa ovat vuosien mittaan muuttuneet, data on aiemmin kirjattu yksityiskohtaisemmin, kaikkea dataa ei enää olekaan saatavilla, data on kerätty tiettyä käyttötarkoitusta varten tai datan omistaja haluaa tietoisesti datan näyttävän tietyn näkökulman mukaiselta. Vastaaja toivoi, että myös *”lukujenkin semmoista puhdasta objektiivisuutta aletaan ihan aiheellisesti vähän haastaa ja kyseenalaistaa. Miten tämä luku on itseasiassa*

muodostettu, mitä se oikeasti kertoo ja mitä tämmöisiä virhelähdemahdollisuuksia siinä edes on”.

Vastaaja toi esiin merkittävän nykyajan dataan liittyvän paradoksin siitä, että vaikka dataa näennäisesti olisi enemmän kuin koskaan, se ei välttämättä ole vertailtavampaa, relevantimpaa tai täydellisempää kuin aiemmin. Vastaajan mukaan datassa voi olla huolestuttavia aukkoja edellä mainituista syistä, mutta myös siksi, että datan kerääminen ihmisen tekemänä on työläämpää ja siten kustannustehottomampaa kuin sellaisen tiedon, joka syntyy automaattisesti esimerkiksi laitteen tuottamana, joten on mahdollista, että esimerkiksi joitain tilastoja ei enää esimerkiksi kustannussäästöjen vuoksi enää kerätä.

Toinen haastatelluista puhui laajemmin dataan, sen omistajuuteen ja valtaan liittyvistä riskeistä: "*Alustat ja kokonaisuudet kerää tietysti todella paljon dataa ja sen datan avulla pystytään kuitenkin näkemään kokonaisuuksia automaattisesti, mihin ihmissilmä ei pysty, sellaisia riippuvuuksia tavallaan rakentamaan, niin tietysti siinä samassa yhteydessä muodostuu sellaisia osaamis- ja tietokeskittymiä”.* Vastaaja oli huolissaan siitä, että tällöin dataa voi myös käyttää väärin, kuten Cambridgen Analytican tapauksessa. Häneen mukaansa oleellista on siis huolehtia ennen kaikkea yksityisyydensuojasta, tietoturvasta, riskienhallinnasta ja lisätä käyttäjien ymmärrystä siitä, miten järjestelmät toimivat. Tieto- ja kyberturvallisuuden osalta tulisi ymmärtää, että yritysten täytyy tuntea oma datapotentiaalinsa – ei vain siitä näkökulmasta, että mitä kaikkea sillä voisi tehdä tai mitä dataa yritykseltä toisaalta puuttuu ja kerrytetäänkö dataa riittävästi liiketoimintaprosesseissa aina tilaisuuden tullen –vaan myös siitä näkökulmasta, mikä on se data, jota yrityksen tulee suojella kaikin keinoin.

5.2.5 Vastuuvelvollisuus

Vastuuvelvollisuuden ja hankintojen osalta esille tuli monitahoinen epäselvyys ja sen mukanaan tuoma suuri riski siitä, ettei vastuuta ole selvästi jaettu: toisaalta vastuu tekoälyalgoritmin tuottamasta haitasta on järjestelmänsuunnittelijalla tai -kehittäjällä, toisaalta järjestelmän myyjälläkin on velvollisuutensa eikä ostajakaan voi sanoa, *”että emme me tienneet, me vain ostimme palvelua”*. Kerrointa lisää myös se, että *”ohjelmistotoimittajat, jotka rakentavat ohjelmistoja ja ne asiakkaat, jotka niitä käyttävät, niin heillä ei välttämättä ole kummallakaan täydellistä näkemystä siitä, tai ymmärrystä, mitkä ne riskitasot on”.* Eri tahot eivät muutoinkaan välttämättä jaa samoja periaatteita tai periaatteet eivät ohjaa sitä tekemistä tai osaaminen ei riitä vastapuolen erikoisalasta. Vastaaja ei osannut vastata, kenellä se *”viimekätinen vastuu”* olisi, mutta hän piti hyvänä sitä, että on olemassa esimerkiksi tietoperustassakin mainittu alusta, jota on käytetty muun muassa Helsingin ja Amsterdamin kaupunkien tekoälyrekisterin pohjana, koska *”se on selkeä, hyvä askel, oikeaan suuntaan, että tavallaan keskustellaan ja on malli, jolla voidaan katsoa näiden nykyisten, vanhojen ja*

myöskin tulevien ohjelmistoversioiden mallia ja niitä auditoida”. Pelkäksi kontrolliksi sekään ei välttämättä riitä mutta vastaaja piti mahdollisena skenaariona, että sellaisesta voisi kehittyä *”oma yritystoimiala, jossa oikeasti näitä auditoidaan jatkuvasti”,* koska kun on

> ”tutkinut kymmeniä, kenties satoja, ohjelmistoja ja niiden toimintaa, niin kyllähän tällaisilla auditoivilla organisaatioilla alkaa olla jo vähitellen näkemys siitä, mitkä on hyviä malleja ja mitkä on keskeisesti vääristyneitä ja voisiko sanoa, jopa huonoja malleja, jopa niin kuin epäeettisiä malleja”.

Yksi auditoinnin malli voisi myös ilmailualan esimerkki, että auditoinnilla katetaan se, että *”tiedetään, mitä minkälaisia odottamattomia tahattomia vahinkoja on tähän mennessä tapahtunut, kuinka ainakin niistä vanhoista voidaan oppia”.* Kun käytössä olisi kaikki juurisyineen selvitetyt tapahtumat kuvattuina, dokumentoituina ja tilastoituina, niin olisi pohjaa, mistä oppia ja mitä säännöllisesti varmistaa auditoinneilla.

Koettiin myös, että esimerkiksi ohjelmistopuolen edustajana olisi eettistä opetella asiakaskunnan erikoisalaa, tavata alan huippuja ja pysyä tietoisena alan uusimmasta kehityksestä voidakseen tarjota entistä parempaa ohjelmistoa parantamalla ohjelmiston toimintaa, ei vain raportoitujen bugien osalta, vaan myös sen toimintalogiikan osalta, jotta ohjelmisto pysyisi kehityksessä mukana ja tekisi entistä laadukkaampia valintoja sekä asiakkaan että loppukäyttäjän näkökulmasta. Tämä edellyttäisi myös rohkeutta myöntää omia virheitään ja *”on aina rehellistä ja välttämätöntä todeta se, että kaikessa aluksi tekniikka vaan ei ole täydellistä”.* Vastauksissa nähtiin laajemmaltikin, että *”mitä lähempänä se* [tekoäly] *on sitä bisnestä, niin sitä paremmat edellytykset sillä on”.*

Mainittakoon, että sama tarve on tunnistettu Auvisen ym. (2019, 7) mukaan myös pöydän toisella puolella julkishallinnossa, jossa pohditaan muun muassa, onko hallinnossa riittävää kehitys-, hankinta- ja ylläpito-osaamista, voisiko resursseja yhdistää poikkitieteellisen osaamisen ja osaajien varmistamiseksi, tulisiko loppukäyttäjien osallistua tarvemäärittelyihin ja millainen on järjestelmiin liittyvä roolijako työkaluineen ja vastuineen.

5.2.6 Yhteisen viitekehyksen ja lainsäädännön puute

Regulaatio tuli haastatteluissa esille. Yksi vastaajista piti harmillisena, että Euroopan komissiolta odotetaan *”lopullista totuutta”,* mitä yrityksissä jäädään odottamaan. Lainsäädäntöä vastaajan mukaan varmasti onkin tulossa, mutta sitä ei tulisi jäädä odottamaan, vaikka vastaaja itsekin toivoi ”*regulatiivisen kentän selkiytystä*”.

Kysyntä nousi haastatteluista myös esiin. Esimerkiksi yksi vastaajista peräänkuulutti tutkimuksesta ja mallienkehittäjiltä teknisiä ratkaisuja ratkomaan mustan laatikon ongelmaa, mutta totesi samalla, että siihen *”pitää olla pull:ia”.* Uuden EU-direktiivin pitäisi esimerkiksi vaatia, *"että mikäli keinoälyä käytetään tällaisessa ja tällaisessa päätöksenteossa, niin sen*

pitäisi täyttää tietyt minimi-, jotkut läpinäkyvyyskriteerit siitä, että pystyttäisiin backtrace:aamaan sitä, miten se päättely on kulkenut."

Vastaaja toi esiin myös kiinnostavan kulman EU:sta toteamalla, että *"vaikka sitä* [EU:ta] *jatkuvasti haukutaan tehottomaksi ja byrokraattiseksi, niin osittain ehkä juuri se byrokraattisuus on tietynlainen vahvuuskin Euroopan unionille. Elikkä Euroopan unionilla on valtava sääntelyvoima."* Hän perustelee ajatustaan tuomalla esiin GDPR:n, saavutettavuusdirektiivin ja päästörajoitukset, joiden vastaaja koki radikaalisti muuttaneen toimintakenttää, vaikka direktiivit eivät välttämättä ota kantaa käytännön toteutuksiin. Vastaaja myös arveli, ettei lainsäätäjillä tai päättäjillä ole riittävää tekoälyosaamista reguloimaankaan tarkempia käytäntöjä ja lisäsi, että yksi sääntelyvaihtoehto voisi olla niin sanottu kollegiaalinen malli, jossa *"yritykset, tutkijat ja muut loisivat sen eettisen koodiston itse ja julkinen valta olisi ainoastaan auttamassa siinä prosessissa, kätilöimässä sitä ja ehkä vaatimassa sen standardin noudattamista tietyissä asioissa niin kuin julkisissa hankinnoissa ja päätöksenteossa ja muualla"*. Esimerkkinä vastaaja käyttää Julkisen sanan neuvostoa journalistisine standardeineen sekä W3C-konsortion luomaa Web Content Accessibility Guidelines -ohjeistusta, joka syntyi alan sisäisenä työnä ja jonka EU nosti juridiselle tasolle ja pakottavaksi saavutettavuusdirektiiviksi. Vastaaja myös tarkensi, että on julkisen vallan tehtävä varmistaa, että alan luoma eettinen koodisto *"saadaan leikkaamaan läpi myös niiden toimijoiden keskuudessa, jotka eivät ole siihen standardityöhön tai sen koodiston rakentamiseen alkuvaiheessa vielä sitoutuneet".*

5.2.7 Eettiset periaatteet

Kysyttäessä, miten esimerkiksi AI HLEG:n periaatteisiin ja luotettavan tekoälyn arviointilistaukseen tulisi suhtautua, yksi vastaajista totesi, että periaatteita ja niitä luovia tahoja on paljon, ja yritysten tulisi ymmärtää, että näiden tahojen tarkoitus on tuoda jatkuvasti esiin vähän uutta ja omaa kulmaa, jolloin yrityksille tärkeintä olisi kuitenkin tarkastella sitä omaa toimintaa ja kontekstia ja löytää sieltä ne omat käytännölliset tavat (hallintamallit, työkalut, dokumentointitavat, uudet rekrytoinnit) edistää eettistä tekoälyn käyttöä. Toisen vastaajan mielestä AI HLEG:n periaatteet olivat *"kohtuullisen hyvä kokoelma",* joka nostaa oleellisia asioita esiin ottamatta kuitenkaan kantaa vielä teknologiaan. Kolmas vastaaja totesi AI HLEG:n periaatteiden sekä Euroopan komission valkoisen kirjan (kts. Euroopan komissio 2020a) olevan *"vasta tietenkin niin kuin teemoituksia siihen poliittiseen keskusteluun"* ja piti tärkeänä, että meillä (viitaten oletettavasti sekä EU-tasoon että kansalliseen tasoon) on yhtenäisiä näkökulmia eettisiin periaatteisiin ja luottamuksenarvoiseen tekoälyn käyttöön, millä saavutettaisiin synergiaetuja.

AI HLEG:n luomaa itsearviointityökalua sivuttiin myös. Yksi vastaajista kritisoi sitä, ettei minkään tahon periaatteet tai niiden varmistamiseksi luodut arviointilistat riittävällä tasolla vastaa siihen, miten tunnistettuihin haasteisiin vastataan. Ehkä merkittävin yksittäinen näkökulma oli yhden vastaajan suhtautuminen yritysten omiin eettisiin periaatteisiin, joista vastaaja totesi, että vaikka Tekoälyaika-hankkeen peräänkuuluttama yritysten omien periaatteiden luonti oli *”tosi tärkeää”,* kuten myös yritysten ylimmän johdon sitoutuminen ja organisaatioiden panostus yhteisten eettisten periaatteiden luontiin, on aika ajanut ohi tarpeesta: *”Joskus vuosi sitten mä puhuin silleen, että viime vuosi oli tekoälyeettisten periaatteiden tekemisen vuosi, ja nyt pitäs toimia, nyt ei erotuta enää niillä vaan, että nyt erottuu jyvät akanoista siinä, että ketkä toimii ja ketkä jättää sen siihen. Ja se on semmosta white washingia sitten vaan, jos sä et tee mitään muuta kuin julkaiset ne periaatteet.”* Vastaaja myös viittasi Gasserin & Schmittin (2019) tutkimukseen, jonka mukaan periaatteet ja muut ammatilliset normit jäävät usein hyvin vähävaikutteisiksi, jos niiden vaikuttavuutta ei vahvisteta lainsäädännöllä, implementointi- ja vastuuvelvollisuusmekanismein ja julkisin painostuksin.

Vastaaja kuitenkin lisäsi, että periaatteiden miettiminen on hyvä ensimmäinen askel, mutta *”ellei se ajatus ole pidemmällä, että mitä niillä halutaan sitten aikaansaada konkreettisesti ja miten ne viedään eteenpäin, niin helpommalla pääsee, kun ottaa jonkun näistä valmiista olemassaolevista”* ja jatkaa, että voisi hyvin suositella AI HLEG:n periaatteita arvokeskustelun ja käytäntöjen rakentamisen pohjaksi varsinkin, kun on tutkimuksinkin todistettu (kts. 3.5.2 Tekoälyetiikan normatiivinen ydin), etteivät eri tahojen periaatteiden ydinteemat juuri poikkea toisistaan. Vastaaja katsoi, että omia periaatteita voisi perustella vain hyvin uniikilla toimialalla tai yrityksen omilla arvoilla ja ohjeistuksilla, joita halutaan tarkastella ja tarkentaa tekoälyn valossa. Kuitenkin sen lisäksi, että yritys joutuu peilaamaan uusien teknologioiden mukanaan tuomia vaatimuksia suhteessa olemassa oleviin periaatteisiin ja ohjeistuksiin, joutuvat yritykset myös miettimään käytännön varmistamista ja vaatimusten painotuksia, sillä vaatimuksia on kuitenkin verrattain runsaasti.

Periaatteita tulisi vastaajien mielestä myös soveltaa eri tavalla eri käyttökohteisiin. Negatiivisena esimerkkinä tekoälyn käytöstä yksi vastaajista toi esiin IB-lukioissa käytetyt algoritmit, joiden avulla pyrittiin korvaamaan Koronan takia peruuntuneita päättökokeita. Toinen spesifi esimerkki oli vakuutuslääkärin tekemän työkyvyttömyysarvion siirtäminen tekoälylle ja kolmas koski Yhdysvaltain poliisin käyttämää kasvojentunnistusohjelmaa. Yleisemmin mainittiin rahoitusala, lääketiede, vakuutusala, terveydenhoitoala, työeläketoimiala sekä oikeuslaitokset ja automaattinen päätöksenteko ylipäätään, jotka tulisi luokitella korkean riskin aloiksi, joita tulisi käsitellä erikseen.

