Tampereen yliopisto

Kasper Järvinen

TEKOÄLYAVUSTEISEN PÄÄTÖKSEN-TEON ETIIKAN TARKASTELU

Tekniikan ja luonnontieteiden tiedekunta
Kandidaatintyö
Huhtikuu 2020

TIIVISTELMÄ

Kasper Järvinen: Tekoälyavusteisen päätöksenteon etiikan tarkastelu
Ethical approaches to artificial intelligence assisted decision making
Kandidaatintyö
Tampereen yliopisto
Tietojohtaminen
Helmikuu 2020

Tämän kandidaatintyön tarkoituksena oli tutkia, että millä keinoilla tekoälyavusteisen päätöksenteon etiikkaa voidaan tarkastella. Tekoälyn käyttö on yleistynyt ja se yleistyy edelleen ihmisten arkipäiväisissäkin sovelluksissa. Päätöksentekoa ja siihen liittyvää tiedon hankintaa tehdään tekoälysovelluksia hyödyntäen. Tämän vuoksi tekoälyn päätöksenteon tulee olla eettisesti kestävää. Tekoälyteknologioita ja käyttötarkoituksia on useita eikä niiden päätöksenteon etiikkaan ole yksiselitteistä lähestymistapaa.

Tutkimus suoritettiin kirjallisuuskatsauksena. Tutkimusaineistona toimi Tampereen yliopiston kirjaston tietokannoista saatavat tieteelliset artikkelit, kirjat ja konferenssijulkaisut. Tutkimuksen alussa tutkittiin tekoälyn käsitettä ja tietoperustaista päätöksentekoa. Sen jälkeen työssä tutkittiin sitä, voiko tekoälysovellusta pitää moraalisena toimijana tai toimijana ylipäätään. Tämän jälkeen tutkittiin tekoälysovelluksia kahden eri normatiivisen etiikan teorian kautta: utilitarismin ja deontologian.

Tutkimuksessa selvisi, että tekoälysovelluksien oppiminen perustuu kolmeen erilaiseen koneoppimismenetelmään: ohjaamattomaan oppimiseen, ohjattuun oppimiseen ja vahvistusoppimiseen. Lisäksi todettiin, että ihmisten ja organisaatioiden päätöksentekoprosessi on monelta osin hyvin samanlainen tekoälysovelluksen päätöksentekoprosessin kanssa. Tämän jälkeen tutkimuksessa todettiin, että tekoälysovellusta ei voida pitää moraalisena toimijana eikä varsinaisena toimijana. Sen sijaan toimijuutta voi luovuttaa tekoälysovellukselle sille osoitettujen tehtävien myötä. Tutkimuksessa kävi ilmi, että tekoälysovelluksen teknisistä kyvykkyyksistä, kuten toimien seurausten arvioinnista ja oppimiskyvystä riippuen sovelluksen päätöksenteon etiikkaa voidaan tarkastella eri normatiivisten etiikoiden keinoin. Kyvykkään, toimiensa seurauksia arvioivan ja niistä oppivan sovelluksen tarkasteluun sopii utilitarismi. Rajatumman, tiettyjä ohjeita seuraamaan kykenevän sovelluksen päätöksenteon eettiseen tarkasteluun soveltuu paremmin deontologia. Tarkasteluun käytetyn etiikan teorian valinta riippuu käytetystä tekoälyteknologiasta, sovelluksen käyttöympäristöstä, sen kehittäjistä ja sen hyödyntäjistä.

Avainsanat: tekoäly, tekoälyn etiikka, eettinen päätöksenteko, koneoppiminen

ALKUSANAT

Tämä tutkimus tehtiin kandidaatintyönä Tampereen yliopiston tietojohtamisen koulutusohjelmaan kevätlukukaudella 2020. Tutkimuksen aihe valikoitui oman mielenkiinnon perusteella. Aihe osoittautui mielenkiintoiseksi, mutta sen rajaaminen kandidaatintyöhön sopivaksi kokonaisuudeksi ja yhdistäminen opiskelemaani alaan aiheutti päänvaivaa. Tutkimusprosessi ja oikeastaan koko kulunut kevät oli raskas ja ajoin tukalakin, mutta lopulta antoisa ja sisällyksekäs. Suuret kiitokset työn opponoijille hirveän laadukkaasta opponoinnista seminaaritilaisuuksissa.

Tampereella, 28.4.2020

Kasper Järvinen

SISÄLLYSLUETTELO

1. JOHDANTO

1.1 Taustaa

Tekoälyn etiikka tuo monelle mieleen itseajavan auton dilemman: jos törmäys on välttämätön, kenen henki pyritään säästämään? Yleisesti ajatellaan, että tämänkaltaisten ongelmien ratkaisu olisi tekoälyä hyödyntävän auton päätettävissä. Todellisuudessa itseajavan auton tekoälyominaisuudet voidaan ohjelmoida tekemään päätöksiä auton käyttäjän, sen valmistajan tai esimerkiksi paikallisen lainsäädännön määrittelemien arvojen ja sääntöjen mukaisesti. Tällöin kyse ei olekaan enää tekoälyä hyödyntävän auton tekemästä päätöksestä, vaan eettisen päätöksen tekee lopulta ihminen. Tämä johtaa helposti keskusteluun oikeasta ja väärästä, jossa ei ole enää suoraan kyse tekoälyn etiikasta.

Tekoälyteknologioita on kuitenkin erilaisia, eikä kaikkien tekoälysovellusten tekemiä päätöksiä ole mielekästä tarkastella samoihin etiikan teorioihin tukeutuen. Usein tekoälysovelluksiin liittyy koneoppimista, jonka avulla sovelluksen tekoälyominaisuudet oppivat tekemistään päätöksistä. Koneoppimismenetelmiä on kehitetty tutkimalla ja matkimalla ihmisen oppimisprosessia, mikä on johtanut esimerkiksi neuroverkkoteknologioiden kehitykseen. Koneoppimismenetelmiä on useita, joista osa liittyy esimerkiksi *ohjaamattomaan oppimiseen*, *ohjattuun oppimiseen* ja *vahvistusoppimiseen.* Näitä teknologioita hyödyntäviä tekoälysovelluksia käytetään eri tarkoituksiin, jolloin tekoälyn tekemien päätöksien eettiset kysymykset muodostuvat käyttötilanteesta, käytön kohteesta ja käytetystä teknologiasta. Tämä herättää kysymyksen, että onko olemassa yksiselitteistä lähestymistapaa tekoälyn käytön eettisiin ongelmiin, vai onko tarkastelukeinoja useita.

Tekoälyä hyödynnetään laajalti päätöksenteossa. Käytännöllisiä esimerkkejä tekoälyavusteisesta päätöksenteosta ovat terveydenhoitoalalla käytetyt tekoälyavusteiset syöpätutkimukset, finanssialalla käytetty tekoälyavusteinen luotonmyöntäminen sekä verkkokaupoissa käytetty dynaaminen hinnoittelu. Tällaisten tekoälysovelluksien tekemät päätökset perustuvat dataan, eli sitä voi verrata ihmisen tekemään tietoperustaiseen päätöksentekoon. Toisaalta tekoälyä hyödynnetään päätöksenteossa usein ennen päätöksentekotilannetta, jolloin tekoälyn avulla saatu tulos tukee ihmisen päätöksentekoa osana tietoperustaista päätöksentekoprosessia.

1.2 Tutkimusongelma ja rajaus

Tässä työssä tutkitaan tekoälyavusteisen päätöksenteon etiikan lähestymistapoja. Tutkimuksen tavoitteena on löytää etiikan teorioiden pohjalta työkaluja tunnistaa eri teknologioihin ja käyttötarkoituksiin liittyviä eettisiä elementtejä ja keinoja näiden tarkasteluun. Tämän kandidaatintyön päätutkimuskysymyksenä on:

- Miten lähestyä tekoälyavusteisen päätöksenteon etiikkaa?

Päätutkimuskysymyksen tukena on alatutkimuskysymyksiä, joita tutkimalla saadaan tukea päätutkimusongelmaan vastaamiseksi. Alatutkimuskysymyksiä ovat:

- Mitä on tekoäly ja millainen on tekoälysovelluksen päätöksentekoprosessi?
- Voidaanko tekoälysovellusta pitää moraalisena toimijana?
- Millä etiikan teorioilla päätöksenteon ongelmia voidaan tarkastella?
- Miten käytetty teknologia ja käyttöympäristö vaikuttavat etiikan tarkasteluun?

Tutkimuksen kohteena on siis eri tekoälymenetelmien eri käyttötarkoituksissa hyödyntämisen etiikka. Tekoälyn etiikkaa on tutkittu paljon, mutta sen asiantuntemus on lähinnä sekä tekniikan että filosofian asiantuntijoiden käsissä. Filosofien tekemä tutkimus tekoälyn etiikasta keskittyy usein tekoälyn sovellusalueisiin tekoälynkentän komponenttien, kuten esimerkiksi käytetyn teknologian ja käyttöympäristön, sijaan (Ollila, 2019). Itse teknistä toteutusta ja sen suhdetta sovellusalueeseen tutkivat taas yleensä tekniset asiantuntijat. Tässä tutkimuksessa pyritään liittämään sekä teknistä että filosofista näkökulmaa tekoälyavusteisen päätöksenteon etiikan tarkasteluun.

1.3 Tutkimuksen rakenne

Tutkimus koostuu johdannon lisäksi viidestä luvusta. Johdannon jälkeen esitellään tutkimusmenetelmä ja tutkimusaineistoa. Tämän jälkeen luvussa 3 käsitellään tekoälyä työkaluna ja tietoperustaista päätöksentekoa. Seuraavaksi luvussa 4 pohditaan, että voidaanko tekoälyä hyödyntävää sovellusta pitää eettisen pohdinnan kannalta moraalisena toimijana. Luvussa käsitellään moraalisen toimijuuden lisäksi tekoälysovelluksen toimijuutta yleisemmin.

Toimijuuden tarkastelun jälkeen luvussa viisi syvennytään päätöksenteon etiikkaan ja pohditaan etiikan teorioita eri tekoälyteknologioiden ja käyttötilanteiden kannalta. Tutkimuksen lopussa tuloksista esitetään yhteenveto, jossa tutkimustulokset esitetään ja arvioidaan. Yhteenvedon yhteydessä pohditaan myös tutkimuksen merkitystä ja kartoitetaan jatkotutkimuksen tarvetta.

2. TUTKIMUSMENETELMÄ JA -AINEISTO

Kandidaatintyö toteutetaan systemaattisena kirjallisuuskatsauksena. Tutkimus toteutetaan Finkin (2005, Salminen 2011 mukaan) määrittelemää tutkimuskirjallisuuteen perustuvan kirjallisuuskatsauksen mallia mukaillen. Mallissa tutkitaan ja seulotaan aihepiirin aiempaa tutkimuskirjallisuutta ja pyritään asettamaan runsas tutkimusmateriaali tiiviissä muodossa oman aiheen ja tieteenalan kontekstiin (Salminen, 2011). Finkin systemaattisessa tutkimusprosessimallissa on seitsemän vaihetta, joista ensimmäinen on *tutkimuskysymyksen asettaminen*. Tässä vaiheessa tutkimusongelma rajataan päätutkimuskysymyksen ja alatutkimuskysymysten avulla. Tutkimusongelma valikoituu analysoimalla olemassa olevaa tutkimuskirjallisuutta aiheen ympäriltä. Tämän työn tutkimuskysymykset on esitetty luvussa 1.