5.2.8 Työeläketoimiala

Työntilaajan ulkopuolelta haastatellut työeläketoimialan asiantuntijat näkivät työeläkealan stabiilina, luotettavana ja pitkäjänteisenä. Työeläketoimialan näkökulmasta datankäsittelyä pidettiin hyvin tärkeänä näkökulmana siksi, että alalla käsitellään niin paljon henkilötietoa. Kokonaisuutena työeläketoimialaa pidettiin hyvin reguloituna ja stabiilina, rooleiltaan täysin vakiintuneena ja monia luottamuksenarvoisuuden näkökohtia (turvallisuus, yksityisyydensuoja) pidettiin jo itsestäänselvyyksinä. Datalähteistä koettiin jo nyt kerrottavan, mistä tiedot ovat peräisin ja miten niitä käytetään. Tekoälyyn suhtauduttiin lisätyökaluna, joka tuo alalle lähinnä haasteita läpinäkyvyyden ja selitettävyyden muodossa: *”ei ymmärretä tai osata kertoa, mitä miten joku malli tai sovellutus toimii, josta sitten jää puuttumaan se läpinäkyvyys, joka syö sen luottamuksen”.* Mallien valinnan tiedettiin ja tulosten tulkinnan ymmärrettiin vaikuttavan merkittävästi lopputulokseen, jolloin asiantuntijalle jää paljon harkintavaltaa, ja viime kädessä datan käytön valvonta nähtiin esimiesten tehtäväksi. Vaihtokauppana pidettiin nopeaa työskentelyä ja datan luottamuksellista käsittelyä, mutta lähinnä tiedostaen, että kun data on vapaammin käytettävää eikä sisällä luottamuksenarvoista henkilötietoa, myös mallin rakentaminen on nopeampaa. Huolissaan oltiin myös perässä laahaavasta lainsäädännöstä: *”ei ole kaikilta osin lainsäädäntö valmista, sillä tavalla valmista, että meillä olisi sitä säädöspohjaa siihen, mitä teknologia jo mahdollistaa, jolloin sitten tullaan tähän, että ei ole yksilön kannalta turvallisuussäädeltyä”.*

Tämän hetken työeläketoimialan tekoälyratkaisujen koettiin olevan yksittäisiä, jolloin ennen kutakin tapausta käydään tarkasti läpi, miten tietoa käsitellään, millaisia sopimuksia tehdään ja mitä myös mahdolliselta kumppanilta edellytetään. Kokeiluissa on lähinnä tavoitteena oppia ja toisaalta jakaa tietoa siitä, mikä on mahdollista ja miten tekoälyä voisi hyödyntää ja sen koettiin myös ruokkivan uusien mahdollisuuksien tunnistamista. Alan ei koettu eroavan tekoälyn suhteen muista aloista, varsinkaan viranomaisten päätöksenteosta tai finanssialasta. Kiinteistöpuolelle nähtiin olevan tekoälyn suhteen valtavaa potentiaalia ja toisaalta sijoitustoiminnassa valtavia kansantaloudellisia riskejä. Myös uhkaa imagolle sivuttiin. Koettiin, ettei luottamus ole vain tekoälystä kiinni vaan yleistä luottamusta koettiin voivan myös murentaa niiden rajojen hakemisella, mikä toiminta kuuluu työeläketoimialalle ja mikä ei, kuten Finanssivalvonnan valvonnan kohteista on voitu mediassakin lukea:

> ”Jos ollaan sen perustekemisen äärellä, mitä joku pitkään toiminut finanssialan toimija esimerkiksi on tehnyt, ja jos sä teet sitä niillä säännöillä ja tavoilla, millä sä olet kertonut ja käytät tietoja niin kuin olet sanonut, niin silloinhan sinä olet sen luottamuksenarvoinen, mutta jos sä teet jotain muuta, joka ei olekaan niin läpinäkyvää, selkeää, niin sähän murennat sitä kulmaa siellä silloin samalla koko ajan.”

Luottamus nähtiin olevan työeläketoimialan ytimessä, mutta koettiin, että siitä myös ollaan tietoisia: *”kyllähän me isoa luottamusviittaa kannamme ja sitähän mitataankin* [eläkebarometrillä]*, että miten paljon työeläkkeeseen, sitä kautta koko sektoriin, luotetaan”.* Työeläketoimialasta todettiin myös, että alan tehtävänä on toimia pitkäjänteisenä työeläketurvan toimeenpanevana tahona ja on myös tekoälyn suhteen huomioitava, että myös tulevat sukupolvet saavat heille kuuluvan eläkkeensä: *”Jos tulee vaikka jotain järjestelmämuutoksia, niin se, että niissä otettais tekoäly käyttöön ihan keskiössä, niin se luultavasti vaatii aika pitkän selvittelyn tai todella pitkän, että tiedetään, että ne metodit on varmasti sitten kurantteja edelleen sen 100 vuoden päästä.”*

Työeläketoimialan osalta alaa ohjaavat lainsäädännön lisäksi esimerkiksi tutkimustoiminnan eettiset periaatteet, joihin ei koettu tekoälyn tuoneen päivittämistarpeita, ja muutoin alan koettiin odottavan EU-tasoisia ja kansallisia linjanvetoja:

> ”olemme mukana ja vaikutamme osaltamme ja sitten kun olemme olleet eri tahoissa mukana ja vaikuttaneet ja sieltä niitä tuloksia syntyy, niin sen jälkeen hän me ollaan niin kuin siinä tilanteessa, että me sovellamme niitä organisaatioissa ja pidämme huolen siitä, että meidän työntekijät myös tietävät ne asiat eli koulutamme, ohjeistamme, opastamme.”

Työeläketoimiala on mukana kansallisessa tekoälyverkostossa ja ulkopuoliset asiantuntijat kokivat, että se on se kanava, josta tietoa asiasta saadaan ja kanava, jota kautta kansalliseen valmisteluun, ja kansallisen valmistelun kautta EU-tason valmisteluun, voidaan vaikuttaa. Koettiin, että alalla on pitkään jo olleet yhteiset lainsäädännön soveltamisohjeet riippumatta ratkaisujen teknisestä toteutustavasta, mutta että itse ratkaisujen teknisestä toteutuksesta eri toimijat ovat vastanneet itse.

5.2.9 Huomiot ulkoisista haastatteluista

Yleisesti ottaen haastattelut vahvistivat käsityksiäni siitä, mitä voidaan pitää tekoälyn etiikan normatiivisena ytimenä. Suurimpia haasteina tekoälyn luottamukselle ulkopuoliset asiantuntijat mainitsivat läpinäkyvyyden ja selitettävyyden, datan ja yksityisyydensuojan, vastuuvelvollisuuden, yhteisen viitekehyksen ja lainsäädännön puutteen, riskien- ja vaikutustenhallinnan mutta myös datan ja tekoälyn kuluttaman energian sekä kulttuurin ja osallistamisen puutteen (kts. liite 4).

Haasteiden ulkopuolelta haastatteluista nousi neljä uutta teemaa, joiden merkitystä jäin miettimään. Ensinnäkin tulin yllätetyksi yhden vastaajan ajatuksella, että tekoäly voitiin nähdä myös *”business as usual”*. Vaikka data-analytiikkaa ja koneoppimista on tehty jo vuosikaudet, implisiittinen näkökulmani opinnäytetyössä oli ollut se, että tekoäly ja sen vaikutusten arviointi ravistelee ajatteluamme ja pakottaa esiin vanhoja konventioita. Teemasta heräsi kysymys, missä vaiheessa tekoälyn "as usual" alkaa vai päättyykö se sitten, kun

tekoälyn autonomisuus tai adaptiivisuus kasvaa. Voi olla, että asia on myös kontekstisidonnainen: suosittelualgoritmi tai merkityksellisiä asianhaaroja etsivä algoritmi toteuttaa tuttua tehtävää vain hieman älykkäämmin, mutta esimerkiksi itseohjautuvien autojen osalta teknologiaa käytetään uuden mahdollistamiseen, jolloin sitä tuskin voi pitää niin "as usual". On myös mahdollista, että asialla tarkoitettiin enemmän datanhallintaan ja käsittelyyn liittyviä seikkoja, sillä tekoälyn etiikkaan liitetään paljon sellaista, mikä ei sinänsä ole uutta tai vaadi uutta – etenkin, jos eettisiä ja juridisia kysymyksiä on totuttu käsittelemään.

On kuitenkin huomautettava, että Ollilan mukaan (2019, 69, 71) liike- ja työelämässä tekoälyä pyritään kuvaamaan työkaluna, jotta tekoälytuotteet eivät tuntuisi niin uhkaavilta, vaikka tekoälytuotteita yleensä leimaakin *”toimijankaltaisuus ja vuorovaikutteisuus”,* jotka jo aiemmin todettiin ongelmallisiksi käsitteiksi.

Toinen selkeä noussut teema oli datan- ja tekoälynlukutaito sekä tietoisuuden lisääminen. Teemat, jotka tämän ympärillä puhututtivat, liittyivät niin datankeruumenetelmiin, eri tekoälytoimijoihin, tekoälyn edellytyksiin kuten läpinäkyvyyteen ja selitettävyyteen kuin energian käyttöönkin. Yhden vastaajan mielestä kuluttajien tulisi olla tietoisempia siitä, mitkä tahot ovat tekoälyratkaisujen takana, punnita luottamustaan myös näihin tahoihin ja ennen kaikkea vaikuttimiin, jotka näiden tahojen toimintaa ohjaavat. Toisen vastaajan mielestä ihmisten tulisi lisätä tietoisuuttaan myös datankäytön ympäristövaikutuksista, kuten valtavasta, tarvittavasta energiamäärästä, ja että dataa hyödynnettäessä tulisi todellista tarvetta ja käyttökohdetta arvioida myös tältä kannalta. Datan-, median- ja tekoälynlukutaito mainittiin usein, mutta esiin tuli myös datan visualisoinnin lukutaito, jotta *”tämmöiset tehokkaat välineet ymmärtää maailmaa ja viestiä siitä -- ei jää vain sen kaikista parhaiten koulutetun ja parhaiten resurssoidun eliitin välineeksi”,* mikä kielii toisaalta uudenlaisen lukutaidottomuuden eriarvoittavasta vaikutuksesta.

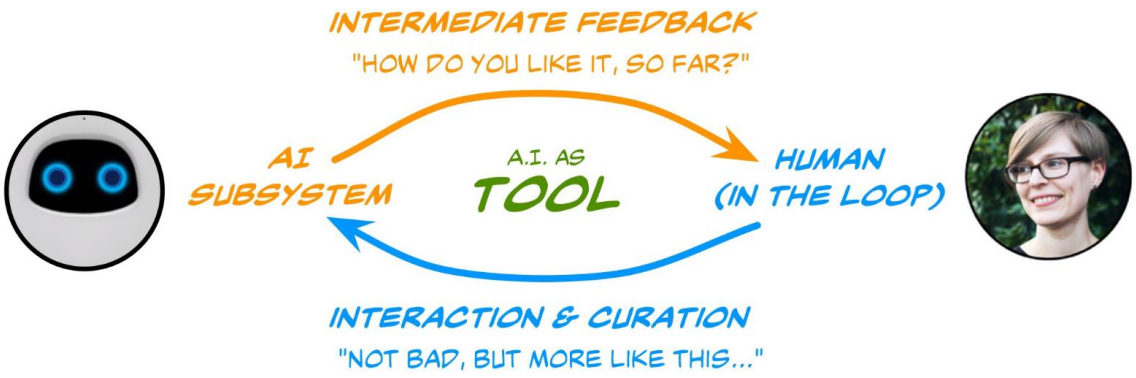
Tietoperustan mukaan kysymys ei ole vain eriarvoistumisesta, vaan seuraukset lukutaidottomuudesta voivat olla vieläkin laaja-alaisempia. Esimerkiksi selitettävyyttä lisäävät visualisointityökalut voivat auttaa havaitsemaan puuttuvia arvoja käytetystä datasta, mutta liiallisen luottamuksen lisäksi ihmiset voivat lukea visualisointeja väärin. Tutkimuksissa on ilmennyt, että käyttäjät eivät välttämättä osanneet edes sanoa, mitä selitettävyyttä lisäävät visualisoinnit näyttivät, mikä johti virheellisiin oletuksiin datasta, malleista ja itse tulkintatyökaluista. Ihmiset luottivat visualisointeihin myös silloin, kun niitä oli manipuloitu näyttämään järjettömiä selityksiä, ja visualisointeja käytettiin tekoälymallien käyttöönottopäätösten tukena myös silloin, kun ihmiset eivät ymmärtäneet niiden takana olevaa matematiikkaa. (Heaven 2020.)

Kolmas selkeä teema oli eri aikaulottuvuudet. Niin termejä, haasteita kuin ratkaisuvaihtoehtojakin punnittiin suhteessa aikaan. Tekoälyn käsitteestä puhuttaessa muistutettiin, että *”eri aikakausina se ymmärretään vähän eri tavoin”* ja että tekoäly voidaan nähdä *”viimeisimpinä tekoälybuumina”* viitaten siihen, ettei kyseessä ole uusi ilmiö vaan *”aihe, joka aina nousee aalloissa”.* Aikaulottuvuuteen liittyi myös se, että koettiin, ettei tämän hetken tekoälymalleilla voida avata mustien laatikoiden toimintaa; ne eivät osaa avata prosessia eivätkä tuloksia selityksen vastaanottavan tahon kannalta relevantilla ja ymmärrettävällä tavalla tässä ajassa: *”nää tekoälyjärjestelmät toistaiseksi ainakin, ennen kuin päästään, jos päästään, siihen vahvaan tekoälyyn, niin ne on tämmösiä black box:eja, ne ei osaa kertoa meille, miksi”.*

Neljäs teema oli ihmisen ja koneen vuorovaikutus. Haastatteluissa vuorovaikutusta pohdittiin niin asiakkaan ja koneen kuin työntekijän ja koneen välillä. Asiakkaiden odotusten palvelun laadun suhteen nähtiin kasvavan tekoälyn käytön lisääntyessä. Jos asiakaspalvelua siirretään koneen tehtäväksi, on haastateltavien mielestä tärkeää tehdä asiakkaalle selväksi, missä asioissa tekoäly asiakasta voi palvella ja toisaalta, mitä se ei kykene tekemään tai mihin sitä ei voida käyttää, jotta odotukset eivät jää täyttymättä. On hyvä muistaa, että vaikka tekoäly tekisi jotain *”uutta ja hienoa”,* niin se ei vielä takaa, että asiakas saisi siitä arvoa. Asiakasta voi myös hämmentää, jos hän ei tiedä asioivansa tekoälyn kanssa. Avoimen viestinnän ja myös tekoälyn rajoitteiden kertomisen lisäksi tekoälyä voisi vastaajien mielestä kehittää enemmän ihmisen rinnalle.

Työntekijän ja koneen välistä vuorovaikutusta pohdittiinkin syvemmin nimenomaan tilanteessa, jossa koneen tulisi toimia avustajana. Turvallisuuden osalta tietoperustassa todettiin, että algoritmin toiminta saattaa johtaa vakaviin seurauksiin tilanteissa, joissa se ei kykene reagoimaan oikein odottamattomissa tilanteissa. Tietoperustassa ei kuitenkaan juuri käsitelty ihmisen ja koneen vuorovaikutusta. Haastatteluissa mainittiin HITL, joten lisään siitä muutaman sanan.

HITL eli human-in-the-loop -menetelmä on yksi vuorovaikutteisen tekoälyn luonnin esimerkki, joka perustuu ihmisen ja koneen vuorovaikutukseen, jossa ihmisen tarkoituksena on auttaa konetta oppimaan. HITL-menetelmää voidaan käyttää esimerkiksi, kun algoritmit eivät ymmärrä syöttödataa tai tulkitsevat sen (esimerkiksi kuvan) väärin, kun algoritmit eivät tiedä, miten suorittaa tehtävää tai kun halutaan tehostaa joko mallien tai ihmisten toimintaa. Sen ajatus on, että sen sijaan, että kone tekisi työn ihmisen puolesta, työn tehokkuus paranee ihmisen ja koneen yhteystyöllä ja nopealla palautesyklillä (kuva 17). (Wang 2019.)



Kuva 17 Esimerkki HITL:stä (Wang 2019)

Ihmisen kognitiivisiin rajoituksiin liittyen myös Merilehto (29.9.2020) on puhunut siitä, että koska ihmiset eivät kykene tiedostamaan kaikkia omia kokemuksiaan, ajatuksiaan, tunteitaan ja tekojaan, eivät ne kykene saattamaan niitä koneidenkaan tiedoksi ja siksi koneilta jää konteksti ymmärtämättä ja siksi ihmisten ja koneiden vuorovaikutuksessa saavutetaan parempia tuloksia. Rusanen & Koskinen (2018, 49) toteavat myös, että nykyiset tekoälyjärjestelmät ovat kompleksisempia ja oppivampia kuin aiemmat työkalut, jolloin *"kognitiivisesta näkökulmasta nykyiset järjestelmät pikemminkin laajentavat tai jatkavat tutkijan tiedonkäsittelyä toisella kognitiivisella järjestelmällä kuin ovat täysin sille alisteisia"*, ja tähän mielestäni pitäisi kiinnittää paljon enemmän huomiota tekoälyn etiikan näkökulmasta.

6 Tulkinta ja johtopäätökset

Tietoperustan ja tutkimusaineiston tulosten perusteella luodut tulkinnat, johtopäätökset ja kehittämistoimenpide-ehdotukset pohjautuvat osin julkisessa opinnäytetyössä esiteltyjen tulosten lisäksi myös työntilaajan haastatteluiden ja dokumenttianalyysin tuloksiin, jotka ovat kokonaisuudessaan salassa pidettäviä tietoja.

6.1 Vastaukset tutkimuskysymyksiin

Lähestyin aihetta tavalla, josta jälkikäteen kuulin eetikko Seppäsen (29.10.2020) varoittavan: tekoälyn etiikkaa ei voi lähestyä tutkimuksen kautta niin, että rakennetaan *”totuutta pikku hiljaa”* ja sitten *”löydetään se viimeinen totuus”* ja saadaan vastauksia, vaan tekoälyn etiikka on luonteeltaan ennen kaikkea dialogia ja neuvottelua. Siksi keskityn johtopäätöksissäni vastaamaan tutkimuskysymyksiin ja kiteyttämään ne kysymykset, joista eri tahojen on eettisiä neuvotteluja mielestäni käytävä sen sijaan, että keskittyisin yksittäisten luottamuksenarvoisen tekoälyn edellytysten esiin nostamiseen.

TK1. Mitkä tekijät jo ohjaavat tekoälyn käyttöä?

Tekoälyn hyödyntämisestä säätelevät niin lait kuin etiikkakin. Niin yritystoimintaa kuin ihmisoikeuksiakin ohjaavat YK:n periaatteet, jotka on integroitu myös esimerkiksi kestävän kehityksen tavoitteisiin ja EU:n toimintaan ja joiden noudattamista voidaan pitää valtioiden ja yritysten ensisijaisena vastuuna (Liappis, Pentikäinen & Vanhala 2019, 144–145).

Viimeisten neljän vuoden sisällä eri tahot – uudenlaisina tahoina niin yritykset kuin kansalaisjärjestötkin – ovat alkaneet luoda tekoälylle omia eettisiä periaatteitaan. Tutkimusten mukaan (Fjeld ym. 2020, 4–5) näissä periaatteissa toistuvat seuraavat teemat: yksityisyys, vastuuvelvollisuus, turvallisuus, läpinäkyvyys ja selitettävyys, oikeudenmukaisuus ja syrjimättömyys, ihmisen hallitsema teknologia, ammatillinen vastuu ja inhimillisten arvojen edistäminen. Myös Euroopan komission nimittämän korkean tason asiantuntijoiden ryhmän AI HLEG:n periaatteet istuvat näihin teemoihin dokumenttianalyysini perusteella. Mittelstadtin mielestä on kuitenkin liian aikaista juhlia saavutettua konsensusta ylätason periaatteiden välillä, koska samaa konsensusta ei ole niiden keskinäisestä priorisoinnista tai käytännön implementoinnista: *”These differences suggest we should not yet celebrate consensus around high-level principles that hide deep political and normative disagreement”* (Mittelstadt 2019, 1–2).

Mittelstadt:n mukaan tekoälyn etiikka muistuttaa lääketieteen neljää klassista periaatetta: oikeudenmukaisuutta, autonomiaa, hyvän tekemistä ja vahinkojen välttämistä, mikä lienee

tarkoituksenmukaista, koska lääketieteen etiikkaa pidetään merkittävimpinä ja tutkituimpana sovelletun etiikan lähestymistapana. Tämä näkyy esimerkiksi AI HLEG:n kirjaamassa neljässä periaatteessa, jotka luovat pohjan tekoälyn vaatimuksille. Sekä lääketiede että tekoälyn kehitys pyrkivät sisällyttämään periaatteet ammatilliseen käytäntöön, mutta ongelmallista kuitenkin on, että tekoälyn kehittämisestä puuttuvat yhteinen tavoite, ammatillinen historia ja normit, tutkitut ja todistetut metodit, joiden avulla periaatteet muutetaan käytännön toimiksi, sekä oikeudelliset ja ammatillisen vastuuvelvollisuuden mekanismit, jotka sisältyvät lääketieteen etiikkaan. (AI HLEG 2019, 2, 10; Mittelstadt 2019, 1–2, 7–8.)

Tekoälyn käyttöä säätelevät kuitenkin jo monet eri lait, joista suurin osa tosin koskee vain julkisen vallan käyttöä ja joista yksikään ei ota kantaa teknologiaspesifisesti tekoälyn hyödyntämiseen. EU:n tietosuoja-asetus on kuitenkin esimerkki teknologianeutraaleista periaatteista, jotka määrittelevät, miten dataa saa hyödyntää. Lainsäädäntöä on myös valmisteilla: oikeusministeriössä pohditaan par aikaa, millä ehdoilla sääntöpohjainen automaattinen päätöksenteko voidaan sallia ja Euroopan komissiolle on lokakuussa 2020 jätetty *"mietintö suosituksista komissiolle tekoälyä, robotiikkaa ja niihin liittyvää teknologiaa koskevien eettisten näkökohtien kehyksestä"* (Oikeudellisten asioiden valiokunta 2020). Toisaalta muun muassa Viskari (26.11.2019) on todennut, ettei lainsäädäntöäkään voida luoda kattavaksi tilanteessa, jossa sovelluskohteista ei ole tarkkaa tietoa ja teknologia on uusi.

Työntilaajan konteksti on myös tekoälyn kannalta oikeudellisesti kiinnostava, koska työeläkevakuutusyhtiöt, eläkesäätiöt ja eläkekassat ovat yksityisoikeudellisia toimijoita, joiden tulee toimeenpanna työntekijän ja yrittäjän eläkelain mukaista työeläketurvaa. Työeläketoimialaa varten on laadittu oma lakinsa, samoin kuin alan toimijoiden, kuten työeläkevakuutusyhtiöiden, eläkesäätiöiden, eläkekassojen ja eläkelaitosten yhteiselimen Eläketurvakeskuksen, sekä valvontaviranomaisten kuten Finanssivalvonnan, tehtäviä ja hallintoa varten omansa. Lisäksi työeläkeyhtiöitä velvoittaa osin vakuutusyhtiölaki sekä perustuslaki, julkisuuslaki, hallintolaki ja kielilaki yhtiöiden hoitaessa julkishallinnollista tehtäväänsä, kuten antaessaan hallintopäätöksiä, vaikka työeläkeyhtiöiden ei katsotakaan olevan osa julkishallintoa eli julkista sektoria. (Sosiaali- ja terveysministeriö 2019, 19, 21–22; Oikeusministeriö 2020, 4.)

Tämä on oleellista siksi, että osasta tekoälyn edellytyksistä on säädetty työeläketoimialaakin koskevassa lainsäädännössä. Esimerkiksi luottamuksesta on säädetty hallintolaissa: *”Luottamuksensuojaperiaate edellyttää, että viranomaisen on toiminnassaan otettava huomioon oikeusjärjestyksen perusteella suojatut odotukset ja turvattava ne. Periaate tarkoittaa, että hallinnon asiakkaalla on tietyin perustelluin edellytyksin oikeus luottaa viranomaisen toimintaan ja hallintopäätöksen oikeellisuuteen ja pysyvyyteen (HE 72/2002 vp, s. 56)”.*

Yhtä lailla hallintolaissa on säädetty päätöksen perusteltavuudesta. (Vainio ym. 2020, 32, 40.)

Rajanveto yksityisoikeudellisen liiketoiminnan ja julkisen tehtävän hoitamisen välillä ei kuitenkaan ole aina helppoa. Hallinto- ja julkisuuslain osalta juridinen kysymys on, mitä työeläkeyhtiön toimintaa, ja siten myös tekoälyn käyttöä, ei ohjaisi hallinto- tai julkisuuslaki esimerkiksi neuvonanto- tai tiedottamisvelvollisuuksineen, sillä hallintolakia on noudatettava myös julkisen tehtävän hoitamisessa eikä pelkästään hallintopäätösten antamisessa. Toisaalta voidaan perustellusti myös kysyä, tuleeko yksityisen ja julkisen sektorin tekoälyn hyödyntämistä ohjata eri vaatimukset. Muun muassa EK:n Osaaminen ja digi -johtaja on todennut, että hallinnon ja yksityisen sektorin tulisi ennemmin *"kirittää toisiaan"* (Heikinheimo 5.10.2020).

Kiinnostavaa on, mitä eri tahot, kuten kansalaiset, kuluttajat, asiakkaat, työntekijät, työnantajat ja lainsäätäjät, alkavat vaatia ja miten eri tahojen odotukset ja vaatimukset saadaan kohtaamaan ja toisaalta myös, mikä vaatimuksia tulee nostattamaan. Kirittäjät ovat tähän asti olleet pitkälti erilaisia julkisuuteen paljastuneita rikkomuksia. Luottamuksen syntyyn liittyvätkin eettisten ja oikeudellisten kysymysten lisäksi oleellisesti myös uuden teknologian sosiaalinen ja yhteiskunnallinen hyväksyttävyys. Hyväksyttävyys on toisaalta sitä, miten teknologian käyttöä voidaan yhteiskunnassa oikeuttaa eli mihin sillä pyritään ja miten pyrkimykset ovat linjassa yhteiskunnallisten arvojen kanssa, ja toisaalta sitä, miten ihmiset kokevat teknologian käytettävyyden ja käytön (Koivisto ym. 2019, 6).

Myös haastatteluissa tuli esiin tekoälykentän jatkuva muuttuminen, sen legitimiteetti, siihen vaikuttavien tahojen moninaisuus ja verkostomainen tekeminen: *”Semmosta balanssointia se arjessa käytännössä sitten vahvasti on, kun tää on kuitenki tämmönen hyvin herkkäluontoinen ja myöskin vahvasti kehityksen alla oleva kenttä, eikä missään muotoa selvä.”* Yhtäältä kuulolla tulee olla, mitä kansalaisyhteiskunta sanoo ja seurata akateemista kehitystä ja regulaation muodostamista, mutta toisaalta samaan aikaan huolehtia oman toiminnan ohjaamisesta, hallintamalleista, dokumentoinnista ja vaatimusten täyttämisestä.

TK2. Mitä eettisillä periaatteilla voidaan ratkoa?

Tekoälyn eettisten vaatimusten normatiivisen ytimen, lakien ja kansainvälisten sopimusten lisäksi tekoälyn kehitystä ohjaavat ammatilliset normit. Gasserin & Schmittin mukaan ammatilliset normit voidaan jakaa kolmeen eri kategoriaan, jotka liittyvät pyrkimykseen, koulutukseen ja sääntelyyn. Osa yrityksistä korostaa velvollisuuttaan yhteiskuntaa ja yleistä hyvinvointia kohtaan, kun taas ammattiliittojen koodit keskittyvät kehitykseen ja teknisiin normeihin sitoutumista ja vastuuvelvollisuutta korostaen, ja osa normeista voidaan muotoilla

sääntelykoodeiksi. On ymmärrettävä, että osin tekoälyn kehitystä määräävät sen kontekstin ammatilliset normit, johon tekoälyjärjestelmä luodaan ja jossa sitä käytetään (esimerkiksi oikeus- tai lääketiede) ja osin ne ammatilliset normit, jotka ohjaavat tekoälyn kehitystä sen elinkaaren eri vaiheissa. Tekoälyn kehityksessä erityyppiset ammatilliset normit: uudet ja jo olemassa olevat, yleiset sekä tekoälyspesifit, sekä eri tieteenalojen normit, ovat siis osin toistensa kanssa päällekkäisiä ja vuorovaikutuksessa muun muassa siksi, että tekoäly kattaa joukon alatieteenhaaroja, erilaisia työkaluja, menetelmiä ja kehityksen vaiheita suunnittelusta ylläpitoon asti. (Gasser & Schmitt 2019, 14–16, 26.)

Luottamuksenarvoisen tekoälyn vaatimusten tai eettisten periaatteiden osalta suuri kysymys EU:n mahdollisen lainsäädännön lisäksi on, mitä ja kenen etua niillä tavoitellaan: onko yritysten eettisyydellä viime kädessä vain välinearvoa vai ohjaako mahdollisimman lokaali itsesääntely parhaiten käytännön eettistä toimintaa. Toinen kysymys on, asetetaanko tekoälylle tiukempia vaatimuksia kuin ihmiselle tai toisille teknologioille. Tätä keskustelua värittää myös tekoälyn käsitteen vakiintumattomuus ja teknologioiden jatkuva kehitys. Esimerkiksi Suomen tekoälykeskuksen eli FCAI:n puheenjohtaja on sanonut, ettei hän edes ammattilaisena osaisi sanoa, mihin teknologiset rajat tekoälylle vedettäisiin ja ettei EU:nkaan tulisi regulaatiossaan lähteä käsitemäärittelystä (Myllymäki 5.10.2020).

Lisäksi luottamuksenarvoiseen tekoälyyn liittyy oleellisesti myös kysymys, missä määrin puhutaan tekoälyspesifeistä haasteista, sillä monet tekoälyyn liitetyt eettiset kysymykset liittyvät laajemmin datan hyödyntämiseen, digitalisaatioon ja automatisaatioon. Eettisten tai juridisten sääntöjen hyveenä onkin usein pidetty teknologianeutraalia lähestymistapaa (Koulu 5.10.2020), mistä poiketen viime vuosina on alettu pohtia juuri nimenomaisesti tekoälyyn liittyviä koodistoja ja lainsäädäntöä. Erilaisissa yhteyksissä onkin aiheellisesti tullut esiin kysymys, onko esimerkiksi syrjintä ihmisen tai automaation tekemänä oikein ja vasta tekoälyn tekemänä rangaistavaa. Keskusteluun on nostettu myös kysymys siitä, onko tekoäly tuonut esille vain joukon laiskanläksyjä, jotka ovat jääneet aiemmin hoitamatta, ja nyt kun teknologia mahdollistaa enemmän, havahdutaankin uudenlaisen tarkkuuden vaatimukseen. Esimerkiksi ennen kuin voidaan pohtia tekoälyä autonomisena päätöksentekijänä, onkin pohdittava vähemmän älyllistä konetta päätöksen toteuttajana, koska jo senkin osalta on edellytykset yhteisesti sopimatta. Keskustelu olisi hyvä pitää teknologianeutraalina siltä osin, että monet vastuuvelvollisuuden ja oikeudenmukaisuuden kysymykset liittyvät myös ihmisen toimintaan ja tekoälyä yksinkertaisempiin tai sääntöpohjaisempiin ratkaisuihin.

Kolmas kysymys liittyy siihen, miten periaatteiden välisiä ristiriitoja ratkaistaan ja mitä arvotetaan. Tutkimustulokset osoittavat, että eettisen tekoälyn käytännön saavuttamiseksi eh-

dotetaan erilaisia, ja usein ristiriitaisia, toimenpiteitä. Esimerkiksi mahdollisimman kattavan ja monipuolisen data-aineiston yhteys puolueettomuuteen ja oikeudenmukaiseen tekoälyyn on ilmeinen, mutta samalla pidetään tärkeänä yksityisyyttä ja yksilön oikeutta hallita oman datansa käyttöä (Jobin ym. 2019, 16). Erilaisten näkemysten tuominen eettiseen dialogiin edellyttää myös yhteistä kieltä, mitä vakiintumaton terminologia vaikeuttaa muutoinkin kuin tekoälyn käsitteen osalta. Tekoälyn periaatteiden normatiivisen ytimen ohessa (luku 3.5.2) sivuttiin, että esimerkiksi termillä yksityisyys voidaan eettisissä periaatteissa viitata niin datan käyttöön liittyvään suostumukseen, datan käytön hallintaan, kykyyn rajoittaa datan käsittelyä, oikaisuoikeuteen, oikeuteen tulla unohdetuksi, yksityisyyteen suunnittelun lähtökohtana ja tietosuojalakien noudattamiseen. Tai näin ainakin tutkijat olivat tulkinneet ryhmitellessään periaatteita (kts. Fjeld ym. 2020, 21– 27).

Tekoälyn eettisten periaatteiden suhteen tulisi siis ensin ratkaista, mitä niillä halutaan ratkoa. On ymmärrettävä, että etiikassa – tässäkin ja ehkä juuri tässä kontekstissa – on kyse siitä, millaiseksi haluamme tulevaisuuden yhteiskunnan muokkaantuvan ja millaisia valintoja se edellyttää (Ollila 23.1.2020). Ollilan mukaan (23.1.2020) on tuomioistuimen tehtävä ratkoa jo tapahtunutta mutta etiikan tehtävä muotoilla tulevaa. Hän kuitenkin muistuttaa, että usein lainsäädäntö pohjaa vallitseviin moraalikäsityksiin, mutta eroaa moraalista siten, että lain noudattamatta jättämisestä rangaistaan, mutta moraali edellyttää *”henkilökohtaista sitoutumista”* (Ollila 23.1.2020). Tätä sitoutumista itse tarkoitan sanoessani, että luottamuksenarvoisuus, sen luominen ja ylläpitäminen, edellyttävät aktiivista toimijuutta niin lain noudattamiseksi kuin tilanteessa, jossa moraaliset kysymykset ovat laissa vielä tuntemattomia. Yksi keskusteluun tuotava iso kysymys onkin, miten periaatteiden tai vaatimusten toteutuminen käytännössä varmistetaan ja millaisin kannustimin, koska eettisyys on ennen kaikkea käytännön valintoja.