Tutkimuskysymyksen asettamista seuraa Finkin mallissa (2005, Salminen 2011 mukaan) niiden tietokantojen ja sivustojen valinta, joista tutkimuskirjallisuutta etsitään. Tässä kirjallisuuskatsauksessa käytetään Tampereen yliopiston tarjoamaa aineistoa, joka kattaa kirjaston kirjojen lisäksi paljon tietokantoja, joista voi hakea tutkimuskirjallisuutta kuten kirjoja, artikkeleita, tutkimuksia ja muita tieteellisiä julkaisuja. Tässä tutkimuksessa hyödynnetään Tampereen yliopiston kirjaston Andor-tietokantahakupalvelua, joka hakee aineistoja useammasta eri tietokannasta. Lisäksi tutkimuksessa hyödynnetään suoraan muutamaa tiettyä tietokantaa, jotka ovat Web of Science, Scopus ja ProQuest Central. Joitain tutkimuksessa käytettyjä artikkeleita löytyy myös Google Scholarista.

Finkin mallin kolmantena vaiheena on hakutermien ja -lausekkeiden valinta (2005, Salminen 2011 mukaan). Hakulausekkeet muodostetaan yhdistämällä keskeisiä termejä Boolen operaattoreilla ”AND” ja ”OR”. Keskeisillä termeillä tarkoitetaan niitä käsitteitä, jotka koskevat ja kuvaavat tutkittavaa aihetta. Taulukossa 1 on esitetty rajoittamattomia hakutuloksia eri hakulausekkeilla tietokantakohtaisesti. Tuloksia on paljon, mikä selittyy sillä, ettei hakutuloksia ole vielä rajattu mitenkään.

Taulukko 1: Tietokantojen rajoittamattomat hakutulokset hakulausekkeittain

Hakulauseke	Scopus	Web of Science	ProQuest	Andor
”artificial intelligence” AND ”moral agent”	46	6	244	304
“artificial intelligence” AND ethics	1 115	409	33 237	21 619
“artificial intelligence” AND “decision making”	19 647	2 575	78 088	85 711
“artificial morality”	18	9	76	110
"machine ethics"	247	138	445	672
"knowledge based decision making" AND “artificial intelligence”	17	2	57	75

Kolmannessa Finkin mallin (2005, Salminen 2011 mukaan) vaiheessa saatuja hakutuloksia rajataan mallin neljännessä vaiheessa. Rajaus suoritetaan tässä työssä asettamalla tuloksille hakukriteereitä kielen, julkaisuvuoden ja artikkelityypin perusteella. Haku rajataan koskemaan vain englannin kielisiä tieteellisiä tutkimusartikkeleita, kirjoja ja konferenssijulkaisuja. Lisäksi hakuja rajataan koskemaan vuosien 2000-2020 julkaisuja, sillä käsitys tekoälystä ja sen käyttösovelluksista on kehittynyt paljon 1900-luvun puolivälistä asti. Taulukossa 2 esitetään taulukon 1 hakutulokset mainituilla hakukriteereillä.

Taulukko 2: Tietokantojen rajoitetut hakutulokset hakulausekkeittain

Hakulauseke	Scopus	Web of Science	ProQuest	Andor
”artificial intelligence” AND ”moral agent”	38	6	218	283
“artificial intelligence” AND ethics	772	349	11 840	13 222
“artificial intelligence” AND “decision making”	15 507	2074	30 803	51 842
“artificial morality”	12	7	40	31
"machine ethics"	149	129	383	514
"knowledge based decision making" " AND “artificial intelligence”	17	2	57	75

Seuraavaksi Finkin (2005, Salminen 2011 mukaan) mallin prosessissa rajataan hakutuloksia asiayhteyksien ja avainsanojen perusteella. Tämän tutkimuksen aineiston rajaaminen asiayhteyden ja avainsanojen perusteella osoittautui kuitenkin liian rajuksi, sillä vaikka asiayhteyksiä ja avainsanoja valitsi suuren määrän, rajautui hakutuloksista pois paljon tutkimuksen kannalta relevanttia aineistoa. Näin ollen tutkimusaineisto valitaan taulukon 2 mukaisista tuloksista otsikon ja sisältökuvauksen perusteella niin, että materiaalia tutkitaan kriittisesti ja arvioidaan sen relevanttiutta suhteutettuna tämän tutkimuksen aiheeseen ja tavoitteeseen.

Tutkimusaineiston keräämiseksi suoritettiin myös muutama suomenkielinen haku taulukon 1 ja 2 mukaisilla, suomeksi käännetyillä hakulausekkeilla. Materiaalia on suomen kielellä huomattavasti vähemmän, minkä vuoksi aineistoa haettiin pääasiassa englanniksi. Varsinaisen tutkimusaineiston lisäksi tätä tutkimusta varten suoritetaan muutamia asiasanahakuja liittyen tekoälyyn ja etiikan teorioihin ja niiden määrittelyihin. Edellä mainitut aikarajaukset eivät koske asiasanahakuja, sillä vanhoista ja vakiintuneista käsitteistä kertovan kirjallisuuden katsotaan olevan edelleen koherenttia.

Tutkimusmateriaalin rajauksen jälkeen Finkin (2005, Salminen 2011 mukaan) mallissa suoritetaan kirjallisuuskatsaus. Tämä tarkoittaa tutkimusprosessia, jossa tutkimusaineistoon tutustutaan laajalti ja verrataan sitä muuhun tutkimusaineistoon ja oman tutkimuk-

sen aihepiiriin. Kun tutkimus on suoritettu, seuraa tätä Finkin (2005, Salminen 2011 mukaan) mallissa tulosten syntetisointi, jossa saatuja tuloksia kootaan yhteen ja analysoidaan.

2.1 Tutkimusaineiston esittely

Löydetty tutkimusaineisto koostuu pääasiassa kirjoista, kirjojen luvuista, artikkeleista ja konferenssijulkaisuista. Tutkimuksessa pyritään hyödyntämään pääasiassa tieteellisesti vertaisarvioitua materiaalia, mutta myös esimerkiksi tietokirjallisuutta ja muita artikkeleita kriittisesti tarkastellen. Tutkimusaineistoa haetaan vaiheittain tässä työssä käsiteltävän teeman mukaan, kuten taulukoissa 1 ja 2 esitetään. Taulukossa 3 on esitelty muutama työssä relevantiksi osoittautunut lähdemateriaali otsikkoineen ja ydinsisältöineen.

Taulukko 3: Esimerkkejä tutkimusmateriaalista

Aihe	Kirjoittaja	Otsikko	Ydinsisältö
Koneoppiminen	Louridas, P. & Ebert, C.	Machine Learning	Koneoppimisteknologiat ja niiden toimintaperiaatteet
Tekoälysovelluksen päätöksenteko	Alaieri, F. & Vellino, A.	Ethical decision making in Robots: Autonomy, Trust and Responsibility	Tekoälysovelluksien päätöksenteon etiikan tarkastelu
Tekoälyn toimijuus	Gunkel, D.J. & Bryson, J.	Introduction to the Special Issue on Machine Morality: The Machine as Moral Agent and Patient	Tekoälyä hyödyntävän koneen pitäminen moraalisubjektina
Tekoälyavusteisen päätöksenteon etiikka	Bench-Capon, T.J.M.	Ethical approaches and autonomous systems	Eri näkökulmia autonomisten systeemien etiikan tarkasteluun

Tutkimusmateriaalia löytyi paljon, joten tutkimuksen kannalta relevantin aineiston löytäminen oli työlästä. Usein rajaukseen ei riittänyt aineiston otsikko tai tiivistelmä, vaan aineistoon tuli tutustua tarkemmin. Lopulta löydetty ja käytetty aineisto osoittautui kuitenkin melko ajankohtaiseksi ja hyvin aiheeseen sopivaksi.

3. TEKOÄLY JA SEN TEKEMÄ PÄÄTÖKSEN-TEKO

Tässä luvussa esitellään tekoälyä työkaluna ja elementtinä, jona tekoälyä ja sen sovelluksia käsitellään myöhempänä tässä tutkimuksessa. Luvussa selvitetään, mitä tekoälyllä tarkoitetaan, mitä elementtejä siihen liittyy ja mitä teknologisia menetelmiä tekoälysovellukset käyttävät. Lisäksi luvussa tarkastellaan, millainen on tekoälysovelluksen päätöksentekoprosessi.

3.1 Tekoälyn määritelmiä

Tekoäly on käsitteenä abstrakti, eikä sille ole olemassa yhtä yleisesti tunnustettua määritelmää. Russel & Norvig kuvailevat kirjassaan *Artificial Intelligence: A Modern Approach* (2003) tekoälyä työkaluna, jonka avulla koneet, laitteet, sovellukset, järjestelmät ja palvelut kykenevät toimimaan tehtävän ja tilanteen mukaisesti järkevällä tavalla. Tämä määritelmä edellyttää yleisesti tunnistettuja tekoälyn ominaisuuksia, joista esimerkiksi Helsingin yliopiston ja teknologiayritys Reaktorin kaikille ilmainen verkkokurssi *Elements of AI* esittelee kaksi: *autonomisuus* ja *adaptiivisuus*. Jotta sovellus, kuten laite, kone tai järjestelmä, pystyisi tekemään järkeviä päätöksiä itse, tulee sen suoriutua tehtävästään ilman käyttäjän ohjausta sekä parantaa suoritustaan saamansa palautteen perusteella. Sovelluksen voidaan siis ajatella hyödyntävän tekoälyä, jos se toimii joiltain osin autonomisesti ja adaptiivisesti.

Yhden määritelmän mukaan tekoälyllä tähdätään keinotekoisten eläinten rakentamiseen (Standard Encyclopedia of Philosophy 2018, Ollila 2019 mukaan). Vaikka joitain tekoälysovelluksia pyritäänkin saamaan matkimaan inhimillistä käyttäytymistä matemaattisin menetelmin (Cervantes et al., 2016), koko tekoälyn tutkimus ja kehitys ei keskity luomaan mahdollisimman inhimillistä tekoälysovellusta. Tässä tutkimuksessa tekoälyä ei käsitellä ihmisyyttä matkivana koneena, vaan työkaluna.

3.2 Tekoälyn elementtejä

3.2.1 Koneoppiminen

Koneoppiminen tarkoittaa tekniikkaa, jossa tietokonetta opetetaan suorittamaan jokin sille osoitettu tehtävä paremmin ja tehokkaammin kuin ihminen. Tämä tapahtuu syöttä-

mällä koneoppimisalgoritmeille suuri määrä opetusdataa opetettavasta tehtävästä. Tämän jälkeen tietokone suorittaa samaa tehtävää uudella, opetusdatasta irrallisella syötteellä. Tyypillisiä koneoppimisen sovellusalueita ovat muun muassa kuva-analyysit, kuvantunnistus sekä kuviontunnistus. (Louridas & Ebert, 2016) Kuvantunnistusta hyödynnetään esimerkiksi robottiautoissa ja kuviontunnistusta käytetään muun muassa automatisoiduissa lainanhaku- ja myöntämisprosesseissa lainanhakijan maksukyvyn arviointiin.

Pääasiallisia koneoppimismenetelmiä on tunnistettu kolme (Wojtusiak, 2012; Louridas & Ebert, 2016):

- Ohjattu oppiminen (engl. supervised learning)
- Ohjaamaton oppiminen (engl.unsupervised learning)
- Vahvistusoppiminen (engl.reinforcement learning)

Ohjatun oppimisen mallissa opetettavalle tietokoneohjelmalle syötetään dataa, jonka lisäksi ohjelmalle kerrotaan metatietoja datasta (Wojtusiak, 2012). Tämä voi tarkoittaa esimerkiksi ohjelmalle syötettävää ääniraitaa pianosta, ja tietoa, että kyseinen ääniraita on pianon ääni. Kun sama toistetaan monella eri pianon ääniraidalla, ohjelma tunnistaa analysoimistaan ääniraidoista samankaltaisuuksia ja liittää ne tietoon siitä, että raita on pianosoittimesta. Jos ohjelmalle syöttää tämän jälkeen ääniraidan esimerkiksi kitarasta, osaa ohjelma todennäköisesti kertoa, että kyseinen ääniraita ei ole pianosoittimesta. Ohjattua oppimista voi siis verrata tilanteeseen, jossa oppilaalle annetaan joukko samankaltaisia tehtäviä ja ratkaisut näihin tehtäviin ja pyydetään tätä keksimään keino ratkaista samankaltaisia tehtäviä tulevaisuudessa (Louridas & Ebert, 2016).

Ohjaamaton oppiminen tarkoittaa menetelmää, jossa opetettavalle ohjelmalle syötetään opetusdataa, mutta ei ”oikeita vastauksia”. Tällöin tietokoneohjelman tulee analysoida dataa ja tunnistaa sen perusteella kullekin dataobjektille tyypillisiä piirteitä. (Wojtusiak, 2012) Tätä kutsutaan *klusteroinniksi*, joka eroaa ohjatussa oppimisessa tehtävästä luokittelusta siten, ettei ohjelmalle kerrota valmiita luokkia, joihin data tulisi lokeroida. (Louridas & Ebert, 2016) Jos ohjaamatonta oppimista hyödyntävälle tietokoneohjelmalle annettaisiin syötteenä paljon vain pianon ääntä sisältäviä ääniraitoja, *klusteroisi* ohjelma todennäköisesti ne omaksi, luokittelemattomaksi joukokseen. Tällöin, jos ohjaamattoman oppimisen ohjelmalle syöttäisi kitaran ääntä sisältävän ääniraidan, tunnistaisi ohjelma todennäköisesti eron kitaran ääniraidassa ja pianon ääniraidassa, vaikka ohjelma ei sinänsä tietäisikään, minkä soittimen ääniraita sille on syötetty.

Vahvistusoppiminen on ohjattua oppimista ja ohjaamatonta oppimista hieman monimutkaisempi koneoppimismenetelmä. Siinä tietokoneohjelman tavoite on löytää yhteys suoritettujen operaatioiden ja maksimaalisen tuotoksen välillä (Wojtusiak, 2012). Tällöin tietokoneohjelmaan määritetään ohjelmoijan toimesta tavoite, joka voi olla esimerkiksi pistesaldo jossain pelissä. Kun tavoitteena on maksimaalinen pistesaldo, suorittaa kone operaatioita ja tarkistaa, johtaako ne maksimaaliseen pistesaldoon, ja muokkaa toimintaansa sen mukaisesti.

Lisäksi on olemassa eri koneoppimismalleja yhdisteleviä menetelmiä. Hady ja Schwenker esittelevät kirjansa *Handbook on Neural Information Processing* kappaleessa *Semi-supervised Learning* (2013) esimerkin puoliohjatusta oppimisesta (semi-supervised learning). Puoliohjatussa oppimisessa koneelle syötetään ihmisen luokitteleman datan lisäksi luokittelematonta dataa. Tällä mallilla pyritään säästämään suuri määrä ihmisen tekemää luokittelutyötä. Eri koneoppimismenetelmien yhdistäminen aiheuttaa yleensä lisähaasteita esimerkiksi koneoppimismallien tarkkuuksiin, jolloin yhdisteleviä koneoppimisen malleja tulee käyttää harkiten ja vain malliin sopivassa sovelluksessa. (Hady & Schwenker, 2013)

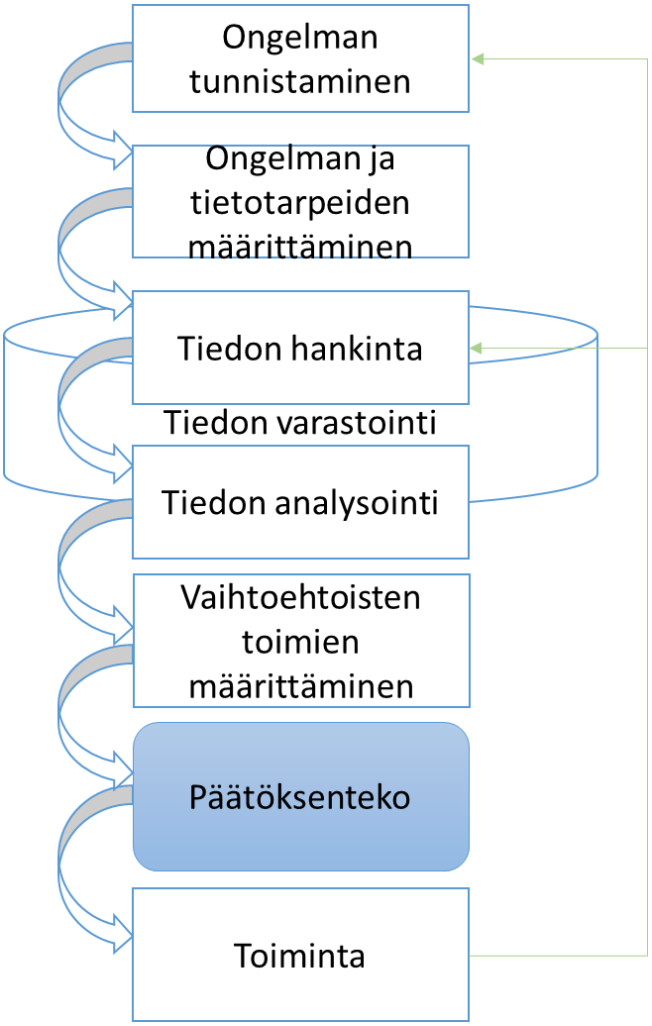
3.2.2 Neuroverkot ja syväoppiminen

Neuroverkot, tai tässä yhteydessä keinotekoiset neuroverkot, ovat biologisten aivojen inspiroimia kompleksisia laskennallisia malleja. Nämä mallit koostuvat useista tietoa prosessoivista neuroneista ja niiden välisistä painotetuista suhteista. Yksittäinen neuroni saa syötteekseen tietyn tyyppistä tietoa ja käsittelee sitä omilla algoritmeillaan, minkä jälkeen neuronin tuloste toimii seuraavan neuronin syötteenä. Nämä suhteet muodostavat verkkomaisen rakenteen, johon on myös kytketty algoritmeja. (Shanmuganathan, 2016) Neuroverkot ovat oppivia ja adaptiivisia (Shanmuganathan, 2016) ja niitä käytetään koneoppimisteknologioissa.

Neuroverkot voivat muodostua monesta eri kerroksesta, mikä johtaa koneoppimisalgoritmien kykyyn ratkaista entistä kompleksisempia ongelmia. Kun neuroverkko muodostuu tarpeeksi monesta oppivasta kerroksesta, voidaan sen katsoa mahdollistavan syväoppimisen (LeCun et al., 2015). Syväoppiminen tarkoittaa datan eri ulottuvuuksien oppimista (LeCun et al., 2015). Jos data on esimerkiksi ääniraita, sen ulottuvuuksilla voidaan tarkoittaa esimerkiksi ääniraidan pituutta, äänen taajuutta, äänenvoimakkuutta ja äänen sävyä.

3.3 Tietoperustainen päätöksenteko

Tietoperustainen päätöksenteko perustuu sananmukaisesti tietoon. Tietoperustainen päätöksenteko kytkeytyy vahvasti tietojohtamiseen, joka tutkii tieteenalana tiedon, osaamisen ja informaation kokonaisvaltaista hyödyntämistä organisaatioissa (Yim et al., 2004). Tietoperustaisessa päätöksenteossa sovelletaan tietojohtamisen menetelmiä muun muassa tietotarpeiden tunnistamiseen, tiedon prosessointiin ja lopulta tiedon hyödyntämiseen (Choo, 2001). Organisaatioissa tietoperustaisen päätöksenteon tukena käytetään yleensä jonkinlaista tietojärjestelmää, jonka avulla voidaan muun muassa hyödyntää kerättyä dataa vaihtoehtoisten päätösten seurausten arviointiin (Alyoubi, 2015).