Sekä haastateltujen asiantuntijoiden että Mittelstadtin mukaan periaatteet edellyttävät sitovia ja näkyviä vastuuvelvollisuuden rakenteita, selkeitä alakohtaisia toteutus- ja valvontaprosesseja, mallien ja data-aineistojen dokumentointia, läpinäkyvyyttä ja riippumatonta eettistä auditointia. Lisäksi kehitettäessä alakohtaisia ja tapauskohtaisia ohjeita tarvitaan käytännön kokeiluja ja tapaustutkimuksia, jotka paljastavat eettisiä kysymyksiä ja antavat empiiristä ja teknistä tietopohjaa muun muassa tekoälyratkaisujen vaikutuksista. Etiikkaan ei kuitenkaan saisi suhtautua teknologisena ratkaisuna vaan prosessina ja periaatteellisiin eettisiin yhteentörmäyksiin tulisi varautua ja suhtautua positiivisesti, koska ne todistavat sekä eettistä harkintaa että ajattelun monimuotoisuutta. (Mittelstadt 2019, 9–10.)

Moraali ei kuitenkaan ”*tarjoa vetoavaa visiota”* (Seppänen 29.10.2020); eikä elämästä tule merkityksellisempää sillä, että tekoälyn normeja noudatetaan. Siksi periaatteiden valinta

edellyttää yrityksissä sekä tekoälystrategiaa että yrityksen sisäistä laajaa arvokeskustelua, mutta myös valtiollisen tason ja EU:n kannanottoja ja regulaatiota. Jos tekoälyn tai datan käytön eettisiä periaatteita ei vielä ole, arvokeskustelun lähtökohdaksi voi ottaa esimerkiksi AI HLEG:n periaatteet, joiden pohjalta tulisi miettiä, mitä periaatteet käytännön työssä tarkoittavat, mitä ne toiminnassa muuttavat ja miten niiden toteutuminen varmistetaan, koska sellaisenaan periaatteilla on uusimpien tutkimusten mukaan vain vähän vaikuttavuutta (kts. esim. Gasser & Schmitt 2019, 18).

Periaatteista tulisi käydä ilmi, kuka ne on kirjoittanut, miten, kenelle ja mihin tarkoitukseen. Niiden tulisi selventää, miksi jokaisen tulee niitä noudattaa, miten niitä noudatetaan ja miten ne on implementoitu käytäntöön. Periaatteiden tulee kertoa myös, mistä tiedetään, että periaatteita noudatetaan, miten toimitaan tilanteessa, jossa periaatteiden tulkinnat ovat keskenään ristiriidassa, mitä tapahtuu, jos periaatteita ei noudata, miten periaatteita voi kyseenalaistaa tai keneltä kysyä niistä selvennyksiä. (Mittelstadt 2019, 3, 19.)

Pelkästään omien arvojen pohjalta luoduilla eettisillä periaatteilla ei siis saavuteta luottamusta ja sen mukanaan tuomaa arvoa (Cap Gemini 2020). Mittelstadtin (2019, 8) mukaan periaatteellinen lähestyminen tekoälyn luottamuksenarvoisuuteen juridisen viitekehyksen puuttuessa voi antaa myös virheellisen kuvan luottamuksenarvoisuudesta:

> "These weaknesses in existing legal and professional accountability mechanisms for AI raises a difficult question: is it enough to define ‘good intentions’ and hope for the best? Without complementary punitive mechanisms and governance bodies to ‘step in’ when self-governance fails, a principled approach runs the risk of merely providing false assurances of ethical or trustworthy AI."

Periaatteellinen lähestymistapa edistää kuitenkin tietoisuutta, ymmärrystä ja keskustelua ja on alku.

TK3. Miten tekoälyä käytetään luottamuksenarvoisesti?

”Vasta yhteisessä keskustelussa voimme löytää arvoja ja periaatteita, joiden avulla kaikkien ihmisten ja elonkehän intressit tulevat otetuiksi huomioon” (Ollila 2019, 180).

Ymmärsin vasta työni edetessä, etten kykene vastaamaan kysymykseen: ”TK3: Miten tekoälyä käytetään luottamuksenarvoisesti?”, vastaamatta ensin varsin kattavaan kysymykseen siitä, mitä on luottamuksenarvoinen tekoäly ja mikä luottamuksenarvoisuutta uhkaa. Yhden ulkopuolisen haastateltavan mukaan luottamuksenarvoisen tekoälyn osalta ollaan jo siirrytty *mitä*-kysymyksestä *miten*-kysymykseen: *”nyt me tiedetään kyllä, mistä asioista pitää huolehtia”,* ja tästä rohkenen olla eri mieltä. State of AI in Finland 2020 -raportin mukaan (Faia 2020, 3, 6) hieman yli kolme prosenttia yli viiden hengen yrityksistä Suomessa käyttää

tekoälyä, ja tekoälyyn investoivat lähinnä ne yritykset, jotka ovat jo investoineet digitalisaatioon, jolloin kuilu eri yritysten osaamisen välillä kasvaa: *”This finding is, however, rather worrisome, as AI seems to increase the knowledge gap, or the digital divide, between organizations”*. Kyseenalaistan siis sitä, voiko kaikissa yrityksissä olla riittävää ymmärrystä tekoälyn luottamuksenarvoisesta käytöstä, jotta yrityksissäkään voitaisiin siirtyä suoraan miten-kysymykseen, olkoonkin, että oleellista on vastata jo molempiin.

Haastatteluissa luottamuksenarvoisuus nähtiin toisaalta suurena ja merkityksellisenä asiana, johon tekoälyn käytössä tulisi pyrkiä, mutta toisaalta myös hyvin subjektiivisena ja haastavana konseptina moniarvoisessa maailmassa: mitään ei voida julistaa luotettavaksi tai eettiseksi, koska luotettavuus ja eettisyys riippuvat näkökulmasta ja eri puolilla maailmaa nähdään myös demokratia ja yhteiskunta eri tavoin. Vastaajat pyörittelivät sitä, kenelle luottamuksenarvoista tekoälyä tehdään, kenen luottamusta arvotetaan ja kehen tai mihin luottamus kohdistetaan. Yhden vastaajan mukaan tekoälyn käytössä pitää pyrkiä niiden asiakkaiden tai ihmisten luottamuksen arvoisuuteen, joihin tekoälyllä pyritään vaikuttamaan. Tätä näkökulmaa itse korostaisin, sillä vaikka olisi itsestään selvää, että luottamuksenarvoisella tekoälyn käytöllä tavoitellaan asiakkaiden, kuluttajien tai kansalaisten luottamusta, on tärkeää olla tietoisia, arvioida ja tutkia sekä miten heihin pyritään vaikuttamaan että miten tekoäly heihin vaikuttaa.

Jobin:n ym. tutkimien kymmenien eettisten koodistojen mukaan luotettavalla tai luottamuksenarvoisella tekoälyllä voidaan viitata luotettavaan tekoälytutkimukseen, luotettavaan tekoälyteknologiaan, luotettaviin tekoälyn kehittäjiin, luotettaviin organisaatioihin tai luotettaviin suunnitteluperiaatteisiin ja myös niissä alleviivataan asiakkaiden luottamuksen merkitystä. Luottamus tekoälyn päätöksiin, suosituksiin ja käyttöön nähdään edellytyksenä tekoälyn lupausten lunastamiselle, vaikka toisaalta varoitetaan luottamasta tekoälyyn liikaa. Luottamus rakentuu periaatetutkimusten mukaan koulutuksesta, vastuuvelvollisuudesta, työkaluista, tekniikoista ja prosesseista, joilla varmistetaan normien ja standardien noudattaminen sekä tekoälyjärjestelmien eheyden seuraaminen ja arviointi. Tekoälyn odotetaan olevan läpinäkyvää, selitettävää, oikeudenmukaista ja ymmärrettävää tai ymmärrettävän sijaan ihmisten odotusten mukaista. Lisäksi se edellyttää vahinkojen välttämisen periaatetta, tietoisuutta henkilötietojen käytön arvosta ja sidosryhmien välistä vuoropuhelua. (Jobin ym. 2019, 12.)

Tuota vuoropuhelua tulisi myös käydä kuitenkin ensin sidosryhmien sisällä ja jotta tuota vuoropuhelua, eli eettistä keskustelua tekoälystä, voitaisiin käydä järkevällä tai tuloksekkaalla tavalla yrityksissä, tulisi kaikkien keskustelijoiden ymmärtää tekoäly kutakuinkin samanlaisena, mikä tuli vahvasti esiin niin tietoperustassa kuin haastatteluissakin. Tämä on

haastavaa muun muassa siksi, etteivät kaikki keskustele saman aikakauden tekoälystä. Asian on kiteyttänyt parhaiten Lehtimäki (29.10.2020), joka sanoo, että osan mielestä tekoäly on suoraan verrattavissa scifi-elokuviin, osa on mielikuvissaan historiassa ja mieltää tekoälyn ihmisen koodaamaksi, osan mukaan heikko tekoäly oppii itsenäisesti jäljitellen datasta ilmenevää ihmisen toimintaa ja osa saattaa olla jo tulevaisuudessa ja uskoo vahvan tekoälyn ratkovan uusiakin ongelmia autonomisesti. On siis ensin ymmärrettävä, että tekoälyllä on historiansa, aaltonsa ja suuntansa, mutta samalla ymmärrettävä, että tällä hetkellä kannattaa ratkoa tämän hetken tekoälyyn liittyviä eettisiä ongelmia ja löydettävä yhteinen ymmärrys siitä, minkä aikakauden tekoälystä ollaan keskustelemassa (Lehtimäki 29.10.2020).

Yhteinen ymmärrys on tärkeää, koska huolimatta edellä mainituista syistä pyrkiä teknologiseen neutraaliuteen, tekoälyllä on ominaisuuksia, kuten adaptiivisuus ja autonomisuus, jotka herättävät myös uudenlaisia, vaikeita eettisiä ja oikeudellisia kysymyksiä, joita voi olla syytä tarkastella myös erillisinä. Luottamuksenarvoisuuden rakentaminen saattaakin olla ennen kaikkea taitavaa tasapainoilua ja sen selvittämistä, missä määrin ja miten esimerkiksi tekoälyn käyttöä on syytä rajoittaa rajoittamatta samalla mahdollisuuksia oppia teknologiasta ja saavuttaa sillä parempia tuloksia tai missä määrin tekoälyn päätelmien tulee olla täysin selitettäviä tarkkuuden kustannuksella tai läpinäkyviä vaarantamatta yksityisyydensuojaa ja liikesalaisuuksia. Teknologinen nopeus näkyy myös kaikkialla: kaikki täytyy suhteuttaa aikaan, ajoitukseen ja ajanhetken mahdollisuuksiin.

Haastatteluvastauksissa nousi esiin myös käyttökohteen huomioiminen luottamuksenarvoisuuden vaatimuksissa. Tällaisiksi korkean riskitason kohteiksi vastaajat mielsivät muun muassa kaikki sellaiset sovellukset, joilla on yhteiskunnallista, sosiaalista, poliittista tai taloudellista merkitystä ja jotka vaikuttavat laajasti ja merkittävästi tavallisten kansalaisten elämään. Yhden vastaajan mielestä päätöksentekoa ei määrätyissä kohteissa tai määrättyjen alojen päätöksenteossa (esimerkiksi työeläketoimiala) saisi koskaan siirtää tekoälylle, vaan tekoälyllä tulisi aina olla vain avustava rooli. Tekoäly voisi esimerkiksi seuloa isosta massasta esiin huomiota vaativia kohteita *”perinteisin menetelmin analysoitavaksi”*, mutta sen ei tulisi edes suositella päätöksiä. Vastaajan mukaan *”käyttöalue voi tulla tulevaisuudessa kasvamaan”* riippuen siitä, saadaanko esimerkiksi läpinäkyvyyden ja selitettävyyden haastetta metodologisesti ratkottua.

Toinen varsin validi ja vaikea kysymys – mikä liittyy laajemmin myös ammatin, ammatillisuuden ja vastuuvelvollisuuden käsitteisiin – on kysymys siitä, voisiko tekoäly tehdä ihmistä paremman päätöksen, ja kuka olisi vastuussa siitä, ettei tekoälyä tällaisissa tilanteissa käytettäisi. Muun muassa Robbins (2019, 509) tuo esiin näkökulman, että mustat laatikot tulisi

sallia esimerkiksi diagnosoinnissa, jos tekoäly antaa diagnosointiin lisätietoa, mutta ei heikennä diagnoosia, jonka potilas olisi saanut ilman tekoälyn hyödyntämistä. Käytännössä kysymys on siitä, tulisiko tietyissä käyttökohteissa kieltää tekoälyn hyödyntäminen ihmisen korvaajana ja sallia sen käyttö ihmisen apuna ja ihmisen kanssa vuorovaikutuksessa.

Toisaalta kysymys on myös siitä, koskisivatko tekoälylle asetetut vaatimukset kaikkia tekoälyn sovelluskohteita, vai ovatko vaatimukset kontekstisidonnaisia tai arvioitaisiinko käyttökohdetta, prosessin osaa, menettelytapaa tai lopputulemaa sen mukaan, onko kyseessä julkisen tehtävän hoitaminen vai yksityisoikeudellinen toiminta vai asetettaisiinko eri kohteille erilaisia vaatimuksia perustuen vaikutus- ja riskiarviointiin. Tällöin kysymys on myös siitä, kuka korkean riskin tai suuren vaikuttavuuden määrittelee ja mihin perustuen: onko se esimerkiksi käyttötapasidonnaista vai toimialakohtaista. Lisäksi on ymmärrettävä, että riskiarvioinnit johtavat erilaisiin tuloksiin riippuen siitä, kenen näkökulmasta niitä tarkastellaan (Jobin ym. 2019, 16).

Eettisessä dialogissa tulisi siis saavuttaa yhteisymmärrys myös siitä, miltä osin tekoälyn käyttö on "business as usual", ja miltä osin se muuttaa esimerkiksi yksityisyyden, turvallisuuden, riskiarvioinnin, vaikutusten arvioinnin, auditoitavuuden tai läpinäkyvyyden vaatimusta tai vuorovaikutusta asiakkaiden, työntekijöiden tai koneen kanssa tai niiden välillä. Sillä voi olla vaikutusta siihen, millä vakavuudella uuteen teknologiaan tai sen eettisiin kysymyksiin suhtaudutaan.

6.2 Tavoitteiden saavuttamisen ja tulosten arviointi

Opinnäytetyön tavoitteena oli selvittää, millaisia haasteita tekoälyn luottamuksenarvoisuuteen liittyy ja miten niitä voidaan ratkoa. Selvityksen perusteella tuli ottaa kantaa siihen, miten työntilaajan tulisi luottamuksenarvoisuutta edistää, tulisiko työntilaajan luoda omat eettiset periaatteet luottamuksenarvoiselle tekoälylle sekä toimittaa jatkokehitysehdotukset mahdollisine alustavine eettisine periaatteineen tutkimuksen perusteella.

Työni vastaa kaikkiin tutkimuskysymyksiin kattavasti; siitä ilmenee, mitä asioita on otettava huomioon luottamuksenarvoisen tekoälyn rakentamisessa, mitä erilaisilla tekoälyn ominaisuuksilla tai sen mukanaan tuomilla haasteilla ja edellytyksillä tarkoitetaan ja miten moninainen on tekoälyyn liittyvä velvoittavan säännöstön ja pehmeiden moraalisten arvojen verkosto.

Työn yhtenä alkuperäisenä tavoitteena oli luoda alustavat eettiset periaatteet, joista arvokeskustelua ja etiikkatyötä jatkaa. Näiden osalta totesin tutkimukseeni perustuen, ettei olisi eettisesti oikein, että loisin periaatteet vastoin tutkimustuloksia. Tekoälyn etiikan tulisi ensisijaisesti olla yksilöstä riippumatonta dialogia ja organisaation yhteistä ymmärrystä. Eetikko

Anna Seppäsen mukaan (29.10.2020) eettistä keskustelua on pidettävä neuvotteluna, jossa on oltava yhteinen kieli ja joka edellyttää myös analyyttistä ajattelua vasten normatiivisia, perusteltuja käsityksiä. Tuota kieltä ja normeja olen sen sijaan yrittänyt opinnäytetyössäni pohjustaa ja tuoda esiin esimerkiksi tekoälyn etiikan normatiivista ydintä ja käsitteiden kirjoa ja tulkittavuutta. Seppäsen (29.10.2020) mukaan etiikka on myös toimintaa; hyvä motiivi tai aie ei riitä takaamaan eettisiä ratkaisuja, minkä vuoksi olen pyrkinyt läpityön konkretisoimaan aihetta ja liittämään käytännön esimerkkejä eettisiin koodeihin.