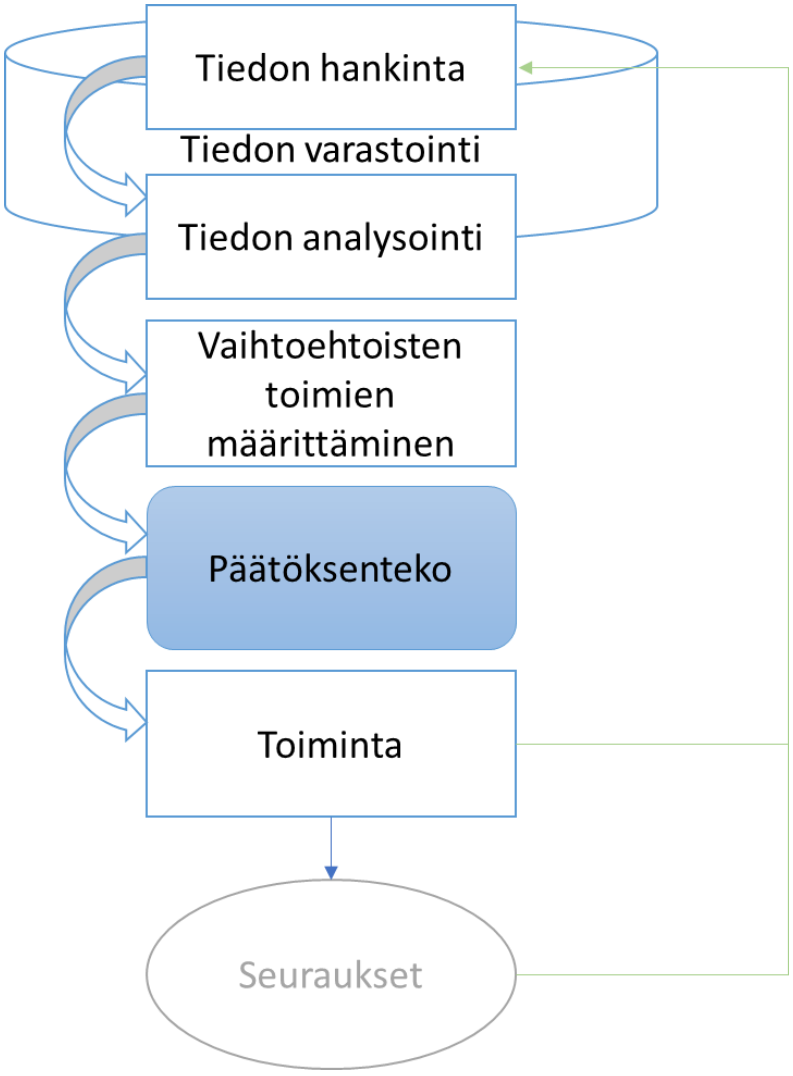
Kuva 1: Tietojärjestelmän tukema päätöksentekoprosessi (muokattu lähteistä Choo 2001 ja Courtney 2001)

Kuvassa 1 esitetään hahmotelma tietojärjestelmän tukemasta tietoperustaisesta päätöksentekoprosessista. Prosessi on rakennettu oppivaksi, mikä näkyy päätöksenteon ja

suoritettujen toimintojen vaikutusten seurantana. Kuvassa 1 näitä niin kutsuttuja feedback-loopeja mallinnetaan vihreillä nuolilla.

3.4 Tekoälysovelluksen päätöksentekoprosessi

Jotta päätöksentekoprosessia voidaan pitää tietoperustaisena, tulee päätöksien perustua tiedon analysointiin. Tekoälysovelluksien päätöksenteko perustuu opittuun tietoon ja sen analysointiin, joten sitä voidaan pitää tietoperustaisena. Kuvassa 2 on esitetty Alaierin & Vellinon (2016) esittelemää robottiparadigmaa – eli yleisesti tunnustettua teoriaa robottien päätöksenteosta – mukaileva tekoälysovelluksen päätöksentekoprosessi.



Kuva 2: Tekoälysovelluksen päätöksentekoprosessi (muokattu lähteestä Alaieri & Vellino, 2016)

Kuvan 2 prosessissa toiminnasta ja siitä johtuvista seurauksista kerätään tietoa, eli prosessi on oppiva. Prosessin feedback-loopien tietovirta on kuvattu kuvassa 2 vihreillä

nuolilla. Tiedon hankinta voi tekoälysovelluksesta riippuen tapahtua dynaamisesti ympäristöstä seurauksien kautta (vahvistusoppiminen) tai se voi perustua kehitysvaiheessa opittuun tietoon (ohjattu oppiminen).

Kuten kuvista 1 ja 2 huomaa, ovat organisaatioiden ja tekoälysovelluksien päätöksentekoprosessit hyvin pitkälti samanlaisia. Molemmat perustuvat tietoon, molemmissa kerätään uutta tietoa ja molemmissa on feedback-loopeja. Olennainen ero liittyy tietoperustaisen päätöksentekoprosessin ensimmäiseen vaiheeseen, eli ongelman tunnistamiseen. Tekoälysovelluksen päätöksentekoprosessi ei pyri tunnistamaan ongelmia, sillä tekoälysovelluksia käytetään yleensä johonkin tietyn jo tunnistetun ongelman tai tehtävän ratkaisemisen apuvälineenä.

Allen et al. (2005) esittävät kaksi eri keinoa rakentaa tekoälysovelluksen päätöksentekoprosessista eettisen: top-down-metodi ja bottom-top-metodi. Top-down-metodissa tekoälysovelluksen kehittäjät ohjelmoivat tekoälysovellukselle päätöksentekoalgoritmeja, joiden perusteella tekoälysovellus tuottaa ennustettavia päätöksiä (Alaieri & Vellino, 2016). Päätöksenteko perustuu tällöin siis ennalta määritettyjen sääntöjen noudattamiseen (Allen et al., 2005).

Bottom-top-mallissa tekoälysovelluksen annetaan oppia päätöksenteon moraaliset säännöt käytännön kokemuksesta (Allen et al., 2005). Tässä metodissa tekoälysovellukseen ei siis asenneta tiettyjä moraalisia sääntöjä noudattavia päätöksentekoalgoritmeja. Bottom-top-metodissa tekoälysovellus muodostaa omat eettisen päätöksenteon säännöt analysoimalla tietoa tekemistään päätöksistä ja niiden seurauksista.

Eri tekoälyteknologiat sopivat eri tyylisen päätöksenteon tueksi. Tekoälyn rooli organisaatioiden ja ihmisten päätöksentekoprosessissa vaihtelee sovelluskohteen mukaan. Kun tekoälysovellusta käytetään työkaluna jonkin tietyn tehtävän suorittamiseen, tekee tekoälysovellus päätökset käytännössä itsenäisesti. Tällöin käyttäjän ja kehittäjän vastuulle jää toiminnan ja seurausten laadun valvonta. Toisaalta monissa varsinkin suoraan ihmisen terveyteen vaikuttavissa päätöksissä tekoälysovelluksen tarjoama tieto on vain apuna ihmiselle, joka päätöksen lopulta tekee – vaikkakin enemmän psykologisista ja sosiaalisista kuin teknisistä tai eettisistä syistä (Goldhahn et al., 2018).

4. TEKOÄLY TOIMIJANA

Tekoälyn etiikan tutkimuksessa eräs merkittävä tutkimuskysymys on, että voidaanko tekoälyä hyödyntävää sovellusta pitää moraalisena toimijana. Tämä metaeettinen kysymys jakaa muuta tekoälyn etiikan tutkimusta pohjautumaan kahteen eri oletukseen: yhteen, jossa tekoälysovelluksia pidetään yleisesti moraaliagenttina ja toiseen, jossa tekoälysovellusta ei pidetä moraaliagenttina. Seuraavissa alaluvuissa tutkitaan niitä perusteita, joilla edellä mainittu jako tehdään.

4.1 Tekoälysovelluksen käsittely moraalisubjektina

Moraalisen toimijuuden ehtoja on useita, eikä oikeastaan ole tunnustettu mitään tiettyä vaatimusten sarjaa, jonka täytettyä toimija täyttäisi moraaliagenttiuuden kriteerit. Moraaliselle toimijuudelle tyypillisiä piirteitä on kuitenkin tunnistettu ja niitä ovat muun muassa vapaa tahto (vapaus valita tekemisen ja tekemättä jättämisen välillä), tieto tai tietämys teon perustana, vastuu teoista, kyky arvioida tekojen seurauksia sekä aikomus teon taustalla (Enwald et al., 2007; Dodig-Crnkovic & Persson, 2008; Himma, 2008; Boersema & Bandini, 2014). Vapaata tahtoa voidaan pitää myös autonomiana (Coeckelbergh, 2009).Yleensä toimijaa voidaan pitää moraaliagenttina, jos kaikki tai ainakin osa näistä ehdoista täyttyy (Enwald et al., 2007). Tämä jättää paljon tulkinnanvaraa tekoälysovelluksien etiikan tutkimukselle, koska monen tekoälysovelluksen voidaan katsoa täyttävän osan näistä ehdoista.

Himman (2008) mukaan ollakseen moraalinen agentti, tulee toimijan itse pystyä vapaasti valitsemaan käyttäytymisensä. Tämä ei Himman (2008) mukaan kuitenkaan suoraan tarkoita sitä, ettei esimerkiksi ihmisen (eli moraalisen toimijan) ohjelmoimaa toimijaa voisi pitää moraalisena toimijana. Himman (2008) mukaan yksi perusteellisimmista haasteista keinotekoisen toimijan pitämisessä moraalisena toimijana liittyy vapaan tahdon filosofian perusteelliseen ongelmaan siitä, että miten vapaa tahto määritellään. Tästä syystä tekoälysovelluksen autonomian ajatellaan usein tarkoittavan vapaata tahtoa (Coeckelbergh, 2009), minkä perusteella vapaan tahdon kriteerin katsotaan täyttyneen.

Kriteeriä kyvystä arvioida päätöksien seurauksia on helpompi lähestyä. Teknisesti tarkasteltuna vain sellaiset tekoälysovellukset, jotka ovat tarpeeksi kyvykkäitä arvioimaan päätöksien ja tekojen seurauksia, voisivat teoriassa täyttää tämän moraalisubjektiuuden kriteerin. Tekoälyteknologioiden kannalta kyseessä olevat tekoälysovellukset voidaan

ajatella rajautuvan niihin, joissa hyödynnetään vahvistusoppimista, sillä se on ainoa koneoppimismenetelmä, jonka toiminta perustuu toiminnan mukauttamiseen päätöksiä ja niiden seurauksia arvioimalla. Toisaalta myös muita koneoppimismalleja hyödyntäviä sovelluksia voidaan ohjelmoida arvioimaan vaihtoehtoisten ratkaisujen seurauksia (Alaieri & Vellino, 2016). Ohjatun oppimisen tapauksessa tämä tapahtuisi kuitenkin sillä kustannuksella, että kaikki vaihtoehtoiset lopputulokset olisivat sekä selvitettävissä että ihmisen arvottamia, jotta sovellus pystyisi lopulta tekemään päätöksen itsenäisesti. Toisin sanoen sovelluksen kehittäjien täytyisi opettaa sovellusta noudattamaan ihmisen arvottamia eettisiä periaatteita opetusdatan avulla (Alaieri & Vellino, 2016). Ohjaamattomassa oppimisessa tekoälysovellus muodostaa Alaierin & Vellinon (2016) mukaan omat eettisen päätöksenteon menetelmät, mikä sisältää samoja piirteitä vahvistusoppimismenetelmän kanssa.

Tekoälysovelluksien päätöksien voidaan yleisesti väittää perustuvan tietoon. Koneoppimismenetelmiä hyödyntävien tekoälysovelluksien osalta päätöksentekoon tarvittava tieto on opittu joko opetusdatasta (ohjattu oppiminen ja osin ohjaamaton oppiminen) tai dynaamisesti sovelluksen käyttöympäristöstä (vahvistusoppiminen ja osin ohjaamaton oppiminen). Joskus autonomiset koneet hankkivat tietoa ympäristöstä esimerkiksi erilaisilla sensoreilla (Alaieri & Vellino, 2016), jolloin päätöksenteossa käytettävä tieto ei perustu pelkästään opetusdataan. Jotta päätöksenteko pysyisi autonomisena, ei päätökseen tarvittavaa tietoa voida antaa päätöksentekotilanteessa esimerkiksi ihmisen toimesta, vaan tekoälysovelluksen tulee pystyä perustaa päätös omaan, saatavilla olevaan tietoon. Siispä monen tekoälysovelluksen voidaan sanoa täyttävän moraalisen toimijuuden tietoperustaisen päätöksenteon kriteerin.

Alaieri & Vellino (2016) esittelevät kaksi lähestymistapaa tekoälysovelluksen päätöksien vastuun jakautumiseen. Ensimmäisen, niin kutsutun klassisen lähestymistavan mukaan tekoälysovellusta ei voitaisi koskaan pitää vastuussa teoistaan, koska se on aina mekaaninen työkalu (Alaieri & Vellino, 2016). Toisen, niin kutsutun pragmaattisen lähestymistavan mukaan tekoälysovelluksia ei tulisi käsitellä eristettyinä itsenäisinä kokonaisuuksina (Dodig-Crnkovic & Persson, 2008). Pragmaattisen lähestymistavan mukaan moraalinen vastuu ei ole yksilöllinen velvollisuus, vaan ”ryhmän ulkopuolisten normien määrittelemä rooli” (Dennett 1973 & Strawson 1974, Dodig-Crnkovic & Persson 2008 mukaan s.166). On siis olemassa lähestymistapoja, joiden mukaan tekoälysovellukselle voidaan ajatella jakautuvan päätöksenteon vastuuta sosiaaliteknologisessa käyttöympäristössä (Dodig-Crnkovic & Persson, 2008), jolloin kriteeri moraalisesta vastuusta katsotaan joissain tapauksissa täytetyksi.

Kenties yksiselitteisin moraalisen toimijuuden kriteereistä on intentio, eli aikomus teon taustalla. Dodig-Crnkovicin ja Perssonin (2008) mukaan yksi moraalisen vastuun pääelementeistä on toimijan aikomus tehdä jokin päätös tai teko. Dodig-Crnkovic & Persson (2008) pitävät aikomusta mielentilana. Mielen itsensä ominaisuuksiin kuuluu muun muassa kyky tehdä vapaaehtoisia päätöksiä sekä itsetietoisuus (Nath & Sahu, 2020). Koska tekoälysovelluksella ei ole itsetietoisuutta (Nath & Sahu, 2020), ei sillä voida katsoa olevan ihmismäistä mieltäkään. Tästä voisi päätellä, ettei tekoälysovelluksella voi olla mielentiloja, kuten aikomusta. Kun tekoälyä hyödynnetään työkaluna, aikomus tulee tavalla tai toisella sen hyödyntäjiltä (Kuflik, 1999). Kriteeri aikomuksesta jää siis tekoälysovelluksilla täyttämättä.

Joidenkin kehittyneiden tekoälysovelluksien voidaan siis katsoa täyttävän osan moraalisen toimijuuden kriteereistä. Tekoälyn etiikkaa tutkivassa kirjallisuudessa puhutaankin usein keinotekoisesta moraalisesta toimijuudesta (engl. artificial moral agency, AMA) (Allen et al., 2006; Coeckelbergh, 2009), joka jo käsitteenä tekee eron ihmismäisen moraaliagenttiuuden ja keinotekoisen moraaliagenttiuuden välille. Toisaalta joskus kysymys moraalisesta toimijuudesta ohitetaan kevyin perustein. Esimerkiksi Bench-Capon (2020) pitää toimijaa eettisenä toimijana, jos se käyttäytyy tavalla, jota tulisi pohtia eettisesti, jos ihminen toimisi samalla tavalla. Alaieri & Vellino (2016) taas pitävät yksinomaan toimijan autonomiaa perusteena eettiselle toimijuudelle. Näissä tapauksissa jää kuitenkin monta moraalisubjektiuuden perusteellisista kriteereistä kaikilla artefakteilla eli ihmisen tuottamilla tekotuotteilla (Himma, 2008; Ollila, 2019) täyttämättä. Tässä tutkimuksessa ei pyritä kuitenkaan ratkaisemaan ongelmaa artefaktien moraalisesta toimijuudesta, vaan tarkastellaan tekoälysovelluksien etiikkaa sillä oletuksella, että nämä tekoälysovellukset eivät ole moraalisia toimijoita.

4.2 Toimijuuden luovuttaminen tekoälysovellukselle

Coeckelberghin (2009) mukaan ”toimijuuden ei tarvitse käsittää vapautta tai rationaalisuutta”, kun taas Himman (2008) ja Ollilan (2019) mukaan vapaus valita vaihtoehtoisten tekojen välillä on edellytys myös toimijuudelle. Varsinainen toimijuus edellyttää hyvin pitkälti samoja ominaisuuksia kuin alaluokkansa moraalinen toimijuus (Himma, 2008; Ollila, 2019). Tekoälysovellusta ei voida pitää moraalisena toimijana, eikä itse asiassa edes varsinaisena toimijana. Käytännössä moni tekoälysovellus kuitenkin näyttäytyy ihmisille toimijan kaltaisena ”ikään kuin” -toimijana (Ollila, 2019). Tällaisesta ikään kuin -toimijasta voidaan puhua termillä keinotekoinen toimija (engl. artificial agent). Aitojen toimijoiden

aikomuksella tuottamia tekotuotteita eli artefakteja pidetään keinotekoisina toimijoina. (Himma, 2008; Ollila, 2019)

Keinotekoisina toimijoina tekoälysovelluksia käytetään työkaluina jonkin tehtävän suorittamiseen. Autonominen tekoälyä hyödyntävä sovellus voi olla esimerkiksi itse ajava auto tai rekrytointirobotti, joka pyrkii karsimaan työnhakijoita ihmisen tekemästä rekrytointiprosessista oppimansa datan perusteella. Näitä, kuten muitakin työkaluna käytettyjä tekoälysovelluksia yhdistää se, että ne on rakennettu suorittamaan jotain tiettyä tehtävää tai tietyntyyppisiä tehtäviä. Itseajavan auton tapauksessa tämä tarkoittaa monen turvallisuuteen ja ympäristön havainnointiin liittyvän asian lisäksi yleensä ihmisen tai tavaran kuljettamista paikasta toiseen. Rekrytointirobotin tehtävä on taas hoitaa rekrytointiprosessin yksi työläimmistä vaiheista, eli hakemusten karsimisesta, ihmisen sijaan tämän aikaa säästämiseksi. Tällaiset tekoälysovellukset ovat siis tavallaan osa sosiaali-teknologista yhteiskuntaa (Dodig-Crnkovic & Persson, 2008), jossa näille sovelluksille osoitettu tehtävä on osa jotain suurempaa kokonaisuutta. Liikenne on osa yhteiskuntaa siinä missä työntekijät ja organisaatiot, ja yksittäisen auton ajaminen ja rekrytointiprosessin karsintavaihe ovat näiden osia.

Paikasta toiseen liikkumisen sekä rekrytointitarpeen taustalla on kuitenkin aidon toimijan eli ihmisen intentio. Osa näiden prosessien toiminnoista voidaan osoittaa tekoälyä hyödyntävän sovelluksen tehtäväksi. Kun tällainen sovellus pystyy suorittamaan tekoja, kuten ajamaan autoa tai tehdä päätöksiä rekrytointiprosessissa, voidaan moraalisen toimijan ajatella luovuttavan toimijuutta tekoälysovellukselle. Tällöin prosessin moraalinen vastuu ei Dodig-Crnkovicin & Perssonin (2008) mukaan enää ole pelkästään tehtävän osoittaneella moraalisella toimijalla, vaan tehtävän osoittamisen myötä tekoälysovellukselle osoitetaan myös vastuuta tehtävän turvallisesta ja oikeaoppisesta suorittamisesta. Tämän näkemyksen mukaan moraalinen vastuu jakautuu siis tehtävän, funktion ja roolin mukaan (Dodig-Crnkovic & Persson, 2008). Dodig-Crnkovic & Persson (2008) toteavat kuitenkin myös, että varsinkin turvallisuuskriittisten tekoälysovelluksien kehittäjiä pidetään vastuussa siitä, etteivät nämä sovellukset aiheuta vahinkoa. Alaierin & Vellinon (2016) mukaan tämän voidaan yleistää pätevän muihinkin kuin turvallisuuskriittisiin tekoälysovelluksiin. Esimerkiksi käytännön tekoälysovelluksissa, kuten tekoälyä hyödyntävässä lainanmyöntämisessä tai rekrytointiprosessissa sovelluksen kehittäjät monitoroivat sovelluksen tekemiä päätöksiä ja voivat puuttua sovelluksen toimintaan tarvittaessa. Voidaan siis päätellä, että on lopulta ihmisen vastuulla rakentaa tekoälysovelluksesta sellainen, että se toimii halutulla tavalla – tai vähintään niin, että se ei toimi eitoivotulla tavalla. Tämä tarkoittaa moraalisen vastuun olevan ihmisellä.

5. PÄÄTÖKSENTEON ETIIKKA

5.1 Normatiivinen etiikka

Etiikan osa-alueista normatiivisen etiikan teoriat pyrkivät tarjoamaan ratkaisuperusteita käytännön ongelmille. Näissä teorioissa etiikkaa lähestytään systemaattisen ajattelun kautta. (Ollila, 2019) Päätöksiä ja tekoja voidaan siis punnita normatiivisen etiikan teorioiden avulla. Kun tekoälysovelluksia käytetään päätöksenteon välineenä, sopii normatiivisen etiikan teoriat hyvin tekoälysovelluksien eettiseen tarkasteluun, ja tekoälyn etiikkaa tutkivassa kirjallisuudessa tukeudutaankin usein normatiivisen etiikan teorioihin (Alaieri & Vellino, 2016; Bench-Capon, 2020).

Normatiivisen etiikan teorioita on useita. Eri teorioissa päätöksenteon etiikka perustuu erilaisiin periaatteisiin. Esimerkkejä normatiivisen etiikan teorioista ovat hyve-etiikka, velvollisuusetiikka, seurausetiikka ja oikeusperustainen etiikka (Ollila, 2019). Seuraavissa alaluvuissa käsitellään tarkemmin utilitarismia ja deontologiaa ja sitä, millä perusteilla nämä teoriat sopivat tekoälysovelluksien päätöksenteon etiikan tarkasteluun.

5.1.1 Utilitarismi

Utilitarismi on yksi seurausetiikan klassinen muoto. Seurausetiikka painottaa sananmukaisesti tekojen seurauksia toiminnan moraalisessa arvioinnissa. (Enwald et al., 2007) Utilitarismissa teon moraalisen arvon katsotaan muodostuvan ainoastaan siitä, kuinka paljon hyvää ja onnellisuutta sen seuraukset aiheuttavat (Enwald et al., 2007; Alaieri & Vellino, 2016; Bench-Capon, 2020). Tämä tarkoittaa sitä, ettei moraalisen tarkastelun kohteena ole teko tai päätös itse, vaan moraalinen arviointi ja päätös tehdään ainoastaan teon seurausten perusteella. Tällöin teko itse voi periaatteessa olla jonkin etiikan teorian tai esimerkiksi yhteiskunnallisen normin mukaan huono tai paha, mutta silti utilitarismin mukaan oikein, jos siitä seuraa eniten hyvää.