Toimitin työntilaajalle dokumenttianalyysin tuloksena (liite 5b) julkista versiota laajemman ja yksityiskohtaisemman periaatevertailun työntilaajan olemassa olevan ohjeistuksen teemoista suhteessa muun muassa tekoälyn etiikan normatiiviseen ytimeen, AI HLEG:n periaatteisiin ja periaatteisiin, johon pohjautuu Euroopan komissiolle lokakuussa 2020 jätetty mietintö koskien tekoälyn, robotiikan ja niihin liittyvän teknologian oikeudellista kehystä. Lisäksi toimitin sisäisten haastattelujen tulokset, johtopäätökset ja jatkotoimenpide-ehdotukset perustuen kaikkeen tutkimaani ja lukemaani ja nostin sieltä esiin teemoja, jotka jatkokehityksessä tulisi ainakin huomioida (liitteet 6–9). Yritin vastata tutkimuksessa myös työnaikaiseen kysymykseen siitä, miksi muut yritykset ovat omat eettiset periaatteensa tehneet, ja toin esiin haastatteluiden tuloksena ne eettiset kysymykset, jotka työntilaajayrityksessä puhuttavat tällä hetkellä eniten. Haastatteluiden sivutuotteena syntyi myös lista potentiaalisista tekoälyn käyttökohteista (liite 10).

Uskon siis työlläni olevan vaikuttavuutta ja uskon niin julkisen opinnäytetyön kuin salattujen liitteidenkin antavan materiaalia ja ajateltavaa opinnäytetyön ulkopuolella jatkuvaa työtä varten ja toimivan myös erillisinä kokonaisuuksina.

6.3 Jatkotutkimusehdotukset

Aiheessa riittää tutkimussarkaa hyvin monialaisesti. Suomalaisten yritysten näkyvin osa lienevät vielä yritysten omat eettisen tekoälyn periaatteet, joten itse lähtisin tutkimaan tarkemmin sitä, että miten jo luodut yritysten omat eettiset periaatteet on saatu näiden muutaman kuluneen vuoden aikana muutettua käytännön tekemiseksi, toimiviksi menetelmiksi ja mitä muut organisaatiot siitä voisivat oppia. Esimerkiksi Anttisen ja Lohilahden tutkimus (2019) ei vielä ottanut selvää, miten asiakkaat kokivat periaatteiden ja käytännön välisen suhteen ja ovatko julkiset periaatteet sellaista viestintää, mitä asiakkaat kokivat tekoälyn eettisyydestä tarvitsevansa.

Tutkisin myös tekoälyprosessiin liittyvää päätöksentekomallia: kuka mistäkin päättää ja mihin päätökset perustuvat, miten eettisesti relevantit kysymykset havaitaan ja millaisia eetti-

siä prosesseja ne käynnistävät. Kiinnostavaa olisi myös ymmärtää paremmin, millaisia mekanismeja tekoälyratkaisujen erilaisten, eritasoisten ja eri ajanjakson vaikutusten arviointiin olisi käytettävissä.

Yksi kiinnostava kulma on myös se, millaisilla toimilla on todellista vaikutusta, voiko esimerkiksi algoritmien julkaisemisella parantaa tekoälyratkaisujen eettisyyttä ja legitimiteettiä, millaisessa kontekstissa ja millä edellytyksin. Toinen kiinnostava lähtö voisi olla tutkia tarkemmin erilaisia tekoälyn tarjoamia selityksiä loppukäyttäjän näkökulmasta, sillä ainakin Adadi & Berrada (2018, 52153) ja Heaven (2020) ovat tuoneet esiin, että selitettävyyden metodit ovat perustuneet aiemmin enemmän kehittäjien intuitioon kuin käyttäjien tarpeeseen ja sillä on ollut vaikutusta selitysten käytettävyyteen, tehokkuuteen ja tulkittavuuteen. Heavenin (2020) mukaan automatisoitujen järjestelmien pitäisi perustella toimintaansa samoin kuin ihminen selittäisi, mikä edellyttää, että suunnittelussa on alusta asti mukana erilaisia selityksiä tarvitsevia loppukäyttäjiä.

Selitettävyydessä ja läpinäkyvyydessä tilaa tutkimukselle olisi myös kaksisuuntaisessa vuorovaikutuksessa, koska esimerkiksi Adadi & Berrada (2018, 52155) tuovat esiin sen, että tekoälyjärjestelmän käyttäjille saattaa herätä järjestelmästä kysymyksiä, joihin tarjottu selitys ei vastaakaan, jolloin käyttäjällä tulisi olla mahdollisuus kysyä lisää. Saman tutkimuksen mukaan tulevaisuudessa voisi olla mahdollista käyttää niin sanottuja mukautuvia selitettäviä malleja, jotka kykenisivät tarjoamaan erilaisia selityksiä eri käyttäjille (asiantuntemus, ala, tausta, kiinnostuksen kohteet ja muut asiayhteyteen liittyvät muuttujat) eri tarkoituksiin (perustelu, opetus, valvonta) ja tämä aiemmin mainitun osallistavan personoinnin kanssa voisi tarjota lisää tutkimuskenttää.

Ulkopuolisten asiantuntijoiden haastatteluissa tuli esiin myös kasvava kansainvälinen trendi palkata AI policy -, AI Governance - ja AI Ethics -alueille tekijöitä sekä perustaa organisaatioihin erilaisia AI Ethics - ja AI oversight -boardeja, joiden tarkoituksena on valvoa teknologian käyttämistä ja varmistaa sisäistä valvontaa ja tekoälyn hallintaa. Tämä ei näy vielä laajalti Suomessa, joten kentällä voisi riittää mielenkiintoista tutkittavaa.

Lisäksi yksi suuri kysymys on, mikä merkitys eri tahojen luomilla periaatteilla ylipäätään on alati digitalisoituvassa maailmassa. Itse tutkisin lisää Gasserin ajatusta digitaalisesta perustuslaillisuudesta, sillä eetikot ja ihmisoikeusaktiivit ovat jo huomauttaneet, ettei ihmisten perusoikeuksia voi kirjoittaa uudelleen esimerkiksi datan tai teknologioiden eettisiksi periaatteiksi eikä etiikkaa tulisi valjastaa esimerkiksi taloudelliseksi ajuriksi.

6.4 Prosessin ja oman oppimisen arviointi

Ilmeinen haaste oli tehdä laajasta aiheesta kompakti ja ymmärrettävä kokonaisuus. Laajuus asetti haasteita erityisesti aiheen rajaamiselle. Pyrin tuomaan esiin lähinnä tällä hetkellä ja tässä ajassa vallitsevia käsityksiä, näkemyksiä, keskusteluita, haasteita ja kysymyksiä. Ollila huomauttaa (2019, 8–9), että julkisessa keskustelussa tekoälyn etiikalla tarkoitetaan usein digitaalista etiikkaa ja *”moraalisilla pulmilla jokseenkin kaikkea sitä, mitä kansalaisten keskustelussa sellaisiksi nimitetään”*. Siksi en itsekään tässä työssä ole kovin tiukasti rajannut, mikä osuus kuuluu tekoälyn etiikan alalajiin dataetiikkaan ja mikä digin eettisyyteen ja mikä on puhtaasti tekoälyn etiikkaa. Käsitteiden vakiintumattomuus oli siis toinen syy rajaamisen vaikeudelle.

Kolmas syy oli se, että toimeksianto laatia alustavat eettiset periaatteet mielestäni edellytti, että ymmärtäisin ennen kaikkea, miksi ne tutkimusten mukaan tulee laatia ja kysymys on monin tavoin vaikea, mitä en onnekseni ymmärtänyt työhön ryhtyessäni. Muun muassa haastatteluni sisälsivät paljon kysymyksiä, joiden kohdalla piti miettiä, millä tasolla ja kenen näkökulmasta asiaa on syytä tarkastella, miten vastaaja voi kysymyksen ymmärtää ja puhutaanko tässä yhteydessä työntilaajayrityksen kontekstista vai tekoälyn luottamuksenarvoisuudesta laajemmassa merkityksessä ja miten tarkkoja kysymyksiä teemahaastattelu ylipäätään edellyttää. Tekoälyä ja sen etiikkaa ei voida luoda ja kehittää kuitenkaan vain yhdestä näkökulmasta tai yhteen näkökulmaan, vaan kehityksessä on huomioitava useamman sidosryhmän edut ja vaatimukset. Laajemman lähestymiskulman avulla selvisi kuitenkin myös seikkoja, joita en kysynyt, kuten miten vähän vastaajat korostivat työntekijän näkökulmaa. En löytänyt siis asiaan oikeaa ratkaisua, ennemminkin sen suuntaista viitettä, että *”tekoäly tuo meille uudenlaisia uhkia, haasteita ja eettisiä kysymyksiä, joiden selvittämiseen ja keskustelemiseen on varattava aikaa ja resursseja”* (Ojanen 2019, 9).

Oman vivahteensa työhön toi myös aiheen ajankohtaisuus: työni aikana julkaistiin useita aiheeseen liittyviä eettisiä, tieteellisiä ja juridisia selvityksiä, arviomuistioita, lausuntoja ja artikkeleita sekä hankkeiden lopputuloksia, joista oli hyvä pysyä tietoisena. Tutkimusaihe on lisäksi hyvin monitieteinen, ja siihen liittyen esimerkiksi oikeustieteen professori Lindroos-Hovinheimo (5.10.2020) on todennut, että hyvän poikkitieteellisen tutkimuksen tekemisessä kaikkein vaikeinta on määritellä työn riittävää laajuutta suhteessa riittävään syvyyteen ja siihen en löytänyt oikeaa vastausta. Siksi yrityksissä tulisi ehdottomasti varmistaa, että tekoälyn ympärillä on riittävää ja monialaista osaamista ja tiivistä keskustelua eri erityisalojen välillä. Myös käsitteiden vakiintumattomuus aiheutti vaikeuksia tiedonkeruussa ja tulkinnassa, mutta toisaalta yksi opeista on, että teknologia, tutkimus ja käsitykset muuttuvat entistä nopeammin ja myös tutkimuksen tekeminen muuttaa muotoaan.

Tietoperustan Vastuuvelvollisuus ja auditointi -kappaleen kohdalla kyberturvallisuusasiantuntijana toimiva ystäväni totesi, että käsittelin tekoälyä *”kuin se olisi jonkin tietojenkäsittelyjärjestelmän osa, kenties osa hallintoa tai päätöksentekoa, tai olisi tuottamassa aineistoa näiden käyttöön”*, kun todellisuudessa vastuukysymykset ovat paljon suurempia. Tämä on totta, että läpi työn konteksti on – työntilaajan toimialasta johtuen – enemmän tietojärjestelmä, joka ei aja autoa eikä ole vastuussa ihmishengistä, enkä ole varma, toinko sen riittävän selvästi esiin. IEEE:kin toteaa (2017, 37), että on tärkeää, että sen kontekstin ja yhteisön sosiaaliset ja moraaliset normit, johon tekoälyratkaisut luodaan, on huomioitu. Siksi tietoperustaan ei voi suhtautua niin, että se kattaisi kaikki kontekstit, vaan siinä ennemminkin esimerkinomaisesti tuodaan esiin, mitä kyseisillä tekoälyn edellytyksillä voidaan tarkoittaa ja kuinka niitä voidaan varmistaa ja näkökulma on enemmän julkisen sektorin, koska se on mielestäni lähinnä työeläketoimialaa teknologisten ratkaisujen tarpeen osalta.

Muistuttaisin myös, ettei luottamuksenarvoisen tekoälyn kysymyksiä tulisi käsitellä vain haasteiden kautta – kuten tässä opinnäytetyössä tein – vaan myös siltä kantilta, missä määrin uusilla teknologioilla voidaan vaikuttaa arvopohjan vaalimiseen, esimerkiksi millainen rooli tekoälyllä on ihmisoikeuksien varmistajana tai ekologisten haittavaikutusten ehkäisijänä. Valtioneuvoston raporttikin toteaa (Koivisto ym. 2019, 9), että varsinkin viranomaistoiminnassa tekoälyä käsitellään usein lähinnä ongelmien aiheuttajana, jolloin saatetaan unohtaa, että tekoäly on ennen kaikkea mahdollistaja, jonka avulla voidaan saavuttaa monia toivottuja laatukriteerejä, kuten *”nopeus, aikariippumattomuus, yhdenmukaisuus, reilu kohtelu, säännönmukaisuus, isojen volyymien käsittelykyky, resurssitehokkuus, uusien palvelujen mahdollistuminen ja datan maksimaalinen hyödyntäminen”.* Raportti muistuttaa myös, että tekoälyn avulla voidaan *”jopa ratkaista eettisyyteen ja yhteiskunnalliseen hyväksyttävyyteen liittyviä ongelmia”* (Koivisto ym. 2019, 9).

Opin matkan varrella ennen kaikkea, että etiikka on toisistaan poikkeavia näkemyksiä ja kompromissejä, jolloin myös ehdottomuuden on suurilta osin karistava. Pyrin läpi työn saturaatiopisteen saavuttamiseen, mutta näin uuden ja monimutkaisen asian äärellä sitä on melko vaikea tavoittaa. Huomasin kuitenkin asiantuntijuuteni kasvaneen miettiessäni, voiko explicability-termin tosiaan kääntää suomeksi selitettävyydeksi ja voiko sitä pitää AI HLEG:n tavoin (2019, 2) edes periaatteena kirjaamatta, miltä osin ja millaisissa käyttökohteissa periaatteen noudattamista edellytetään.

Työni ei välttämättä tarjoa vastauksia, mutta toivon sen johdattelevan aiheen piiriin, innostavan tutustumaan aiheeseen lisää, madaltamaan kynnystä tarttua toimeen ja herättävän kysymyksiä ja eriäviä ajatuksiakin, koska niistä syntyy aitoa eettistä pohdintaa. Itse pääsin

näiden kuukausien aikana nollasta tähän; samaan pienuuteen kuin haastattelemani tekoälyetiikan asiantuntija: *”Mä oon hirveän varovainen lähtemään sille tielle, että jotain pystyisi julistamaan eettiseksi tai luotettavaksi. -- Koen tosi paljon pienuutta näiden termien edessä”.*

7 Yhteenveto

Tekoäly on ymmärrettävä joukkona alati kehittyviä teknologioita, joiden vaikutukset ulottuvat kaikkialle ja joiden kanssa olemme vuorovaikutuksessa monin eri tavoin. Niiden älykkyyttä ilmentävät ominaisuudet, kuten oppivuus, adaptiivisuus, autonomisuus ja korkea suorituskyky, ja toisaalta taas ihmisen kognitiiviset rajoitteet, asettavat uudenlaisia eettisiä, juridisia ja teknisiä kysymyksiä. Esimerkiksi teknologian etiikassa ja laissa toimija on ihminen, ei autonominen kone. Tekoälyratkaisujen reagointi yllättävällä tavalla odottamattomiin tilanteisiin tai dataan vaikeuttaa niin riskien kuin vaikutustenkin hallintaa. Näitä kysymyksiä yritykset ovat pyrkineet omien eettisten tekoälyä koskevien periaatteidensa avulla itselleen ja asiakkailleen selventämään, mutta kysymys kuuluu, miten ne on tuotu käytäntöön.

Koska tekoälyn hyödyntämisen edellytyksenä on data, luottamuksenarvoisuuden haasteet ulottuvat jo sen käyttöön: esimerkiksi tietosuojan vaatimukset datan anonymisoinnista ovat osin vastoin koneoppimisen traditioita. Tasa-arvoon ja oikeudenmukaisuuteen liittyvät kysymykset voivat kuitenkin palata myös datan laatuun ja ihmisen tekemiin epäoikeudenmukaisiin päätöksiin, vääriin tulkintoihin tai valintoihin, jotka tekoälyn hyödyntäminen tuo esiin.

Tekoälyn eettisistä edellytyksistä on tutkimusten mukaan saavutettu yhteisymmärrys. Normatiivinen ydin koostuu yksityisyydestä, vastuuvelvollisuudesta, turvallisuudesta, läpinäkyvyydestä, selitettävyydestä, oikeudenmukaisuudesta, syrjimättömyydestä, ammatillisesta vastuusta, inhimillisten arvojen edistämisestä ja siitä, että ihminen kontrolloi konetta. Yhteisymmärrys on kuitenkin vain periaatteellista. Sen lisäksi, että tulkinnat ja painotukset voivat vaihdella, tekoälyn kehittämisellä ei ole ammatillista historiaa ja normeja, oikeudellisia ja ammatillisia vastuuvelvollisuuden mekanismeja, yhteisiä standardeja tai tutkittuja menettelytapoja ja ohjeistuksia, joiden avulla periaatteet muutetaan käytännön toimiksi. Yleinen lainsäädäntökään ei pysy teknologisen kehityksen vauhdissa ja sääntely on vaikeaa: Suomessa lainsäätäjien pöydällä on vasta sääntöpohjaisten hallintopäätösten automatisointi. Perustellusti voidaan kysyä myös, ovatko tekoälyyn liittyvät eettiset ja juridiset kysymykset vasta muotoutumassa ja miten jo pelkkää kognitiivista datan- ja tekoälynlukutaitoa sekä ymmärrystä ja tietoisuutta lisätään, ettei niiden puute lisää ihmisten välistä eriarvoisuutta.