Tekoälysovelluksien kannalta utilitarismia voidaan soveltaa siis vain sellaisiin sovelluksiin, jotka ovat teknisesti tarpeeksi kyvykkäitä muodostamaan vaihtoehtoisia skenaarioita ja arvioimaan niiden seurauksia (Alaieri & Vellino, 2016; Bench-Capon, 2020). Esimerkiksi itseajavan auton, joka kykenee sensoreillaan tunnistamaan jalankulkijoita tiellä, ja joka pysähtyy tai väistää jalankulkijoita *välttääkseen* vahingon aiheuttamisen heille, voidaan siis ajatella käyttäytyvän eettisesti oikein utilitaristisesta näkökulmasta (Alaieri & Vellino, 2016). Tämä esimerkki edellyttää kuitenkin sitä, että itseajava auto todella arvioi

jalankulkijoiden väistämättömyyden seurauksia ja niiden hyvyyttä kokonaisuuden kannalta. Auton päätös väistää jalankulkijoita ei siis utilitaristisessa tarkastelussa voi perustua esimerkiksi auton kehittäjien, sen käyttäjän tai esimerkiksi lainsäädännön päätökseen siitä, ettei jalankulkijoiden yli saa missään tapauksessa ajaa.

Bench-Caponin (2020) mukaan seurausetiikassa tekoja arvotetaan nimenomaan niiden *odotettujen* seurausten perusteella, eikä todellisten seurausten perusteella. Yksinkertaistetusti tekoälysovelluksen päätös tai teko voi siis olla utilitarismin mukaan moraalisesti oikein, vaikka sen seuraukset olisivat ei-toivottuja ja huonoja, mutta jos sen arvioidut seuraukset olisivat olleet hyviä. Tämä hankaloittaa rajan määrittämistä sille, että mitä pidetään luotettavana arviona seurauksista. Tekoälysovellukset kykenevät arvioimaan seurauksia melko tehokkaasti rajatuissa ympäristöissä, kuten esimerkiksi lautapeleissä (Bench-Capon, 2020). Silti kehittyneenkin tekoälysovelluksen on vaikeaa arvioida todellisen maailman ei-rajattujen vaihtoehtojen seurauksia (Bench-Capon, 2020), sillä tekoälysovelluksia on mahdotonta opettaa kaikilla mahdollisilla vaihtoehdoilla tai testata kaikissa mahdollisissa käyttötapauksissa (Dodig-Crnkovic & Persson, 2008). Tekoälysovelluksen päätöksentekoprosessin tarvitsee siis olla tarpeeksi läpinäkyvä, jotta päättelyketjuja ja tietoperustaisuutta voidaan arvioida. Alaierin & Vellinon (2016) mukaan päätösten perustelu on tärkeää myös tekoälysovelluksen luotettavuuden kannalta.

Utilitaristisessa tarkastelutavassa yhdeksi olennaiseksi kysymykseksi muodostuu myös se, että miten tekoälysovellus määrittelee hyvän ja pahan. Käytännössä tekoälysovellukseen tulisi siis jotenkin määrittää, mikä lasketaan missäkin tilanteessa hyväksi ja mikä pahaksi. Useimmissa tapauksissa tekojen hyvyyden arvioinnin voisi ajatella lopulta perustuvan ihmisen tai yhteiskunnan määrittämiin arvoihin, jotka joko ohjelmoidaan tekoälysovellukseen tai annetaan tekoälysovelluksen oppia ne itse. Tekoälysovelluksen itse oppimat arvot perustuvat kuitenkin lopulta ihmisten määrittelemiin arvoihin, sillä tekoälysovellus oppii arvot ihmisten käyttäytymisestä joko opetusdatasta tai dynaamisesti siitä ympäristöstä, jossa tekoälysovellusta käytetään.

Utilitaristinen lähestymistapa soveltuu tarkastelutavaksi sellaisiin tekoälysovelluksiin, joiden päätöksentekoprosessi on kolmosluvussa esitellyn bottom-up-metodin mukainen (Alaieri & Vellino, 2016). Tämä perustuu siihen, että bottom-up-metodissa tekoälysovellus pyrkii muun muassa ennustamaan tekojensa seurauksia ja valitsemaan tekonsa niiden perusteella (Alaieri & Vellino, 2016). Tekoälyteknologioiden näkökulmasta tämä tarkoittaa kompleksisia neuroverkkoteknologioihin perustuvia sovelluksia, joilla on kyky syvään adaptiiviseen oppimiseen (Alaieri & Vellino, 2016). Adaptiivisesti oppivat, eli vahvistusoppimista hyödyntävät järjestelmät soveltuvat utilitaristiseen tarkasteluun siksi, että ne arvioivat vaihtoehtoisia päätöksiä ja niiden seurauksia.

5.1.2 Deontologia

Deontologia tarkoittaa velvollisuusetiikkaa ja se on yksi normatiivisen etiikan teorioista (Enwald et al., 2007; Ollila, 2019). Deontologisessa etiikassa teon moraalinen arvo määräytyy ennalta määrättyjen sääntöjen ja ohjeiden perusteella (Alaieri & Vellino, 2016; Bench-Capon, 2020). Näitä ohjeita voi olla monia ja ne voivat perustua esimerkiksi lainsäädäntöön tai soveltajan omiin arvoihin. Toisin kuin utilitarismissa, deontologiassa teon arvo on siis riippumaton tekojen seurauksista (Alaieri & Vellino, 2016; Bench-Capon, 2020). Deontologia edellyttää siis yksittäisten sääntöjen ja tekojen hyvyyden moraalista arviointia täysin riippumatta siitä, mitä seurauksia ne aiheuttavat. Deontologisissa säännöissä taustalla on jo tehty pohdinta siitä, onko jokin tietty teko tai sääntö itsessään hyvä vai paha ja miksi. Deontologisen etiikan teorian mukaan moraalisesti oikeita sääntöjä ovat esimerkiksi seuraavat: älä tapa, älä aiheuta kipua, älä vammauta, pidä lupauksesi, älä valehtele ja älä riistä vapautta tai nautintoa (Gert, 1998).

Deontologista etiikkaa on utilitarismia suoraviivaisempaa soveltaa tekoälysovelluksiin. Alaierin & Vellinon (2016) mukaan käytännössä deontologisesta näkökulmasta tekoälysovellukselta vaaditaan vain kyvykkyyttä seuraamaan ohjeita. Bench-Capon (2020) puolestaan ajattelee, että tekoälysovelluksen tulisi pystyä määrittämään kaikki vaihtoehtoiset tilanteet ja polut, mikä tarkoittaa hyvin rajattua ja pientä toimintaympäristöä. Näin ollen deontologista etiikkaa voidaan soveltaa laskentakyvyltään paljon rajallisempiin sovelluksiin kuin esimerkiksi utilitaristista etiikan teoriaa. Deontologiaa voidaan siis soveltaa sellaiseen tekoälysovelluksen päätöksentekotilanteeseen, jossa sen tulee valita ennalta määritellyistä vaihtoehdoista paras ennalta määrättyjen ohjeiden ja sääntöjen perusteella. Tekoälysovellus siis käytännössä seuraa suunnittelijansa määrittelemiä ohjeita (Bench-Capon, 2020).

Yksi deontologian ongelmista on mahdollisuus ristiriitaan ohjeiden välillä (Tzafestas, 2016). Itseajavan auton tapauksessa tällainen ristiriita voisi tarkoittaa esimerkiksi tilannetta, jossa auto ei ennalta määriteltyjen sääntöjen mukaan saa koskaan tappaa jalankulkijaa, mutta jossa auton tulee aina suojella kuljettajaa. Törmäystilanteessa olisi kaksi vaihtoehtoa: aiheuta kuolema jalankulkijalle törmäämällä tähän tai aiheuta kuolema kuljettajalle törmäämällä seinään. Nyt auton säännöt ovat selvässä ristiriidassa, jonka selvittämiseksi auto tarvitsisi oman mekanisminsa (Allen et al., 2006; Alaieri & Vellino, 2016). Jos tilanne tulisi ihmiskuljettajan eteen, olisi tällä ratkaistavanaan täsmälleen sama dilemma. Tekoälysovellukset ja ihmiset voivat siis törmätä samankaltaisiin moraalisiin ongelmiin (Alaieri & Vellino, 2016).

Toinen deontologisen etiikan soveltamisen ongelma on siinä, että tekojen seurauksia ei oteta huomioon tekojen moraalisessa arvioinnissa. Joskus säännön noudattamisella voi olla ei-toivottuja seurauksia (Bench-Capon, 2020). Itseajavan auton esimerkissä tällainen tilanne voisi syntyä, jos auton tulee sääntöjen mukaan aina väistää eteen juoksevaa olentoa, minkä seurauksena auto ajautuu seinään surmaten tai vammauttaen kuljettajan. Monen säännön noudattamisella voi siis olla ilmiselviä huonoja seurauksia, joita deontologinessa lähestymistavassa ei oteta huomioon. Tällaisia tilanteita voidaan kuitenkin ratkoa poikkeuksilla (Bench-Capon, 2020). Poikkeustilanteissa sovellus voisi toimia vastoin jotain tiettyä sääntöä. Tämä edellyttäisi kuitenkin poikkeustilanteiden tunnistamista, joka puolestaan vaatisi jonkinasteista tekojen seurauksien arvioimista. Lisäksi tämä johtaa samaan ongelmaan utilitarismin kanssa: niin ihmisen kuin tekoälysovelluksenkin on mahdotonta arvioida kaikkia mahdollisia seurauksia teolle (Bench-Capon, 2020).

Päätöksentekoprosessina deontologiseen tarkastelutapaan soveltuu kolmosluvussa esitetty top-down-metodi (Alaieri & Vellino, 2016). Metodi soveltuu deontologiseen tarkasteluun siksi, että siinä päätöksenteko perustuu tekoälysovelluksen suunnittelijan määrittelemiin päätöksentekoalgoritmeihin (Alaieri & Vellino, 2016). Käytännössä tekoälysovelluksen suunnittelijat ohjelmoivat siihen ennalta määritellyt säännöt ja arvot, joita sovellus noudattaa. Tekojen moraalinen arviointi on siis täysin tekoälysovelluksesta ja sen senhetkisestä käyttöympäristöstä riippumatonta.

Jos tekoälysovelluksen päätöksenteko perustuu ennalta määriteltyihin arvoihin ja sääntöihin, ei sen toteuttamiseksi tarvita esimerkiksi vahvistusoppimista tukevaa tekoälyteknologiaa. Tällaisen sovelluksen kehittäjät voivat siis opettaa sovellukselle ennalta määriteltyjä moraalisia sääntöjä esimerkiksi ohjatun oppimisen keinoin (Alaieri & Vellino, 2016). Deontologista etiikkaa voidaan siis soveltaa yksinkertaisempiin tekoälysovelluksiin kuin utilitaristista etiikkaa.