Epäselvää on myös, onko eri tahoilla tekoälyn etiikan normatiivisesta ytimestä huolimatta yhteinen intentio ja osaavatko tahot priorisoida myös välttämättömien tekoälyn edellytysten välillä ja mitä teknologiaa tai millaisia käyttökohteita edellytykset koskevat.

”Ethics is a process, not a destination. The real work of AI ethics begins now: to translate and implement our lofty principles, and in doing so to begin to understand the real ethical challenges of AI.” (Mittelstadt 2019, 10.)

Lähteet

Adadi, A. & Berrada, M. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access-lehti vol. 6 s. 52138-52160. Luettavissa: https://ieeexplore.ieee.org/document/8466590. Luettu. 30.10.2020.

AI HLEG 2020. The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment. Luettavissa: https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment. Luettu 30.9.2020.

AI HLEG 2019. Luotettavaa tekoälyä koskevat eettiset ohjeet. Luettavissa: https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMIT-TEES/JURI/DV/2019/11-06/Ethics-guidelines-AI_FI.pdf. Luettu 29.8.2020.

Ailisto, H. (toim.), Neuvonen, A., Nyman, H., Halén, M., & Seppälä, T. 2019. Tekoälyn kokonaiskuva ja kansallinen osaamiskartoitus – loppuraportti. Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja 4/2019. Luettavissa: https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/161282/4-2019-Tekoalyn%20kokonaiskuva.pdf?sequence=1&isAllowed=y. Luettu 27.7.2020.

Ailisto, H. (toim.), Heikkilä, E., Helaakoski, H. Neuvonen, A. & Seppälä, T. 2018. Tekoälyn kokonaiskuva ja osaamiskartoitus. Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja 46/2018. Luettavissa: http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160925/46-2018-Tekoalyn%20kokonaiskuva.pdf. Luettu 27.7.2020.

Anttinen, T. & Lohilahti, A-M. 2019. Katsaus tekoälyyn ja sen eettisiin periaatteisiin. Opinnäytetyö. Luettavissa: https://www.theseus.fi/bitstream/handle/10024/261267/Anttinen_Terhi%20Lohilahti_Anna-Maija.pdf?sequence=2&isAllowed=y. Luettu 5.5.2020

Anttiroiko, A. 2004. Yhteiskuntavastuu ja sen määrittelyprosessi. Luettavissa: https://trepo.tuni.fi/bitstream/handle/10024/68200/yhteiskuntavastuu_ja_sen_maarittely-prosessi_2004.pdf?sequence=1&isAllowed=y. Luettu 23.10.2020.

Arrieta, A. B., Diaz-Rodrıguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R. & Herrera, F. 2019. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Luettavissa: https://arxiv.org/pdf/1910.10045.pdf. Luettu: 29.3.2020.

AuroraAI:n Etiikka-verkosto, Haataja, M. & Latvanen, M. 2019. AuroraAI-esiselvityshanke. Etiikka-työkokonaisuuden suositukset. Luettavissa: https://www.lausuntopalvelu.fi/FI/Proposal/Participation?proposalId=6bca7323-c799-4956-a92e-7965deed5f61. Luettu 30.10.2020.

Bellamy, R. K. E, Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M. Varshney, K. R. & Zhang, Y. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. Luettavissa: https://arxiv.org/pdf/1810.01943.pdf. Luettu 10.11.2020.

Burke, B. 2020. Top Strategic Technology Trends for 2021. Luettavissa: https://emtemp.gcom.cloud/ngw/globalassets/en/information-technology/documents/insights/top-tech-trends-ebook-2021.pdf. Luettu 13.11.2020.

Capgemini 2020. Organizations must address ethics in AI to gain public's trust and loyalty. Consumers, employees and citizens will reward organizations that proactively show their AI systems are ethical, fair and transparent. Luettavissa: https://www.capgemini.com/us-en/news/organizations-must-address-ethics-in-ai-to-gain-publics-trust-and-loyalty/*. Luettu 5.11.2020*

Capgemini Reseacrh Institute 2020. The AI-powered enterprise: Unlocking the potential of AI at scale. Luettavissa: https://www.capgemini.com/wp-content/uploads/2020/07/State-of-AI_Report_Web.pdf. Luettu 5.11.2020.

Capgemini Research Institute 2019. Why addressing ethical questions in AI will benefit organizations. Luettavissa: https://www.capgemini.com/wp-content/uploads/2019/08/AI-in-Ethics_Web.pdf.

Cheatham, B., Javanmardian, K ja Samandari, H. 2019. Confronting the risks of artificial intelligence. McKinsey Quarterly -artikkeli 26.4.2019. Luettavissa: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/confronting-the-risks-of-artificial-intelligence. Luettu 4.10.2020.

Cognilytica Research 2020. Global AI Adoption Trends & Forecast 2020. Patterns of Worldwide Adoption of AI. Luettavissa: https://www.cognilytica.com/download/global-ai-adoption-trends-forecast-2020/. Luettu 30.5.2020.

Dain Studios s.a. Elo parantaa asiakaskokemustaan datan avulla. Luettavissa: https://www.itewiki.fi/p/elo-parantaa-asiakaskokemustaan-datan-avulla. Luettu 25.9.2020.

Eläketurvakeskus s.a. Laskuperusteet ja vakuutusehdot. Luettavissa: https://www.etk.fi/suomen-elakejarjestelma/hallinto-ja-valvonta/tyoelakejarjestelman-saantely/laskuperusteet-ja-vakuutusehdot/. Luettu 6.11.2020.

Euroopan komissio 2020a. Tekoälystä – Eurooppalainen lähestymistapa huippuosaamiseen ja luottamukseen. Valkoinen kirja 19.2.2020. Luettavissa: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_fi.pdf. Luettu: 14.6.2020.

Euroopan komissio 2020b. Statement by Executive Vice-President Margrethe Vestager on the launch of a Sector Inquiry on the Consumer Internet of Things. Luettavissa: https://ec.europa.eu/commission/presscorner/detail/en/speech_20_1367. Luettu 22.7.2020.

Euroopan komissio 2019. Tekoäly: Komissio edistyy eettisiä ohjeita koskevassa työssään. Lehdistötiedote 8.4.2019. Luettavissa: https://ec.europa.eu/commission/presscorner/api/files/document/print/fi/ip_19_1893/IP_19_1893_FI.pdf. Luettu 13.8.2020.

Faia 2020. State of AI in Finland. Luettavissa: https://faia.fi/state-of-ai-in-finland/. Luettu 15.10.2020.

Fjeld, J., Achten, N., Hilligoss H., Nagy, A. & Srikumar, M. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rightsbased Approaches to Principles for AI. Berkman Klein Center for Internet & Society. Luettavissa: https://dash.harvard.edu/handle/1/42160420. Luettu 28.8.2020

Frey, T. 2020. Taking care of business with Responsible AI. Luettavissa: https://cloud.google.com/blog/products/ai-machine-learning/taking-care-of-business-with-responsible-ai. Luettu 6.10.2020.

Gall, R. 2018. Machine Learning Explainability vs Interpretability: Two concepts that could help restore trust in AI. Luettavissa: https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html. Luettu 1.11.2020.

Gasser, U. & Schmitt, C. 2019. The Role of Professional Norms in the Governance of Artificial Intelligence. Luettavissa: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3378267. Luettu 29.9.2020.

Government of Canada 2020. Algorithmic Impact Assessment. Luettavissa: https://open.canada.ca/aia-eia-js/. Luettu 3.11.2020.

Government of Canada 2019. Directive on Automated Decision-Making. Luettavissa: https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592. Luettu 3.11.2020.

Government Digital Service 2020. Data Ethics Framework. Luettavissa: https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework-2020. Luettu 4.10.2020.

Haataja, M. 29.9.2020. Saidot.ai:n perustaja ja toimitusjohtaja. What should we all know about the AI we interact with? Aalto-yliopiston Internet Forum-yleisöluentosarja: Teknologian tulevaisuus ja Suomi. Katsottavissa: https://www.youtube.com/watch?v=_QTSjc5wGZI. Katsottu 20.10.2020.

Halila, H. 2005. Julkinen ja yksityinen toiminta työeläkelaitoksissa. Teoksessa Juhlajulkaisu Juhani Wirilander 1935–30/11–2005, s.1–13. Suomalaisen Lakimiesyhdistyksen julkaisuja C-sarja N:o 37. Helsinki.

Heaven, W. D. 2020. Why asking an AI to explain itself can make things worse. Luettavissa: https://www.technologyreview.com/2020/01/29/304857/why-asking-an-ai-to-explain-itself-can-make-things-worse/. Luettu 1.10.2020.

Heikinheimo, R. 5.10.2020. EK:n Osaaminen ja digi -johtaja. Automaattinen päätöksenteko: kone päättää vaiko ei? HiDATA, Helsinki University Legal Tech Lab ja Suomen tekoälykeskus FCAI. Keskustelutilaisuus. Helsinki. Katsottavissa: https://www.helsinki.fi/en/news/data-science-news/automaattinen-paatoksenteko-kone-paattaa-vaiko-ei-5.10.2020. Katsottu 10.10.2020.

IEEE (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems) 2017. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2. Luettavissa: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf. Luettu 1.11.2020.

Iivonen, A. 2019. Vauvakato leimannut työeläkevuotta – Politiikka kiinnosti työeläketoimialaa. Työeläkelehti 5:2019.

Ilmarinen 2018. Ilmarisen yritysvastuu 2018. Luettavissa: https://www.ilmarinen.fi/siteassets/liitepankki/ilmarinen/taloudellisia-tietoja/vuosikertomus/2018/yritysvastuuraportti-2018_fi.pdf. Luettu: 25.9.2020.

Jobin, A., Ienca, M. & Vayena, E. 2019. Artificial Intelligence: the global landscape of ethics guidelines. Luettavissa: https://arxiv.org/ftp/arxiv/papers/1906/1906.11668.pdf. Luettu 7.10.2020.

Karkiainen, A. 2018. Tekoäly apulaisena – Kokeilemisen kulttuuri. Työeläke-lehti 3:2018. Luettavissa: https://tyoelakelehti.fi/digilehti/032018/teko-ly-apulaisena-el-keturvakeskuksessa-l-hestyt-n-teko-ly. Luettu 15.8.2020.

Khaleghi, B. 2019a. The How of Explainable AI: Pre-modelling Explainability. Luettavissa: https://towardsdatascience.com/the-how-of-explainable-ai-pre-modelling-explainability-699150495fe4. Luettu 13.3.2020.

Khaleghi, B. 2019b. The How of Explainable AI: Explainable Modelling. Luettavissa: https://towardsdatascience.com/the-how-of-explainable-ai-explainable-modelling-55c8c43d7bed?sk=998bbb1d6d73722fd0d633a3cbc86b53. Luettu 13.3.2020.

Khaleghi, B. 2019c. The How of Explainable AI: Post-modelling Explainability. Luettavissa: https://towardsdatascience.com/the-how-of-explainable-ai-post-modelling-explainability-8b4cbc7adf5f?gi=a27e98a7839e. Luettu 13.3.2020.

Kiviniemi, K. 2018. Laadullinen tutkimus prosessina. Teoksessa Valli, R. Ikkunoita tutkimusmetodeihin 2, s. 73–87. PS-Kustannus. Jyväskylä.

Knowit 2018. Case: Finanssivalvonnan älykkäät robotit työkavereina. Luettavissa: https://www.knowit.fi/referenssit/pankki-rahoitus-vakuutusala/finanssivalvonta/finanssivalvonta-rpa-tekoaly/. Luettu 25.9.2020.

Koivisto, R., Leikas, J., Auvinen, H., Vakkuri, V., Saariluoma, P., Hakkarainen, J., Koulu, R. 2019. Tekoäly viranomaistoiminnassa – eettiset kysymykset ja yhteiskunnallinen hyväksyttävyys. Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja 2019:14. Luettavissa: http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/161345/14-2019-Tekoaly%20viranomaistoiminnassa.pdf. Luettu: 8.10.2020.

Korhonen, S. 2020. ”Ei jätetä datatieteilijöitä yksin päättämään asioita, jotka ei ole teknologisia” – Meeri Haataja herättelee johtajia johtajuuteen. Tivi 18.2.2020. Luettavissa: https://www.tivi.fi/uutiset/ei-jateta-datatieteilijoita-yksin-paattamaan-asioita-jotka-ei-ole-teknologisia-meeri-haataja-herattelee-johtajia-johtajuuteen/5fe8d0bb-9557-442a-bdee-5dda928c89b2. Luettu 15.9.2020.

Koskinen, I. 2017. Koneoppiminen EU:n yleisen tietosuoja-asetuksen valossa – Etenkin automaattisen päätöksenteon näkökulmasta. Pro gradu -tutkielma. Helsingin yliopisto. Luettavissa: https://helda.helsinki.fi/bitstream/handle/10138/229709/Pro%20gradu_Koskinen%20Ida.pdf?sequence=2&isAllowed=y. Luettu 22.7.2020.

Koulu, R. 5.10.2020 Helsinki University Legal Tech Lab:n johtaja. Automaattinen päätöksenteko: kone päättää vaiko ei? HiDATA, Helsinki University Legal Tech Lab ja Suomen tekoälykeskus FCAI. Keskustelutilaisuus. Helsinki. Katsottavissa: https://www.helsinki.fi/en/news/data-science-news/automaattinen-paatoksenteko-kone-paattaa-vaiko-ei-5.10.2020. Katsottu 10.10.2020.

Kurunmäki, K. 2007. Vertailu. Teoksessa Laine, M., Bamberg, J. & Jokinen, P. (toim.) Tapaustutkimuksen taito, s. 74–92. Gaudeamus. Helsinki.

Kääriäinen, J. (toim.), Aihkisalo, T., Halén, M., Holmström, H., Jurmu, P., Matinmikko, T., Seppälä, T., Tihinen, M. & Tirronen, J. 2018. Ohjelmistorobotiikka ja tekoäly – soveltamisen askelmerkkejä. Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja 65/2018. Luettavissa: http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/161123/65-2018-Ohjelmistorobotiikka%20ja%20tekoaly.pdf. Luettu: 8.9.2020.

Lehtimäki, P. 29.10.2020. Johtava konsultti. Tekoälyn etiikka: ajattelun apuvälineitä eettisesti kestäviin ratkaisuihin. Gofore Oyj. GTalks-webinaari. Katsottavissa: https://www.youtube.com/watch?v=6qYwKhy4oOQ. Katsottu 29.10.2020.

Lehtiniemi, T. & Ruckenstein, M. 2019. Eettinen tekoäly toteutuu punnituissa käytännöissä. Blogikirjoitus 7.2.2019. Luettavissa: https://etiikka.fi/eettinen-tekoaly-toteutuu-punnituissa-kaytannoissa/. Luettu: 7.9.2020.

Leinvuo, M. 2020. Kelan tekoälyn eettiset periaatteet. Työpajakokonaisuus 2019. Luettavissa: https://drive.google.com/drive/folders/1q7UsPxAb_wCa5Y0yD8KRAaemvZsWuhL7. Luettu 19.9.2020.

Leslie, D. 2019. Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. Luettavissa: https://doi.org/10.5281/zenodo.3240529. Luettu 12.9.2020.

Liappis, H., Vanhala, A., Pentikäinen, M. & Vanhala, A. 2019. Menesty yritysvastuulla: Käsikirja kokonaisuuteen. Edita. Helsinki.

Lindroos-Hovinheimo, S. 5.10.2020. Helsingin yliopiston oikeustieteen professori. Automaattinen päätöksenteko: kone päättää vaiko ei? HiDATA, Helsinki University Legal Tech Lab ja Suomen tekoälykeskus FCAI. Keskustelutilaisuus. Helsinki. Katsottavissa: https://www.helsinki.fi/en/news/data-science-news/automaattinen-paatoksenteko-kone-paattaa-vaiko-ei-5.10.2020. Katsottu 10.10.2020.

Lipton, Z.C. 2017. The Mythos of Model Interpretability. Luettavissa: https://arxiv.org/pdf/1606.03490.pdf. Luettu 13.3.2020.

Loukides, M., Mason, H. & Patil, DJ 2018. Ethics and Data Science. Kindle Edition. Saatavilla: Amazon.com. Luettu 6.11.2020.

Magoulas, R. & Swoyer, S. 2020 AI adoption in the enterprise 2020. Luettavissa: https://www.oreilly.com/radar/ai-adoption-in-the-enterprise-2020/. Luettu 30.5.2020.