5.2 Tarkasteltujen etiikan teorioiden valitseminen

Tekoälyteknologioita tarkastelemalla voidaan todeta, että eri teknologiaa hyödyntävillä tekoälysovelluksilla on erilaiset päätöksentekoprosessit. Normatiivisen etiikan teorioita tarkastelemalla voidaan puolestaan huomata, että erityyppisten päätöksentekoprosessien eettiseen arviointiin soveltuu erityyppiset etiikan teoriat. Esimerkiksi deontologia soveltuu nopeaan ja spontaaniin päätöksentekotilanteeseen, jossa päätöksen moraalinen arvo muodostuu ainoastaan teon hyvyyden perusteella. Utilitarismia voidaan sen sijaan soveltaa monimutkaisempiin päätöksentekotilanteisiin, joissa päätösten moraalinen arvo

määräytyy päätöksen seurausten hyvyyden perusteella. Voidaan siis päätellä, että tekoälysovelluksen päätöksenteon eettiseen arviointiin voidaan soveltaa siinä käytetystä teknologiasta riippuen eri etiikan teorioita.

Tarkasteltu etiikan teoria voidaan siis valita tekoälysovelluksen päätöksentekokykyjen (Alaieri & Vellino, 2016) eli lopulta siinä käytettyjen teknologioiden perusteella. Jos tekoälysovellus hyödyntää vahvistusoppimista ja kykenee arvioimaan päätöksien seurauksia ja oppimaan niistä, voidaan sen eettiseen arviointiin soveltaa utilitarismia. Jos tekoälysovellus sen sijaan pystyy vain oppimaan moraaliset säännöt ohjattua oppimista hyödyntäen ja toimimaan niiden perusteella, sopii sen eettiseen tarkasteluun paremmin deontologia. (Alaieri & Vellino, 2016; Bench-Capon, 2020)

Joskus tekoälysovellukset voivat hyödyntää useita tekoälyteknologioita. Tällöin päätöksentekoprosesseissa saatetaan hyödyntää sekä vahvistusoppimista että jotain tiettyjä sääntöjä päätöksien taustalla. (Alaieri & Vellino, 2016) Tällöin sovellus voisi toimia tiettyjen periaatteiden pohjalta, mutta myös oppia uusia eettisiä periaatteita kokemuksistaan (Alaieri & Vellno, 2016). Tällaista normatiivisen etiikan teorioiden yhdistämistä sovelletaan todellisessa maailmassa ihmistenkin puolesta. Esimerkiksi toisen ihmisen vahingoittaminen on väärin, mutta jos sen tekee esimerkiksi puolustaakseen itseään, voi se olla oikeutettua. Samaa voitaisiin soveltaa tekoälysovelluksiin, mikä tekisi kuitenkin etiikan teorioiden valitsemisesta kompleksisempaa.

6. YHTEENVETO

6.1 Tulosten esittely

Tutkimuksessa selvitettiin ensimmäisenä tekoälyn käsite sekä tekoälyn tyypillisiä ominaisuuksia. Järjestelmän voidaan sanoa hyödyntävän tekoälyä, jos se on autonominen ja adaptiivinen. Tekoälysovelluksen tulee siis pystyä suorittamaan toimintoja ja päätöksiä ilman ihmisen ohjaamista. Sovelluksen tulee lisäksi pystyä oppimaan, eli parantamaan suoritustaan saamansa palautteen perusteella. Oppimismenetelminä tekoälysovellukset käyttävät koneoppimista, joka jakautuu ohjattuun oppimiseen, ohjaamattomaan oppimiseen ja vahvistusoppimiseen. Nämä oppimismetodit vaativat järjestelmältään eri tyyppistä arkkitehtuuria ja teknologiaa. Edistykselliset tekoälysovellukset hyödyntävät neuroverkkoteknologioita ja pystyvät täten syväoppimaan, eli yhdistämään datan eri ulottuvuuksia oppimaansa.

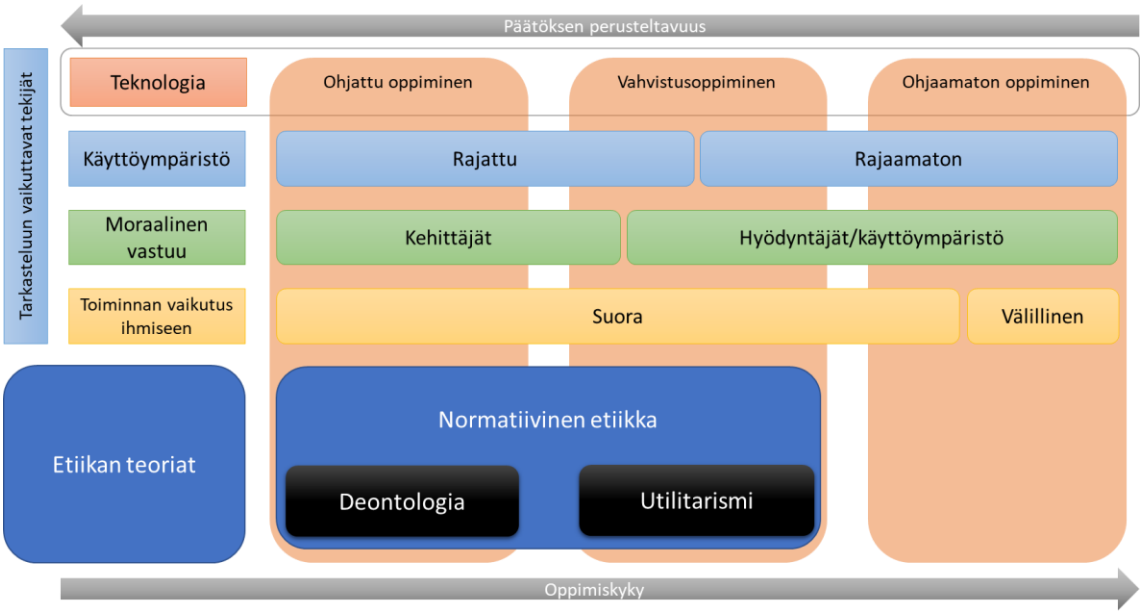
Tutkimuksessa havaittiin, että ihmisen ja organisaatioiden päätöksentekoprosessi on monilta osin samanlainen tekoälysovelluksen päätöksentekoprosessin kanssa. Molemmat päätöksentekoprosessit ovat tietoperustaisia ja oppivia. Tutkimuksessa selvisi myös, että tekoälysovelluksia käytetään työkaluna sekä ihmisen tai organisaation päätöksenteon tukena että niille osoitettujen tehtävien suorittamisessa (esimerkiksi automatisoitu lainanmyöntäminen).

Tekoälysovelluksella voi olla monia toimijan ja jopa moraalisen toimijan kaltaisia piirteitä, kuten autonomia, kyky tuottaa vaihtoehtoisia ratkaisuja ja valita paras niiden joukosta sekä kyky oppia tekemistään päätöksistä. Yleisesti ottaen kaikilta tekoälysovelluksilta puuttuu kuitenkin useampi toimijuuden kannalta välttämätön kriteeri. Tekoälysovelluksella ei voi olla itsetietoisuutta eikä mieltä. Tämän vuoksi tekoälysovelluksella ei voi olla mielentiloja, kuten intentiota eli aikomusta. Näin ollen edes edistyksellistä tekoälysovellusta ei voi pitää moraalisena toimijana, eikä edes varsinaisen toimijuuden kriteerit täyty.

Tekoälysovelluksille voidaan kuitenkin ajatella luovutettavan toimijuutta niille osoitettujen tehtävien mukana. Tekoälysovelluksia käytetään joskus suoraan ihmisten elämään vaikuttavien asioiden käsittelyssä, joten niiden päätöksenteon tulee olla eettistä ja kestävää. Tutkimuksessa todettiin, että vaikka vastuuta ja toimijuutta luovutetaan tekoälysovellukselle, on vastuu eettisestä päätöksenteosta ja toiminnasta ja sen valvomisesta aina sovelluksen kehittäjillä ja hyödyntäjillä. Vastuu jakautuu sen mukaan, millaisessa roolissa tekoälysovellus on käytetyssä kontekstissa ja päätöksenteossa.

Tutkimuksen keskeisin löydös on se, että eri tekoälysovelluksien päätöksenteon etiikkaa voi ja kannattaa lähestyä eri etiikan teorioiden näkökulmasta. Tähän lähestymistavan valintaan vaikuttaa muun muassa käyttöympäristö, käyttötarkoitus sekä käytetty teknologia. Tekoälysovelluksessa käytetty teknologia vaikuttaa tekoälyn päätöksenteon etiikan tarkasteluun merkittävästi. Etiikan teorian valitsemiseen vaikuttaa erityisesti tekoälysovelluksen tekniset kyvykkyydet, kuten kyky arvioida päätöksiensä seurauksia ja kyky oppia niistä.

Etiikan teorioista tekoälyavusteisen päätöksenteon tarkastelemiseen sopivaksi osoittautui normatiivisen, eli ohjailevan etiikan teorioita. Tutkimuksessa tarkasteltiin normatiivisen etiikan teorioista yksityiskohtaisemmin utilitarismia ja deontologiaa. Utilitarismin todettiin soveltuvan teknisesti edistyksellisiin ja syvästi oppiviin järjestelmiin. Utilitaristista etiikan teoriaa voidaan soveltaa sellaisiin tekoälyjärjestelmiin, jotka kykenevät arvioimaan päätöksiensä seurauksia ja niiden hyvyyttä sekä oppimaan päätöksistään. Teknologian kannalta tämä tarkoittaa järjestelmiä, jotka hyödyntävät vahvistusoppimista. Deontologia sopii puolestaan sellaisiin käyttötilanteisiin ja järjestelmiin, joissa joko kaikki mahdolliset vaihtoehdot on laskettavissa tai niitä ei lasketa lainkaan. Deontologiassa teon moraalinen arvo määräytyy ainoastaan teon itsensä hyvyyden perusteella. Näin ollen deontologinen etiikka sopii sellaisten järjestelmien tarkasteluun, jonka moraaliset arvot ovat ennalta opetettuja eivätkä ne muutu käytännön tilanteiden myötä. Deontologia soveltuu siis lähinnä ohjatun oppimisen koneoppimismenetelmiä hyödyntävien tekoälysovelluksien päätöksenteon etiikan tarkasteluun.



Kuva 3: Eettisen lähestymistavan valinta ja siihen vaikuttavat tekijät

Kuvassa 3 on havainnollistettu tekoälysovelluksen niitä elementtejä, joiden perusteella päätöksenteon etiikan pohtimiseen sovellettava teoria voidaan valita. Kuvassa olevat horisontaaliset palkit koneoppimismenetelmien päällä kuvaavat sitä, kuinka vahvasti niiden ajatellaan koskevan kyseessä olevaa koneoppimismenetelmää. Esimerkiksi vahvistusoppimista hyödyntävän sovelluksen käyttöympäristö voi olla sekä rajattu että rajaamaton. Rajatulla käyttöympäristöllä tarkoitetaan sellaista ympäristöä, jonka kaikki mahdolliset tilat on selvitettävissä. Tällaisia voivat olla esimerkiksi laboratoriot ja virtuaalimaailmat. Rajaamattomassa käyttöympäristössä taas kaikkia mahdollisia käyttötilanteita ja tiloja ei voida tuntea eikä laskea. Todellisessa maailmassa sovelletut tekoälysovellukset toimivat käytännössä siis rajaamattomassa käyttöympäristössä.

Lopulta moraalinen vastuu ei kuulu tekoälysovellukselle, vaan siihen vaikuttavat sen käyttöympäristö, hyödyntäjät sekä kehittäjät. Jos tekoälysovellusta käytetään rajatussa ympäristössä ja se ei opi päätöksiensä seurauksista, on sovelluksen eettisten käytäntöjen noudattaminen kehittäjän vastuulla. Taustalla voi kehittäjästä riippumatta olla esimerkiksi lakeja ja muita sääntöjä. Adaptiivinen ja toimintatapojaan oppimansa mukaan muuttava tekoälysovellus puolestaan oppii eettiset toimintatavat käyttöympäristöstään. Tällä tarkoitetaan sitä, että esimerkiksi ohjaamatonta oppimista hyödyntävä kone oppii objektiivisesti ympäristönsä käyttäytymismalleja ottamatta kantaa niiden merkitykseen ja oikeellisuuteen. Vahvistusoppiminen toimii osin samalla tavalla: jos tekoälysovellus oppii moraaliset arvot ja toimintatavat hyödyntäjiltään ja/tai käyttöympäristöstään, ei tekoälysovelluksen toiminnan vastuu ole enää sen kehittäjällä.

6.2 Tulosten arviointi ja lisätutkimuksen tarve

Tutkimuksen tavoite oli selvittää, että millä keinoilla tekoälyavusteisen päätöksenteon eettistä kestävyyttä voidaan tarkastella. Tutkimusongelmaa tukivat alakysymykset, joihin vastaamalla saatiin lopulta koottua vastaus päätutkimusongelmaan. Sekä päätutkimusongelma että alakysymykset ovat ajankohtaisia ja valtaosasta niistä on saatavilla suuri määrä koherenttia tutkimusta, mikä helpotti tutkimuksen suorittamista. Tutkimuksessa yhdistyi tekoälyteknologian käsitteistöä, päätöksenteon prosessit sekä etiikan tutkiminen.

Vaikka aihetta on aiemmin tutkittu paljon, jouduttiin tutkimuksen tietyillä osa-alueilla luottamaan paljon muutamiin yksittäisiin lähteisiin. Tätä voidaan selittää sillä, että juuri tämän tutkimuksen keskittymistä osa-alueista ei ole tehty vielä paljoa tutkimusta. Tutkimus keskittyi kahteen normatiivisen etiikan teoriaan siksi, että niitä on tutkittu tekoäly-ympäristössä enemmän kuin muita etiikan teorioita. Näistä teorioista kertovat tutkimukset olivat päätyneet pitkälti samoihin tuloksiin, joten tämänkin tutkimuksen tuloksia voidaan pitää niiltä osin luotettavana. Toimijuudesta sekä deontologian ja utilitarismin soveltamista tekoälyavusteisen päätöksenteon etiikkaan löydettiin tässä tutkimuksessa hyvin tuloksia.

Tutkimuksessa jäi selvittämättä muiden normatiivisen etiikan teorioiden lähestymistapoja sekä muiden kuin normatiivisten etiikan teorioiden lähestymistapoja tekoälyn etiikan tutkimukselle. Tutkimuksessa ei löydetty yhtä suoraviivaista lähestymistapaa ohjaamattoman oppimisen sovelluksiin kuin ohjatun ja vahvistusoppimisen sovelluksiin. Jatkotutkimukselle olisi siis useita aiheita. Jatkotutkimuksissa voitaisiin lisäksi selvittää vielä enemmän käytännön tasolla, että mikä etiikan teoria sopii millaisenkin sovelluksen tarkasteluun. Lisäksi käytännön tekoälysovelluksissa yhdistyy moni koneoppimismenetelmä ja käyttötarkoitus ja -ympäristö, mikä tarjoaa tutkimusaiheita eri teorioita ja teknologioita yhdistävistä lähestymistavoista.

LÄHTEET

Alaieri F., Vellino A. (2016) Ethical Decision Making in Robots: Autonomy, Trust and Responsibility. In: Agah A., Cabibihan JJ., Howard A., Salichs M., He H. (eds) Social Robotics. ICSR 2016. Lecture Notes in Computer Science, vol 9979. Springer, Cham.

Allen, C., Smit, I., Wallach, W. (2005) Artificial morality: top-down, bottom-up, and hybrid approaches. Ethics Inf. Technol. 7.

Allen, C., Smit, I., Wallach, W. (2006) Why Machine Ethics? IEEE Intelligent Systems, vol. 21, no. 4, pp.12-17.

Alyoubi, B. A. (2015) Decision support system and knowledge-based strategic management. Procedia Computer Science, No. 65, pp. 278-284. Saatavilla www-muodossa: https://www.sciencedirect.com/science/article/pii/S1877050915029099 (viitattu 16.3.2020)

Bench-Capon, T.J.M. (2020) Ethical approaches and autonomous systems, Artificial Intelligence, vol. 281.

Boersema, D. & Bandini, C. (2014) Dimensions of moral agency. Newcastle upon Tyne, England: Cambridge Scholars Publishing.

Cervantes, J., Rodríguez, L., López, S., Ramos, F., Robles, F. (2016) Autonomous Agents and Ethical Decision-Making. Cogn Comput 8, pp. 278–296. Saatavilla www-muodossa: https://doi.org/10.1007/s12559-015-9362-8 (viitattu 7.3.2020)

Coeckelbergh, M. (2009) Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents. AI & Society, vol. 24, no. 2, pp. 181-189.

Courtney, J. F. (2001) Decision making and knowledge management in inquiring organizations: toward a new decision-making paradigm for DSS. Decision support systems, 31(1), pp. 17-38.

Dodig-Crnkovic, G., Persson, D. (2008) Sharing moral responsibility with robots: A pragmatic approach. Frontiers In Artificial Intelligence And Applications.

Enwald, M., Keinänen, J., Vadén, T. (2007) Etiikan haasteet. Keuruussa: kustannusosake-yhtiö Atena.

Gert, B. (1998) Morality: Its Nature and Justification. Oxford University Press, USA

Goldhahn, J., Rampton, V., Spinas, G. A. (2018) Could artificial intelligence make doctors obsolete? BMJ. saatavilla www-muodossa: http://bmj.com/content/363/bmj.k4563.full.pdf (viitattu 11.3.2020)

Gunkel, D.J. & Bryson, J. (2014) Introduction to the Special Issue on Machine Morality: The Machine as Moral Agent and Patient, Philosophy & Technology, vol. 27, no. 1, pp. 5-8.

Guo, Y., Wang, N., Xu, Z.-. & Wu, K. (2020) The internet of things-based decision support system for information processing in intelligent manufacturing using data mining technology. Mechanical Systems and Signal Processing, vol. 142.

Hady M.F.A., Schwenker F. (2013) Semi-supervised Learning. In: Bianchini M., Maggini

Himma, K. E. (2008) Artificial agency, consciousness, and the criteria

for moral agency: what properties must an artificial agent have to

be a moral agent? Ethics and Information Technology (2019) 11. pp. 12-29.

Kuflik, A. (1999) Computers in control: Rational transfer of authority or irresponsible abdication of autonomy?. Ethics and Information Technology 1. pp. 173–184.

Lecun, Y., Bengio, Y., Hinton, G. (2015) Deep learning. Nature. 521 (7553), pp. 436–444. Saatavilla www-muodossa: https://search.proquest.com/docview/1685003444?rfr_id=info%3Axri%2Fsid%3Aprimo (viitattu 14.4.2020)

Louridas, P. & Ebert, C. (2016) Machine Learning. IEEE Software. 33 (5), pp. 110–115. Saatavilla www-muodossa: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7548905 (viitattu 9.3.2020)

Maggini, M., Bianchini, M., Jain, L. C. (2013) Handbook on neural information processing. Heidelberg: Springer.

Nath, R., Sahu, V. (2020) The problem of machine ethics in artificial intelligence. AI & Soc 35, 103–111. Saatavilla www-muodossa: https://link.springer.com/article/10.1007/s00146-017-0768-6 (viitattu 7.4.2020)

Ollila, M.-R. (2019) Tekoälyn etiikkaa. Helsingissä: Kustannusosakeyhtiö Otava.

Russell, S. J. & Norvig, P. (2003) Artificial intelligence: a modern approach. 2nd ed. Upper Saddle River (NJ): Prentice Hall.

Salminen, A. (2011) Mikä kirjallisuuskatsaus? Johdatus kirjallisuuskatsauksen tyyppeihin ja hallintotieteellisiin sovelluksiin. Vaasan yliopisto, Vaasa.

Sachan, S., Yang, J.-., Xu, D.-., Benavides, D.E. & Li, Y. (2020) An explainable AI decision-support-system to automate loan underwriting. Expert Systems with Applications, vol. 144.

Shanmuganathan, S. & Samarasinghe, S. (2016) Artificial Neural Network Modelling. 1st ed. 2016. Cham: Springer International Publishing. Saatavilla www-muodossa: https://link.springer.com/chapter/10.1007/978-3-319-28495-8_1 (viitattu 14.4.2020)

Standard Encyclopedia of Philosophy. (2018) Atificial Intelligence. Saatavilla www-muodossa: https://plato.stanford.edu/entries/artificial-intelligence/#Bib (viitattu 7.3.2020)

The Elements of AI. Helsingin yliopiston ja Reaktorin verkkokurssi. Saatavilla www-muodossa: https://www.elementsofai.com

Tzafestas, S.G. (2016) Roboethics. A Navigating Overview, vol. 79. Springer, Heidelberg

Wei Choo, C. (2001) The knowing organization as learning organization, Education + Training, Vol. 43 No. 4/5, pp. 197-205. Saatavilla www-muodossa: https://doi.org/10.1108/EUM0000000005482 (viitattu 16.3.2020)

Wojtusiak J. (2012) Machine Learning. In: Seel N.M. (eds) Encyclopedia of the Sciences of Learning. Springer, Boston, MA.

Yim, N. H., Kim, S. H., Kim, H. W., & Kwahk, K. Y. (2004) Knowledge based decision making on higher level strategic concerns: system dynamics approach. Expert Systems with Applications, 27(1), pp. 143-158.