Mattila, J. 2020. Algoritmi – Ketä kiinnostaa. Tekoälyn etiikka -seminaari 4.2.2020. Luettavissa: https://docs.google.com/presentation/d/1S4knAoDh1kIzTy7EteTS3fp_KQ0ZkcPCEb-VcVJHyvM/edit. Luettu 15.9.2020.

McKinsey 2019. Global AI Survey: AI proves its worth, but few scale impact. Luettavissa: https://www.mckinsey.com/featured-insights/artificial-intelligence/global-ai-survey-ai-proves-its-worth-but-few-scale-impact. Luettu 3.9.2020.

Merilehto, A. 29.9.2020. Houston Analytics:n Chief Growth Officer. Artificial Intelligence - a tool or a threat? Aalto-yliopiston Internet Forum-yleisöluentosarja: Teknologian tulevaisuus ja Suomi. Katsottavissa: https://www.youtube.com/watch?v=_QTSjc5wGZI. Katsottu 20.10.2020.

Merilehto, A. 2018. Tekoäly: Matkaopas johtajalle. Alma Talent. Helsinki.

Mittelstadt, B. 2019. Principles Alone Cannot Guarantee Ethical AI. Nature Machine Intelligence 11/2019. Luettavissa: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3391293. Luettu 1.11.2020.

Myllymäki, P. 5.10.2020. FCAI:n varajohtaja. Automaattinen päätöksenteko: kone päättää vaiko ei? HiDATA, Helsinki University Legal Tech Lab ja Suomen tekoälykeskus FCAI. Keskustelutilaisuus. Helsinki. Katsottavissa: https://www.helsinki.fi/en/news/data-science-news/automaattinen-paatoksenteko-kone-paattaa-vaiko-ei-5.10.2020. Katsottu 10.10.2020.

O’Brien, C. 2020. Finnish city Espoo pioneers civic AI with education and explainability. Luettavissa: ”https://venturebeat.com/2020/02/20/finnish-city-espoo-pioneers-civic-ai-with-education-and-explainability/. Luettu 3.9.2020.

Oikeudellisten asioiden valiokunta 2020. Mietintö suosituksista komissiolle tekoälyä, robotiikkaa ja niihin liittyvää teknologiaa koskevien eettisten näkökohtien kehyksestä.

Luettavissa: https://www.europarl.europa.eu/doceo/document/A-9-2020-0186_FI.html. Luettu 30.10.2020.

Oikeusministeriö 2020. Automaattiseen päätöksentekoon liittyvät yleislainsäädännön sääntelytarpeet. Esiselvitys 14.2.2020. Luettavissa: https://api.hankeikkuna.fi/asiakirjat/ff3444f4-24c9-4ee8-8c9d-7bc581c0021a/796dac3f-4527-45c0-a7b8-d63024345ac8/JULKAISU_20200214084153.pdf. Luettu 8.3.2020.

Ojanen, A., Oljakka, N., Sahlgren, O., Tuikka, A-M & Vaiste, J. 2019. Opas tekoälyn etiikkaan. Turku AI Societyn julkaisu. Luettavissa: https://aisociety.fi/sites/aisociety.fi/files/opas_tekoalyn_etiikkaan_v1.pdf. Luettu 5.9.2020.

Ojasalo, K., Moilanen, T. & Ritalahti, J. 2015. Kehittämistyön menetelmät. Uudenlaista osaamista liiketoimintaan. Sanoma Pro. Helsinki.

Ollila, M-R. 23.1.2020. Filosofi. Dataetiikan haasteet. Tietokiristä maratoniksi -seminaari. Helsinki. Katsottavissa: https://www.youtube.com/watch?v=y-R-PfLI2lg. Katsottu 29.9.2020.

Ollila, M-R. 2019. Tekoälyn etiikkaa. Otava. Helsinki

Perrault, R., Shoham, Y., Brynjolfsson, E., Clark, J., Etchemendy, J., Grosz, B., Lyons, T., Manyika, J., Mishra, S. & Niebles, J-C. 2019. The AI Index 2019 Annual Report. AI Index Steering Committee. Human-Centered AI Institute. Stanford University. Luettavissa: https://hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf. Luettu 30.5.2020.

Pervilä, M. 2020. Siri, Alexa ja Google Assistant joutuvat EU:n tiukkaan syyniin. Lehtiartikkeli Tivi 21.7.2020. Luettavissa: https://www.tivi.fi/uutiset/siri-alexa-ja-google-assistant-joutuvat-eun-tiukkaan-syyniin/da9f5ddd-451d-4e8b-b5b3-e867adc3c29e. Luettu 22.7.2020.

Pietikäinen, M. & Silvén, O. 2019. Tekoälyn haasteet – koneoppimisesta ja konenäöstä tunnetekoälyyn. Konenäön ja signaalianalyysin keskus Oulun yliopisto. Luettavissa: http://jultika.oulu.fi/files/isbn9789526224824.pdf. Luettu 13.3.2020.

PwC 2018. Explainable AI. Driving business value through greater understanding. Luettavissa: https://www.pwc.co.uk/audit-assurance/assets/pdf/explainable-artificial-intelligence-xai.pdf. Luettu 5.9.2020.

Ratsula, N. 2016. Compliance. Eettinen ja vastuullinen liiketoiminta. Talentum. Helsinki.

Rissanen, V. 2018. Eläketurvakeskus opetti koneelle puolen miljoonan ihmisen tietojen avulla, millaiset ominaisuudet ennakoivat työkyvyttömyyttä – Tulos: Kone osasi ennustaa sen kaksi vuotta etukäteen. HS 17.4.2018. Luettavissa: https://www.hs.fi/teknologia/art-2000005645109.html. Luettu 15.8.2020.

Robbins, S. 2019. A Misdirected Principle with a Catch: Explicability for AI. Minds & Machines 29, s. 495–514. Luettavissa: https://doi.org/10.1007/s11023-019-09509-3. Luettu 30.9.2020.

Rusanen, A-M. & Koskinen, I. 2018. Tiede, tekoäly ja tiedolliset riskit. Futura, 4, s. 48–53. Artikkeliluonnos luettavissa: https://www.researchgate.net/publication/330089291_Tiede_tekoaly_ja_tiedolliset_riskit?channel=doi&linkId=5c2cbdd2299bf12be3a82c25&showFulltext=true. Luettu 21.6.2020.

Rusanen, A-M. & Lappi, O. s.a. Tekoäly ihmisen kognitiivisena avustajana: Kysymys tiedollisista riskeistä. Luettavissa: https://vm.fi/documents/10623/10841416/Rusanen-Lappi-kognitiivinen-avustaja.pdf/b6d168dd-c79e-59ee-81a2-50ead1c63aaf/Rusanen-Lappi-kognitiivinen-avustaja.pdf. Luettu 21.6.2020.

Saaranen, P., 2014. Tutkimuksen validiteetti ja reliabiliteetti. Haaga Helian YAMK-opinnäytetöiden menetelmätyöpajan materiaali.

Saidot 2019. Is it too early to think about responsible AI development? Luettavissa: https://www.saidot.ai/post/is-it-too-early-to-think-about-responsible-ai-development. Luettu 30.10.2020.

Schmelzer, R. 2019. Understanding Explainable AI. Forbesin artikkeli 23.7.2019. Luettavissa: https://www.forbes.com/sites/cognitiveworld/2019/07/23/understanding-explainable-ai. Luettu: 14.3.2020.

Seppänen, A. 29.10.2020. Eetikko. Tekoälyn etiikka: ajattelun apuvälineitä eettisesti kestäviin ratkaisuihin. GTalks-webinaari. Katsottavissa: https://www.youtube.com/watch?v=6qYwKhy4oOQ. Katsottu 29.10.2020.

Siukonen, T. & Neittaanmäki, P. 2019. Mitä tulisi tietää tekoälystä? Docendo. Jyväskylä.

Solita 2019. The Impact of AI. How to navigate the ethical challenges of using AI in business and society. Luettavissa: https://hub.solita.fi/the-impact-of-ai. Luettu 28.8.2020.

Sorsa, V-P. 2006. Työeläkevakuutusyhtiön yhteiskuntavastuullinen osakesijoittaminen. Institutionalistinen näkökulma työeläkeyhtiöiden sijoitustoiminnan legitimiteettiin. Pro

gradu. Luettavissa: https://jyx.jyu.fi/bitstream/handle/123456789/12847/URN_NBN_fi_jyu-2006187.pdf?sequence=1. Luettu 3.10.2020.

Sosiaali- ja terveysministeriö 2019. Työryhmän raportti 19.2.2019. Sosiaali- ja terveysministeriön muistioita ja raportteja 2019:15. Luettavissa: https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/161385/STM_R15_Elakejarjestelmien_erillisyys_loppuraportti.pdf?sequence=1&isAllowed=y. Luettu 21.5.2020.

Sosiaali- ja terveysministeriö s.a. Lainsäädäntö. Luettavissa: https://stm.fi/vakuutusasiat/lainsaadanto. Luettu 14.6.2020.

Steniche, R. 2019. Myth Busting AI. Luettavissa: https://neurospace.io/blog/2019/10/myth-busting-ai/. Luettu 29.3.2020.

Tietosuojavaltuutetun toimisto s.a. Pseudonymisoidut ja anonymisoidut tiedot. Luettavissa: https://tietosuoja.fi/pseudonymisointi-anonymisointi. Luettu 15.10.2020.

Turek, M. s.a. Explainable Artificial Intelligence (XAI). DARPA-BAA-16-53. Luettavissa: https://www.darpa.mil/program/explainable-artificial-intelligence. Luettu 13.3.2020.

Työ- ja elinkeinoministeriö 2019. Edelläkävijänä tekoälyaikaan – Tekoälyohjelman loppuraportti 2019. Luettavissa: http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/161447/23_19_Tekoalyraportti.pdf. Luettu 7.9.2020.

Vainio, N. Tarkka, V. & Jaatinen, T. 2020. Arviomuistio hallinnon automaattiseen päätöksentekoon liittyvistä yleislainsäädännön sääntelytarpeista. Oikeusministeriön julkaisuja, Selvityksiä ja ohjeita 2020:14. Luettavissa: http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/162355/OM_2020_14_S0.pdf?sequence=1. Luettu 18.9.2020.

Valtioneuvosto 2019. Osallistava ja osaava Suomi – sosiaalisesti, taloudellisesti ja ekologisesti kestävä yhteiskunta. Pääministeri Sanna Marinin hallituksen ohjelma 10.12.2019. Valtioneuvoston julkaisuja 2019:31. Luettavissa: http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/161931/VN_2019_31.pdf?sequence=1&isAllowed=y. Luettu 7.9.2020.

Varis, J. 2018. Eläketurvakeskuksen koneoppimiskokeilu – näin se tehtiin! Eläketurvakeskuksen blogikirjoitus 17.4.2018. Luettavissa: https://www.etk.fi/blogit/elaketurvakeskuksen-koneoppimiskokeilu-nain-se-tehtiin/. Luettu 15.8.2020.

Varma 2019. Tekoälyä hyödyntänyt Varman selvitys: Asiakastyön vaatimukset yhteydessä työkyvyttömyyteen. Luettavissa: https://www.varma.fi/muut/uutishuone/uutiset/2019-q1/selvitys-tyokyvyttomyyselakehakemuksista/. Luettu 25.9.2020.

Vatanen, K. 2017. Ihmisälyn valmentamaa salkunhoitoa. Varman blogikirjoitus 22.11.2017. Luettavissa: https://www.varma.fi/muut/blogi/postaukset/2017-q4/ihmisalyn-valmentamaa-salkunhoitoa/. Luettu 25.9.2020.

Viskari, J. 26.11.2019. Digi- ja väestötietoviraston pääjohtaja. Tekoälyn ja datan vastuullinen hyödyntäminen? – Toiveuni vai painajainen. Finanssivalvonta. Digitalisaatio, tekoäly ja datan käyttö muuttavat finanssipalveluja, kuinka käy asiakkaan? -vuosiseminaari. Helsinki. Katsottavissa: https://www.youtube.com/watch?v=MMk0TtZzd2k. Katsottu 24.10.2020.

Vuori, M. 2019. Tekoälyn kehittämisen ja soveltamisen etiikasta systeemisestä näkökulmasta. Luettavissa: https://www.mattivuori.net/julkaisuluettelo/liitteet/tekoalyn_kehittamisen_etiikasta.pdf. Luettu 30.9.2020.

Wang, G. 2019. Humans in the Loop: The Design of Interactive AI Systems. Blogi-kirjoitus. Luettavissa: https://hai.stanford.edu/blog/humans-loop-design-interactive-ai-systems. Luettu 30.8.2020.

Wennberg, S. 2020. Tekoäly jouhevoittaisi vakuutus- ja eläkepäätöksiä. Luettavissa: https://www.finanssiala.fi/uutismajakka/Sivut/Tekoaly-jouhevoittaisi-vakuutus--ja-elakepaatoksia.aspx. Luettu 25.9.2020.

Liitteet

Liite 1. Käsitemäärittely

Aggregointi	Aggregoinnilla tarkoitetaan tässä yhteydessä tietojen yhdistämistä niin, ettei yksittäistä henkilöä kyetä tunnistamaan henkilöryhmästä, joka aineistossa kuvaa yksittäisiä henkilöitä.
Algoritmi	Algoritmi on täsmällinen, vaiheittainen kuvaus matemaattisen tai jonkin muun ongelman ratkaisemiseksi, ja joka koostuu yksittäisistä, selkeistä, mekaanisista vaiheista.
Attribuutti	Attribuutteja kutsutaan koneoppimisessa usein ominaisuuksiksi, jolloin tarkoitetaan attribuuttia ja sen saamaa arvoa. Attribuutteja voivat olla esimerkiksi ikä, sukupuoli ja diagnoosi, jotka liittyvät johonkin tapahtumaan tai tapaukseen, kuten henkilöön.
Auditointi	Auditoinnilla tarkoitetaan esimerkiksi tekoälyjärjestelmän riippumatonta ja puolueetonta hyväksyntätarkastusta määriteltävää auditointikriteeristöä vasten. Auditoinnin tarkoituksena on todistaa, että käytännöt vastaavat vaatimuksia. Auditointi voi kohdistua myös dataan.
Demografinen	Demografisella datalla tai demografisilla tekijöillä tarkoitetaan yksilöiden ominaisuuksia kuten esimerkiksi ikää, sukupuolta, koulutusta, tulotasoa, asuinaluetta tai perhetietoja.
Kognitiiviset toiminnot	Kognitiivisilla toiminnoilla tarkoitetaan ihmisen kykyä esimerkiksi havainnoida, tunnistaa, ymmärtää, muistaa ja oppia. Älykkyyttä voidaan ajatella kognitiivisena kyvykkyytenä.
Legitimiteetti	Legitimiteetillä tarkoitetaan hyväksyttävyyttä tai traditioihin, moraaliin tai lakiin perustuvaa oikeutusta eli esimerkiksi kansalaisten luottamusta järjestelmää, toimintaa tai tahoa kohtaan.
Parametri	Parametrit ovat koulutusdatan ominaisuuksia, esimerkiksi painoja, lukumääriä tai kertoimia, joiden arvon, algoritmi oppii harjoitellessaan, joita säädetään optimointialgoritmein tai jotka voivat olla ihmisen etukäteen asettamia ja staattisia. Parametri voi olla esimerkiksi ihmisen ennalta määrittelemä lukumäärä algoritmin muodostamia klustereita.
Python	Python on ohjelmointikieli, jota käytetään esimerkiksi koneoppimisessa.
R	R on toinen suosittu avoimen lähdekoodin ohjelmointikieli.
Sertifikaatti	Sertifikaatti on todistus siitä, että esimerkiksi jokin tuote, palvelu tai toiminta on validoitu jonkin laatujärjestelmän mukaan.
Validointi	Menettely, jolla varmistetaan, että esimerkiksi opetettu malli toimii käyttötarkoituksessaan eli että se toimii esimerkiksi erilaisilla, riippumattomilla data-aineistoilla.

Verifiointi	Menettely, jolla todennetaan, että määritellyt vaatimukset toteutuvat.
Vinouma	Vinoumalla tarkoitetaan tässä yhteydessä perusteetonta, suosivaa, puolueellista tai kohtuutonta implisiittistä tai eksplisiittistä taipumusta, arvotusta, suuntausta tai mielipidettä.

Liite 2. Ulkopuolisten asiantuntijoiden haastattelukysymykset ja teemat

Haastatteluiden pääteemat olivat:

- Mistä rakentuu luottamuksenarvoinen tekoäly?
- Mitkä ovat kolme suurinta haastetta tekoälyn käytössä luottamuksenarvoisuuden näkökulmasta?
- Mitkä ovat tämän hetken tärkeimmät toimet (tahot, hankkeet, ekosysteemit, näkökulmat, tuotteet, hyödyntäjät, palvelut) edistää luottamuksenarvoisuutta tekoälyn käytössä?

Kaikissa haastatteluissa käytiin läpi muun muassa seuraavia tarkempia kysymyksiä:

- Mitä on tekoäly? Mistä koostuu luottamuksenarvoinen tekoäly?
- Mitkä ovat kolme suurinta haastetta tekoälyn opettamisessa tai hyödyntämisessä luottamuksenarvoisuuden näkökulmasta?
- Miten edellä mainittuja haasteita tulisi ratkoa?
- Minkä tekijöiden tai tahojen pitäisi ohjata datan ja tekoälyn hyödyntämistä? / Mitkä tekijät ja tahot ohjaavat tekoälyn hyödyntämistä omalla alallasi?
- Mitkä ovat tämän hetken tärkeimmät toimet ja tahot luottamuksenarvoisen tekoälyn edistämisessä?

Myös esimerkiksi seuraavia kysymyksiä käsiteltiin:

- Näetkö alakohtaisia uhkia tai mahdollisuuksia liittyen tekoälyyn?
- Kohtaatko datan tai tekoälyn luottamuksenarvoisuuteen liittyviä kysymyksiä omassa työssäsi ja jos kyllä, niin millaisia?
- Mitkä ovat tekoälyn vaikutukset tieto- tai kyberturvaan?
- Miten suhtaudut yritysten omiin tekoälyn eettisiin periaatteisiin?
- Mitä tekoälyn saralla tapahtuu seuraavan viiden vuoden sisällä? Entä luottamuksenarvoisuuden saralla?

Liite 3. Työntilaajan asiantuntijoiden haastattelukysymykset ja teemat

Haastatteluiden pääteemat olivat tekoälyn hyödyntäminen yrityksessä, tekoälyn käyttöä ohjaavat tekijät, yksityisen toimijuuden ja lakisääteisen tehtävän välinen ero tekoälyn kannalta, tekoälyn haasteet ja luottamuksenarvoisuuden kysymykset erityisesti työntilaajan kontekstissa ja suhtautuminen tekoälyn eettisiin periaatteisiin.

Kaikissa haastatteluissa käytiin läpi seuraavia tarkempia kysymyksiä:

- Mitä on tekoäly? Mistä koostuu luottamuksenarvoinen tekoäly?

- Millaisissa yhteyksissä tekoäly on tullut puheeksi yrityksessä? / Miten näette tekoälyn mahdollisuudet toiminnoissanne?
- Mitkä tekijät ohjaavat tekoälyn kehittämistä ja hyödyntämistä yrityksessä? / Millainen käsitys on tekoälyn käytön nykytilasta?
- Onko yksityisen toimijuuden ja lakisääteisen tehtävän hoitamisen välinen ero selvä? Asettaako se rajoitteita tekoälyn käytölle?

- Mitkä ovat kolme suurinta haastetta tekoälyn käytössä luottamuksenarvoisuuden näkökulmasta?
- Miten ja kenen edellä mainittuja haasteita tulisi ratkoa?
- Näkyvätkö kolme suurinta haastetta yrityksessä?
- Minkälaisia muita tekoälyn eettisyyteen, luotettavuuteen tai vastuullisuuteen liittyviä kysymyksiä on tullut tai voisi tulla vastaan?

- Minkä tekijöiden tai tahojen pitäisi ohjata datan ja tekoälyn hyödyntämistä?
- Tulisiko yrityksen luoda omat eettiset periaatteet tekoälyä varten? Miksi?
- Millaisena yrityksen tulisi tekoälyn hyödyntäjänä profiloitua ja viestiä?

Liite 4. Haasteet ulkopuolisten asiantuntijoiden mielestä

Haaste	Kuvaus / seuraus	Ratkaisu
Läpinäkymättömyys, selitettävyys (H1, H2, H3)	– Ihmiset eivät ymmärrä, mitä tekoälysovellukset tekevät: *"oikeastaan kukaan ulkopuolelta ei pysty osallistumaan siihen, ottamaan mitään kantaa"* – Tekoäly ei osaa kertoa, miksi se päätyy sellaiseen tulokseen kuin päätyykään sille annetuilla inputeilla, kykenemättömyys itsereflektioon – Vaikka kone antaisikin selityksen, onko asiantuntijalla saati loppukäyttäjällä kykyä lukea monimutkaisia selityksiä prosessista – Perustelun vaade, muutoin mielivaltaista ja riskialtista – Riski myös manipulointiin: kouluttaminen datasetillä, jolla voidaan saavuttaa haluttuja päätöksiä – Tarve kontekstisidonnainen – Tärkeää ymmärtää myös, että jos malli ei toimi oikein 1%:ssa tapauksia, mitä silloin tapahtuu – Mallin valinta, jolloin *"jää aika paljon sille yhdelle henkilölle sitä harkintavaraa"*	– Yhteinen, selkeä hallintamalli, rakenne osaksi kehitysprosessia, jolloin prosessissa huomioidaan eri näkökulmat (yhdenvertaisuus ym.), ohjeistus ja hallintamalli linjassa esim. EU:n kanssa – Dokumentointivaatimukset – Kun läpinäkyvyys on prosessissa, *"se väistämättä ohjaa semmoseen järjestelmälliseen, hallittuun tapaan toimia ja sit se haastaa meitä ittee"* – Riskien valvonta – Avoin viestintä, tiedot dataseteistä, millaisia asioita on huomattu ja miten ne on huomioitu – Tekninen ongelma: esimerkiksi selitettävien mallien tai muiden teknisten ratkaisujen tarve, mutta tarvitaan toisaalta myös kysyntää niille – Modulaarinen ajattelu: kokonaisuuksien pilkkominen esimerkiksi *"sarja neuroverkkoja, jotka tekee pienempiä päätöksiä"* – Ihmisen harkinnan lisääminen päätöksentekoprosessiin – Rajataan mustien laatikoiden käyttöä riskialttiissa kohteissa – Visualisoinnilla voi olla rooli osoittaa vinoumia, piileviä kuvioita, rakenteellista rasismia, jonka tekoäly on automatisoinut malliin tai kuvata, miten eri data verkostossa tai nettisivulla liikkuu ja mitkä toimijat keskustelevat keskenään
Datan käsittely luottamuksellisesti ja teknisesti oikein, GDPR ja yksityisyydensuoja (H3, H2, H4)	– Ei tietoa, käsittelevätkö kaikki toimijat dataa samojen sääntöjen mukaan – Lähtödatassa on ja sinne jää vinoumia – Yksityisyydensuojan mureneminen – Käyttäjät eivät ymmärrä, ettei kyse ole vain kissavideoista, vaan riski on suurempi eikä välttämättä henkilökohtainen – Käytetty aineisto on vanhaa – Datan ja vallan keskittyminen	– Henkilötiedon pseudonymisointi – Dataan pääsyn rajoittaminen, datan käyttöoikeuden varmistaminen, henkilötietojen käytön valvonta – Datan asianmukainen suojaaminen, myös tietoturvan kannalta – Tietosuojalain noudattaminen – Kuluttajan ymmärryksen lisääminen – Riskin ymmärtäminen – AI:n rajoitteiden ymmärtäminen – Läpinäkyvyyden vaatimus – *"Piirteet on muokattu oikein"* – Esimiehen vastuu datankäytöstä ja käyttökohteesta – Etukäteen sovitaan, mitä tietoa käytetään ja miten sitä käytetään – Tietoa ei viedä käsiteltäväksi ulkopuolelle – Kumppani noudattaa sovittuja valintoja – Henkilötietojen käytön valvonta – Käyttökohteen eettisyyden varmistaminen
Vastuuvelvollisuus (H1, H4)	– Jonkun täytyy olla vastuussa tekoälyjärjestelmästä – Järjestelmän suunnittelijan, kehittäjän, ostajan, käyttäjän ja myyjän yhteisvastuu	– Läpinäkyvä viestintä – Tunnistetaan, nimetään ja viestitään henkilöistä, jotka kantaa vastuun: sovelluksen omistaja (yritys/toiminto/osasto), kontaktihenkilö ulkopuoliseen viestintään (rooli/yhteystiedot), liiketoimintaomistaja (kehitys ja riskit), tekninen omistaja (tekninen kehitys ja näkökulma) – Vertaisarvioija organisaatiossa, sisäinen auditoija, teknisen omistajan tukena – Tunnistetaan toimitusketju, tunnistetaan ulkoiset toimittajat – Ulkopuolinen auditoija (nimetään kuka) – Luodaan konkretiaa, että vastuullisetkin voivat olla luottavaisin mielin – Kehittäjän rehellisyys ja omien näkemyksien esille tuonti – Hankintaosaaminen, että voi kantaa vastuun – Saidot.ai:n alusta hyvä alku
Yhteisen viitekehyksen ja lainsäädännön puute (H1, H3)	– Paljon puhetta ja näkökulmia, mutta kenttä on epäselvä ja yhteiset vaatimukset puuttuvat – Epäselvää, miten toimitaan luottamuksenarvoisesti – Ei ole säädöspohjaa sille, mitä teknologia jo mahdollistaa – Kaikki ei ole yksilön kannalta turvallisuussäädeltyä – Lainsäätäjillä ei välttämättä ole myöskään riittävää tekoälyosaamista	– "O*lemme mukana* [lainsäädännön valmistelussa] *ja vaikutamme osaltamme"* – Kansallinen tekoälyverkosto kanava, jota kautta kansalliseen valmisteluun, ja kansallisen valmistelun kautta EU-tason valmisteluun, voidaan vaikuttaa.

Riskienhallinta ja vaikutusten arviointi (H4)	– Esimerkiksi tiedon suojaaminen ja datanhallinta ei saa innostuksessa unohtua – On ymmärrettävä, että tekoälyjärjestelmässä on *"virheitä jatkuvasti, on virheitä, jotka minä olen löytänyt ja laittanut kuntoon, siellä on varmasti myös virheitä, joita minä en ole vielä löytänyt", "tekniikka ei ole täydellistä"* – Datanhallinta, tiedetään *"mitä emme voi menettää",* esimerkiksi tietomurroissa – Suomalaisilla olisi hyvä olla aito kiinnostus katsoa konepellin alle, mutta ajatus enemmän on ollut, että *"ei se tavallaan mulle kuulu"* – Ihmisen ja yhteiskunnan, hyvinvoinnin ja sosiaalisten vaikutusten ymmärtäminen ja analysoiminen – Algoritmit voivat olla täysin rationaalisia ja samalla täysin epäeettisiä, *"algoritmit eivät välttämättä miellä sitä asiaa sillä tavalla kuin meidän ihmisyhteisö sen ajattelee"*	– Hyvä kehysjärjestelmä, pidetään yötä päivää huolta – Riskienhallinnan tulisi kasvaa suhteessa riskin määrään – Vahinkojen minimointi – Odottamattomiin tapahtumiin varautuminen ja auditointi: odottamattomat ja kohtuuttomat vahingot täytyy olla minimoitu – Riskienhallinnan jatkuvuus – Ymmärretään kokonaisuutena: prosessit, työtehtävät, kontrollit ja teknologiat minimoi riskejä – Pitää varmistaa, millaiseen lopputulokseen algoritmi ei ainakaan tietyssä tilanteessa saa päätyä, vaikka jotain puuttuisi opetusdatasta – Korkeiden vaatimusten tulee täyttyä, jos kyseessä korkean riskin sovellus, esimerkkinä syöpädiagnoosi – Standardointi – Ohjelmistoversioiden malli – Vuosikellon mukainen tietosuojan ja tietoturvan tarkistus ja vaikutusten arviointi – Domain-osaaminen – Pitäisi ymmärtää kyllä-ei-päätösten harmaa alue, että jos ennalta päätetään jokin lopputulema tai tavoite, niin siihen tavoitteeseen pääsemisen keinot tulee ymmärtää ja *"tässä tulee tietysti tämä järjestelmävastuun ja vastuuvelvollisuuden mekanismit siihen"* – Harmaalla alueella pitäisi aina olla mieluummin se vaihtoehto, josta on vähemmän haittaa kohteelle – Kaikkien tahojen tulisi tietää ja ymmärtää, kuinka paljon tekoälyjärjestelmään voi luottaa
Kulttuuri, osallistamisen puute (H1)	– Asiakkaiden ääni kehitysprosessissa: *"lähtökohtaisesti näitä sovelluksia kehitetään hyvin pienissä, hyvin homogeenisissä tiimeissä"* – *"meidän pitäisi ymmärtää, että me vaikutamme ihmisiin, ja mitä isommin me vaikutetaan, mitä vaikuttavampi se sovellus on, niin sitä tarkempia meidän pitää olla siinä, että me aidosti ymmärretään, kuullaan niiden ihmisten kontribuutio, -- kenen elämää tässä ollaan muuttamassa"*	– Design thinkingiin liittyvät menetelmät, mallia digitaalisten palveluiden kehityksestä – palvelumuotoilu – Erilaisten ihmisten näkemyksiä ja palautetta kehitystiimiin – Eri tahojen yhteistyö – Asiakas osaksi kehittämistä -> arjen vastuullisuutta demokratisoidaan tekoälyä, vähennetään pelkoa
Tekoälyn kuluttama energia (H2)	– Monipuolisten tekoälymallien kuluttama laskentateho ja energiamäärä – Rajoittaa myös mahdollisuuksia käyttää tekoälyä laboratorioiden ulkopuolella, koska ei ole taloudellisesti eikä ympäristön kannalta kestävää – Energian tarve kasvaa voimakkaasti ja jatkuvasti (data ja tekoäly) – Resurssien tarve voi kasvaa nopeammin kuin datakeskusten energiatehokkuus	– Ihmisten tietoisuuden lisääminen – Datankäytön tarveharkinta – Toisaalta energiatehokkuus on jo parantunut ja sähköä tuotetaan ympäristöystävällisemmin ja – datakeskukset sijaitsevat pohjoisessa, jossa jäähdyttäminen on halvempaa ja datakeskusten lämpöä otetaan jo talteen

Liite 5a. Periaatteiden vertailu

Anttinen & Lohilahti (2019, 41) olivat ansiokkaasti jo verranneet tutkimiensa EK:n alaisten yritysten eettisiä periaatteita AI HLEG:n eettisiin periaatteisiin. Jatkoin siitä vertaamalla niitä julkisen sektorin AuroraAI-ohjelman eettisiin periaatteisiin, Fjeld:n ym. (2020, 4–5) löytämään periaatteiden normatiiviseen ytimeen, työntilaajan code of conductiin, yritysvastuuohjelmaan ja digitaalisen sisällön suunnitteluperiaatteisiin sekä yhteenvetoon suosituksista, jotka on jätetty Euroopan komissiolle liittyen tekoälyn etiikan oikeudelliseen kehitykseen. Tietoperustan mukaisesti periaatteiden jaottelu on tulkinnanvaraista, mutta huomionarvoista on silti ylätason periaatteellinen yhdenmukaisuus (liite 5a).

Taulukko 8. Sekundäärisen tutkimusaineiston vertailu

Taho / Periaate	Monimuotoisuus, syrjimättömyys, oikeudenmukaisuus, yhdenvertaisuus, tasa-arvo ja puolueettomuus	Yksityisyydensuoja, datanhallinta, tietosuoja, tiedollinen itsemääräämisoikeus ja datan laatu	Tekninen luotettavuus, turvallisuus, jatkuva arviointi, riskien arviointi, vahinkojen välttäminen ja tietoturva	Läpinäkyvyys, avoimuus, selitettävyys, ymmärrettävyys ja avoin viestintä	Ammatillinen vastuu, vastuuvelvollisuus, muutoksenhakuoikeus ja omistajuus	Ihmisen toimijuus, ihmisjohtajuus ja ihmisen suorittama valvonta ja kontrolli	Yhteiskunnallinen, sosiaalinen ja ekologinen hyvinvointi, yhteiskuntavastuu ja vastuullisuus	Ihmiskeskeisyys, ihmisarvoja kunnioittava tarkoitus, itsemääräämisoikeus, ihmisoikeudet ja inhimillisten arvojen edistäminen	Lähde
AI HLEG:n vaatimukset (& periaatteet)	x	x	x	x	x	x	x	x	AI HLEG 2019
EK:n alaisten yritysten periaatteet	x	x	x	x	x	x	x	x	Anttinen & Lohilahti 2019
AuroraAI	x	x	x	x	x			x	AuroraAI:n Etiikka-verkosto, Haataja & Latvanen 2019
Normatiivinen ydin	x	x	x	x	x	x	x	x	Fjeld ym. 2020
Yhteenveto suosituksista Euroopan komissiolle eettisten näkökulmien kehityksestä	x	x	x	x	x	x	x	x	Oikeudellisten asioiden valiokunta 2020
Työntilaajan code of conduct									
Työntilaajan yritysvastuuohjelma									
Työntilaajan digitaalisen sisällön suunnitteluperiaatteet									